Baro2Talk: Reconstructing Spectrograms from Ear Canal Pressure for Voice-free Communication

Luo Zhou*, Shan Chang*, Han Wang†, Xianbo Wang*, Hongzi Zhu‡
* Donghua University, China
† University of Electronic Science and Technology of China, China
‡Shanghai Jiao Tong University, China
{zhouluo, wxb}@mail.dhu.edu.cn, changshan@dhu.edu.cn, wang_han@std.uestc.edu.cn, hongzi@cs.sjtu.edu.cn

***Abstract*—The increasing demand for private and noise-resilient speech interaction has motivated the development of Silent Speech Interfaces (SSIs) that infer user intent without vocalization. Existing SSI solutions face limitations such as intrusiveness, privacy leakage, environmental sensitivity, deployment complexity, and motion vulnerability. In this work, we present Baro2Talk, a wearable SSI system that reconstructs speech content from TMJ-dominated Pressure Variation Sequences (TPVSs) captured by miniature barometers embedded in standard earbuds. Baro2Talk is inspired by two key observations. First, silent articulation induces consistent ear canal deformation via temporomandibular joint (TMJ) movements, producing pressure fluctuations that reflect articulatory patterns associated with speech. Second, TPVSs exhibit repeatable temporal and articulatory structures within phrases, offering structured signals to support semantic modeling without acoustic input. We develop a lightweight in-ear pressure sensing prototype and propose a set of modules that first perform articulatory event detection and generalization enhancement, followed by a three-stage reconstruction pipeline: Semantic Encoding, Coarse Mel-spectrogram Construction, and Phonetic Enhancement. The resulting spectrograms are decoded into text using a pre-trained automatic speech recognition (ASR) model (e.g., Whisper). Baro2Talk achieves a 6.5% CER, 9.9% WER, and a 0.081 spectral convergence score, demonstrating robust performance in silent, mobile, and noisy environments.**

***Index Terms*—silent speech interface, ear canal deformation, barometric signal, spectrogram reconstruction.**

I. Introduction

Speech is one of the most natural and efficient modalities for human-computer interaction (HCI), and increasingly becoming the dominant input interface for mobile devices, wearable systems, and intelligent assistants. According to recent forecasts [1], the global voice recognition market is projected to grow rapidly in the coming years, reaching 18.41 billion dollars by 2025 and surging to 77.97 billion dollars by 2032, with a compound annual growth rate of 22.9% during this period. Despite its prevalence, audible speech faces significant challenges in practical applications, particularly in noisy surroundings, privacy-sensitive situations, or among individuals with impaired speech ability. **Silent Speech Interfaces (SSIs)** are designed to accurately understand the content without vocalization. This presents a promising solution for natural and resilient HCI through voice-free communication in scenarios where vocal speech is inaccessible or inappropriate. This has attracted widespread attention in academia and industry.

Corresponding author: Shan Chang.

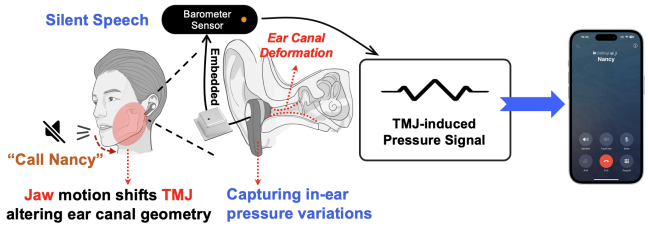


Fig. 1: Conceptual illustration of Baro2Talk.

In order to ensure practical usability in real-world scenarios, ideally, a successful SSI should satisfy the following key requirements including *non-intrusiveness, privacy preservation, robustness to environmental variability, zero-effort deployment, and resilience to user motion*, thereby ensuring practical usability across diverse real-world scenarios. Existing SSIs can roughly be classified into three categories: 1) *Vision-based* SSIs [2]–[4] rely on cameras to capture facial or lip movements, inferring semantic intent through image sequence analysis. This requires clear visual observations and is susceptible to the effects of obstruction, changes in perspective and light variations. Moreover, they raise privacy concerns in public settings. 2) *Wireless-based* SSIs [5]–[11] detect articulatory motion through acoustic or millimeter-wave signals. This not only requires continuous signal emission, but also is often inconvenient to use (e.g., intrusive) and motion sensitive. 3) *IMU-based* SSIs [12], [13] employ one or more inertial sensors positioned near articulatory regions—such as the jaw, chin, throat, or beneath the tongue—to capture subtle movements of the temporomandibular joint (TMJ). Since the TMJ connects the temporal bone and mandible and plays a crucial role in speech-related motions, this has the potential to accurately capture speech. However, this requires additional sensor deployment and is susceptible to motion artifacts, which may not be practically feasible.

In this paper, we present ***Baro2Talk***, a novel SSI that leverages in-ear pressure sensing to reconstruct speech content. As illustrated in Fig. 1, *the key insight in the design of Baro2Talk is that speech articulation—even when performed silently—induces subtle fluctuations in ear canal pressure. These pressure variations—primarily driven by articulatory motions dominated by the TMJ (along with the mandible, tongue, and other surrounding oral structures)—exhibit consistent and semantically informative patterns.* We refer to these sequences as *TMJ-dominated Pressure Variation Sequences*

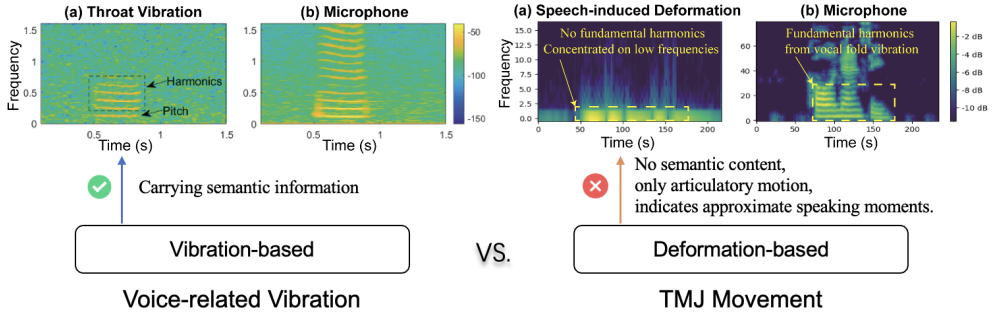


Fig. 2: Problem statement of Baro2Talk. Unlike vibration-based approaches (left) that capture pitch and harmonics from vocal fold vibrations, deformation-based signals (right) induced by TMJ motion carry only articulatory patterns without semantic.

(TPVSs). This presents an opportunity to establish a mapping model between ear pressure variations and their underlying cause—TMJ movements—and ultimately semantics. In other words, it is possible to reconstruct speech content from ear pressure signals. To this end, we develop a prototype with miniature barometers embedded in standard earbuds to capture TPVSs, and use them to reconstruct Mel-spectrograms rather than directly recover textual content. This is because the sampling rate of pressure signals, approximately 100 Hz, is much lower than that of audio signals, creating a significant dimensional mismatch between the low-frequency TPVSs and the high-dimensional textual embeddings. Directly mapping TPVSs to text would require an enormous amount of data. In contrast, Mel-spectrograms transform TPVSs into high-frequency representations that align with speech characteristics, preserving critical features such as formants and pitch contours while bridging the modality gap.

Furthermore, since Mel-spectrograms serve as the standard input for general automatic speech recognition (ASR) systems (e.g., Whisper [14]), these become ideal intermediate representation. Hence, our main target in this work is to leverage TPVS dataset (paired with corresponding transcripts) to pre-train a user-independent Mel-spectrogram synthesizer. However, this turns out to be non-trivial and brings out three key challenges:

1) Low Signal-to-Noise ratio and Interference-prone Pressure Signals: Pressure sequences are inherently low in magnitude and susceptible to drift and interference caused by altitude variations, earbud seal inconsistencies, and subtle head movements. These factors result in low signal-to-noise ratios and unstable deformation patterns, posing challenges for reliably identifying segments of speech-relevant sequences.

Our Solution: We propose a preprocessing pipeline with DC drift removal, band-pass filtering, and signal amplification to reduce noise and clarify deformations. A short-term energy event detect method, enhanced by a local stability check, is then applied to accurately extract silent speech deformation events from continuous pressure signals (§ IV).

2) Inter-user Variation and Rhythmic Diversity: TPVS patterns differ across users due to anatomical and behavioral variability, and speaking rhythms may fluctuate for the same user across time and contexts. These factors reduce model generalizability in real-world deployments.

Our Solution: We pre-train a *Baro-Encoder* using domain adversarial learning to extract user-invariant semantic features. Meanwhile, we design a rhythm-based data augmentation strategy that generates temporally warped variants of TPVSs, enhancing robustness to speaking pace variation (§ V).

3) Non-acoustic Modality and Lack of Fine-grained Supervision: As shown in Fig. 2, unlike vibration-based signals [15]–[17], TPVSs stem from internal articulatory motion without acoustic energy. Their spectrograms are not inherently meaningful in the speech domain, rendering direct mapping to audio spectrograms infeasible. Moreover, silent speech lacks aligned phoneme or frame-level labels, preventing the use of traditional supervised regression or alignment-based models.

Our Solution: We propose a three-stage reconstruction pipeline that decouples semantic understanding from spectrogram generation. First, a semantic encoder named S-Former maps the entire TPVS of a sentence into a latent semantic space shared with its textual counterpart, avoiding the need for alignment. Second, the latent vector is then used to progressively generate Mel-spectrograms via MS-GAN. Finally, a phonetic enhancement via adaptive residual learning (PEARL) refines them, enabling high-fidelity reconstruction without acoustic supervision (§ VI).

We evaluate Baro2Talk on a dataset spanning six months, collected from 25 participants under both silent and acoustic articulation conditions. The system outperforms representative SSI baselines in text prediction (WER 9.90%, CER 6.50%, PER 2.05%) and spectrogram reconstruction (SC 0.081, MSE 0.014). Ablation studies confirm the effectiveness of semantic encoding, adversarial learning, and phonetic enhancement. Extensive real-world experiments show Baro2Talk's robustness across user generalization, unseen commands, varying speaking rhythms, noise, mobility, and command lengths. Long-term evaluation shows consistent performance without retraining, validating Baro2Talk's practicality and generalizability.

II. Preliminaries

A. TMJ-dominated Pressure Variation Sequence

Definition 1. *A **TMJ-dominated Pressure Variation Sequence (TPVS)** refers to a sequence of barometric pressure fluctuations caused by articulatory movements during silent speech, even without vocalization, primarily driven by the temporomandibular joint, with contributions from other related components such as the tongue, jaw, and facial muscles.*

TPVSs vary across phrases and users: different phrases produce distinct deformation patterns, while repeated utterances of the same phrase by the same user yield highly consistent

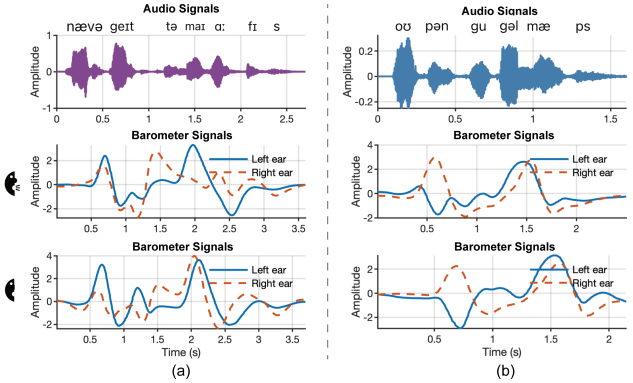


Fig. 3: **TPVS examples from two phrases.** Each row shows the audio (top) and TPVS under voiced (middle) and silent (bottom) conditions. The phrases are (a) *Navigate to my office* and (b) *Open Google Maps*.

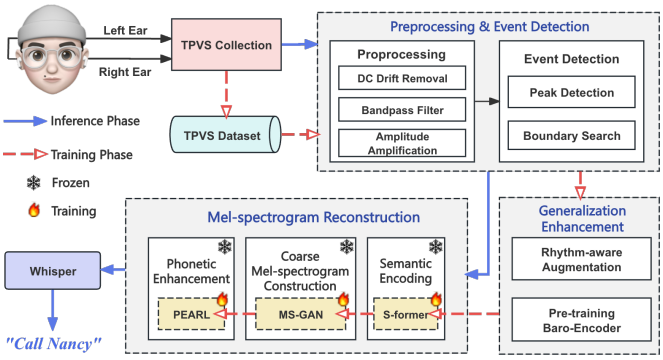


Fig. 4: System overview of Baro2Talk.

trajectories. This fine-grained articulatory encoding, despite lacking acoustic harmonics, underpins Baro2Talk's speech reconstruction. Prior studies [18], [19] reported canal volume changes of 10-25 mm^3 and diameter variations up to $2.5mm$, confirming feasibility of pressure-based articulatory sensing.

B. Pressure–Articulation Modeling

Unlike the idea $pv = k$ in Boyle's Law—where p, v, k represent pressure, volume and a thermodynamic constant—the ear canal's deformable and dynamic nature causes pressure to depend not only on v but also on tissue elasticity, sealing quality, and sensor coupling. Therefore it must be modeled as a nonlinear function of articulatory motion, as shown below:

$$p(t) = f(M(t); C) + n(t), \tag{1}$$

where $p(t)$ denotes the pressure value captured at time t; $M(t)$ denotes latent articulatory dynamics (e.g., jaw and tongue motion); $f(\cdot)$ is a nonlinear mapping modulated by a configuration parameter C; and $n(t)$ represents additive noise. C reflects user-specific characteristics such as articulation habits, ear canal geometry, and sealing quality.

C. Feasibility Study

We conduct a controlled study where participants articulate predefined commands under both voiced and silent conditions. Fig. 3 shows representative waveforms for two phrases: *"Navigate to my office"* and *"Open Google Maps"*, including audio (top), TPVSs from voiced speech (middle) and silent articulation (bottom).

The results show that TPVSs are consistent for repeated instances of the same phrase but vary significantly across different phrases and users. Moreover, TPVSs often exhibit temporal misalignment with the acoustic waveform—either leading or lagging—due to the inherent asynchrony between articulation and sound production. This misalignment makes phoneme- or word-level alignment ambiguous and error-prone, especially in the context of silent speech.

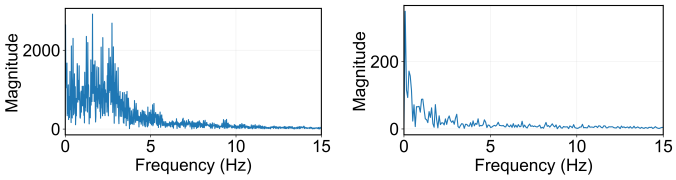
III. Baro2Talk Overview

As depicted in Fig. 4, Baro2Talk operates in two phases: (i) an offline training phase, which leverages both TPVS-text pairs and TPVS–Mel-spectrogram pairs derived from voiced speech to supervise different stages of the model; and (ii) an online inference phase that relies solely on TPVS to generate silent speech text output in real time. Baro2Talk consists of the following three key modules:

(1) Preprocessing and Event Detection. This module preprocesses the raw TPVSs through DC drift removal, band-pass filtering, and amplitude squaring to enhance articulatory deformation signals. A short-term energy function is then computed to amplify potential articulation events. Once energy peaks are identified, event boundaries are delineated based on the stability of local energy patterns.

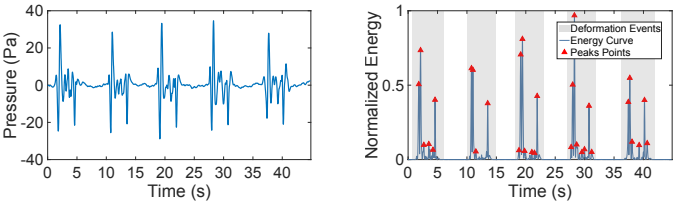
(2) Generalization Enhancement. To tackle the dual challenges arising from variations in speaking tempo and discrepancies in articulatory patterns due to individual differences in articulatory habits and ear canal structures (as discussed in Sec. II-B), this module incorporates two key strategies. First, we introduce a *Rhythm-aware Augmentation* technique that expands the training set by generating multiple tempo variants of the same TPVS sequence, thereby improving the model's robustness to temporal fluctuations. Second, we adopt the Domain Adversarial Neural Network (DANN) framework to pre-train a TPVS encoder—termed the *Baro-Encoder*—which learns speaker-invariant embeddings of TPVS signals corresponding to the same silent speech content.

(3) Mel-spectrogram Reconstruction. This module reconstructs high-fidelity Mel-spectrograms from TPVS inputs through a three-stage process. 1) *Semantic Encoding:* TPVSs are encoded into latent vectors via the Baro-Encoder, while transcripts are embedded by a Sentence Encoder. Both are projected into a shared space, where a contrastive loss aligns each TPVS embedding with its paired text while separating it from unrelated samples—bridging the modality gap without relying on frame-level alignment. 2) *Coarse Mel-spectrogram Construction:* uses a GAN-based module, MS-GAN, to convert latent vectors above into coarse Mel-spectrograms via a U-Net architecture with self-attention and adversarial training. The U-Net generator synthesizes spectrograms, while the discriminator ensures realism. 3) *Phonetic Enhancement:* introduces PEARL, a lightweight residual module is trained to predict high-frequency residuals of the coarse Mel-spectrogram. The refined output is the element-wise sum of the coarse spectrogram and the learned residual. Finally, the refined Mel-spectrograms are fed into a downstream ASR model, e.g., Whisper [14], which decodes them into textual transcriptions.



(a) Silent speech (b) Sensor noise

Fig. 5: Power spectral density of ear canal pressure signals.



(a) Band-pass filtered TPVS (b) Detected events

Fig. 6: Preprocessing and segmentation of TPVS.

IV. Preprocessing and Event Detection

A. Preprocessing

1) DC Drift Removal: To eliminate the low-frequency offset or baseline drift in the original TPVS p, we subtract the mean of the signal over a sliding window of size N , yielding the baseline-corrected signal:

$$\bar{p}(t) = p(t) - \frac{1}{N}\sum_{i=-\alpha}^{\alpha} p(t+i), \tag{2}$$

where $p(t)$ and $\bar{p}(t)$ represent the values at the t-th sampling point in sequence p before and after correction, respectively, and $\alpha = N/2$. This operation removes the direct current (DC) component and centers the signal around *zero*, mitigating long-term fluctuations that may interfere with downstream processing.

2) Band-pass Filter: To mitigate sensor-induced drift and motion-related artifacts, we apply a second-order Butterworth bandpass filter with a passband of 0.5–10 Hz. This range is chosen to retain weak yet informative low-frequency components (0.5–1 Hz), while effectively suppressing very slow drifts (below 0.5 Hz) and high-frequency disturbances typically caused by abrupt body movement. As illustrated in Fig. 5a, the power spectral density of silent speech is predominantly concentrated in the 1–6 Hz, with additional low-amplitude energy extending slightly below 1 Hz. In contrast, sensor noise (Fig. 5b) exhibits a strong spectral presence below 0.5 Hz, without any discernible structure. This validates the selection of the 0.5–10 Hz as a balanced choice for preserving speech-induced deformation signals while eliminating irrelevant noise.

3) Amplitude Amplification: To prepare the signal for reliable peak detection, we compute the point-wise squared magnitude of the preprocessed pressure waveform. This operation transforms all negative values into positive ones while preserving the relative intensity of deformation-related fluctuations. As a result, the signal becomes non-negative, making subsequent short-term energy calculation and peak detection more robust.

B. Event Detection

To identify silent articulation from continuous ear canal pressure signals, Baro2Talk employs a short-term energy-based detection strategy. As shown in Fig. 6a, the signal contains distinct high-energy segments corresponding to silent speech, while the rest remains relatively flat, indicating inactivity. The short-term energy $E(t)$ of the preprocessed signal $\bar{p}(t)$ is computed over a sliding window of M samples:

$$E(t) = \sum_{m=0}^{M-1} \bar{p}^2(t+m), \tag{3}$$

The resulting energy curve is normalized and smoothed to suppress noise and emphasize salient deformation peaks. Local maxima exceeding a predefined threshold are identified using a peak detection algorithm [20]. Around each peak, we extract a 20 ms window (M = 4), and locate the point of maximum energy, used as the anchor for boundary search.

To determine event boundaries, we expand bidirectionally from each anchor point until three consecutive local minima with amplitude differences below 3% are identified, indicating convergence to a stable baseline. Each detected segment is then temporally normalized via linear interpolation [21], yielding fixed-length inputs suitable for downstream processing. Fig. 6b illustrates the detected peaks (red triangles) and corresponding segmented regions (gray shading).

V. Generalization Enhancement

A. Rhythm-aware Augmentation

The speaking rhythm of silent speech varies significantly across utterances and users, which can negatively impact the model's generalization. To enhance generalizability, we design a rhythm transformation-based data augmentation strategy that leverages the paired audio-pressures data collected during voiced speech sessions.

1) Upsampling. Since the audio signal a_i has a much higher sampling rate than the corresponding pressure signal $\bar{p}_i$, we first upsample $\bar{p}_i$ to match the audio's sampling rate using linear interpolation for subsequent alignment, resulting $\hat{p}_i$.

2) Tempo Adjustment. To simulate diverse speaking speeds, we apply time-domain stretching and compression to a_i using a phase vocoder [22], which preserves spectral characteristics while altering the temporal rhythm. We generate multiple variants a_i^β by applying scaling factors $\beta \in [0.7, 1.3]$ with a step size of 0.1.

3) TPVS Synchronization. For each newly generated a_i^β, we apply the same stretching or compression to its corresponding TPVS segment $\hat{p}_i$ to obtain a new version of it, i.e., $\hat{p}_i^\beta$, forming an audio-pressure pair $\mathbb{S}_i^\beta = \{(a_i^\beta, \hat{p}_i^\beta)\}$. These augmented pairs serve as valuable training samples to improve rhythm invariance in downstream models.

B. Pre-training the Baro-Encoder

To extract user-invariant and semantically discriminative representations from TPVSs, we pre-train an encoder *Baro-Encoder* (as shown in Fig. 7), denoted as $G_\mathbf{f}$, using a Domain-Adversarial Neural Network (DANN) architecture [23]. The encoder is jointly trained with a command classifier G_y for semantic supervision and a user identifier G_i for domain

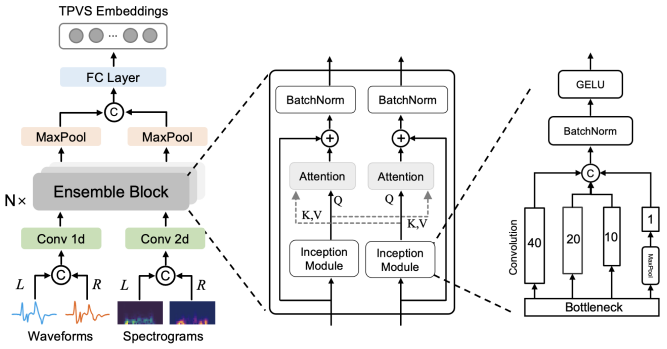


Fig. 7: The Architecture of *Baro-Encoder*.

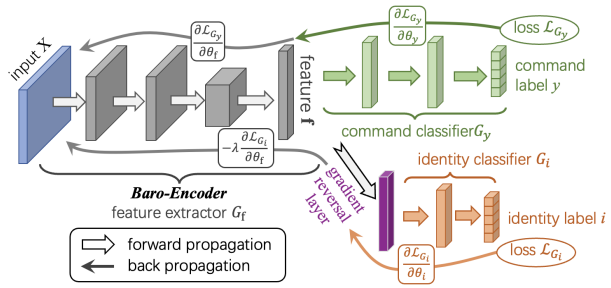


Fig. 8: Adversarial pre-training strategy for the *Baro-Encoder*.

regularization, as illustrated in Fig. 8. A **Gradient Reversal Layer (GRL)** is inserted before G_i, which acts as an identity function during the forward pass but multiplies the gradient by $-\lambda$ during backpropagation, thereby discouraging user-specific features.

During training, the input $\mathbf{X}$ of encoder $G_\mathbf{f}$ consists of two modalities: a pair of TPVSs (the outputs of V-A) from both ears and the corresponding Mel-spectrograms, denoted as $\{p_l, p_r\}$ and $\{\mathcal{S}(p_l), \mathcal{S}(p_r)\}$ (for notational simplicity, we omit hats and subscripts). l and r represent left and right ear, respectively, and $\mathcal{S}(\cdot)$ is the Mel-spectrogram. They are processed by parallel 1D/2D convolutional layers and fused through stacked Inception blocks with cross attention mechanism, yielding a latent representation $\mathbf{f} \in \mathbb{R}^d$.

The training objective jointly minimizes a command classification loss and an identity classification loss:

$$\mathcal{L} = \mathcal{L}_{G_y} + \lambda \mathcal{L}_{G_i}, \tag{4}$$

$$\mathcal{L}_{G_y} = \mathrm{CE}(G_y(G_\mathbf{f}(\mathbf{X})),\, y)\,, \tag{5}$$

$$\mathcal{L}_{G_i} = \mathrm{CE}(G_i(\mathrm{GRL}(G_\mathbf{f}(\mathbf{X}))),\, i)\,. \tag{6}$$

where y and i represent command and identity label, respectively. $\theta_\mathbf{f}$, $\theta_\mathbf{y}$, and $\theta_\mathbf{i}$ in the Fig. 8 represent model parameter of $G_\mathbf{f}$, $G_\mathbf{y}$, and $G_\mathbf{i}$, respectively. Since the GRL internally applies the negative scaling during backpropagation, so the training objective retains a positive weight λ on the adversarial loss. This adversarial mechanism ensures that the latent features $\mathbf{f} = G_\mathbf{f}(\mathbf{X})$ remain discriminative for command (i.e., coarse semantic) prediction while being invariant to user identity.

VI. Mel-spectrogram Reconstruction

Building upon the pre-trained $G_\mathbf{f}$, we design a three-stage reconstruction pipeline—comprising **Semantic Encoding**, **Coarse Mel-spectrogram Construction**, and **Phonetic Enhancement**—to generate Mel-spectrograms from the TPVS inputs $\mathbf{X}$ (see Fig. 9).

A. Semantic Encoding

We propose S-Former, a contrastive learning module designed to align TPVS embeddings with textual semantics, thereby bridging the gap between discrete classification tasks and the continuous nature of semantic space. As illustrated in Fig. 10, S-Former encourages each TPVS embedding $\mathbf{f}_i$ to be pulled closer to the corresponding textual embedding $\mathbf{e}_i$ in a shared latent space, while pushing it away from mismatched embeddings $\mathbf{e}_j$ where $j \neq i$. Specifically, $\mathbf{X}$ is encoded by the pre-trained encoder $G_\mathbf{f}$, and the corresponding ground-truth transcript is encoded by a *Sentence Encoder* (e.g., Sentence-BERT [24]). Both representations are projected into the same latent space to facilitate semantic alignment. Note that $G_\mathbf{f}$ is fine-tuned, while the Sentence Encoder is kept frozen.

Objective: We adopt a contrastive learning objective that encourages matched pairs to be close and mismatched pairs to be distant in the latent space. The loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{contrastive}} = \frac{1}{\mathcal{N}^2} \sum_{i=1}^{\mathcal{N}} \sum_{j=1}^{\mathcal{N}} \Big[& y_{ij} \cdot \log \sigma\left(\mathbf{f}_i^\top \mathbf{e}_j\right) \\ & + (1 - y_{ij}) \cdot \log\left(1 - \sigma\left(\mathbf{f}_i^\top \mathbf{e}_j\right)\right) \Big],\end{aligned} \tag{7}$$

where $\mathcal{N}$ is the number of training samples, $y_{ij} \in \{0, 1\}$ indicates whether the pair (i, j) is semantically aligned, and $\sigma(\cdot)$ denotes the sigmoid function. This objective encourages each TPVS embedding to closely match its paired text embedding while remaining distinguishable from unrelated ones, thus preserving semantic alignment across modalities.

B. Coarse Mel-spectrogram Construction

We introduce a generative module MS-GAN to transform the semantically aligned latent vector $\mathbf{f}$ (output of fine-tuned $G_\mathbf{f}$) into a coarse Mel-spectrogram $\hat{S}_\mathbf{f}$ that preserves global acoustic structure (As illustrated in Fig. 9). To ensure both semantic faithfulness and perceptual realism, MS-GAN is trained with a composite objective that combines adversarial and reconstruction losses.

Generator Objective. G_m is optimized using a hybrid loss:

$$\begin{aligned}\mathcal{L}_{G_m} = \; & \mathbb{E}\left[\|G_m(\mathbf{f}) - \mathrm{GT}_m\|^2\right] \\ & + \lambda\, \mathbb{E}_{\mathbf{f} \sim \mathbb{P}(\mathbf{f})}\left[\log(1 - D(G_m(\mathbf{f})))\right]\end{aligned} \tag{8}$$

where GT_m is the corresponding ground-truth Mel-spectrogram, $\mathbb{P}$ and D represent probability distribution and the discriminator, respectively. A balancing coefficient λ is used to control the trade-off between the two objectives, we empirically set $\lambda = 0.2$ in our experiments.

Discriminator Objective. The discriminator is trained to distinguish between real and generated spectrograms:

$$\mathcal{L}_\mathrm{D} = \mathbb{E}_{x \sim \mathbb{P}_\text{real}}[\log D(x)] + \mathbb{E}_{\mathbf{f} \sim \mathbb{P}(\mathbf{f})}[\log(1 - D(G_m(\mathbf{f})))] \tag{9}$$

This adversarial loss pushes D to improve its discriminative ability while guiding G_m toward more realistic synthesis.

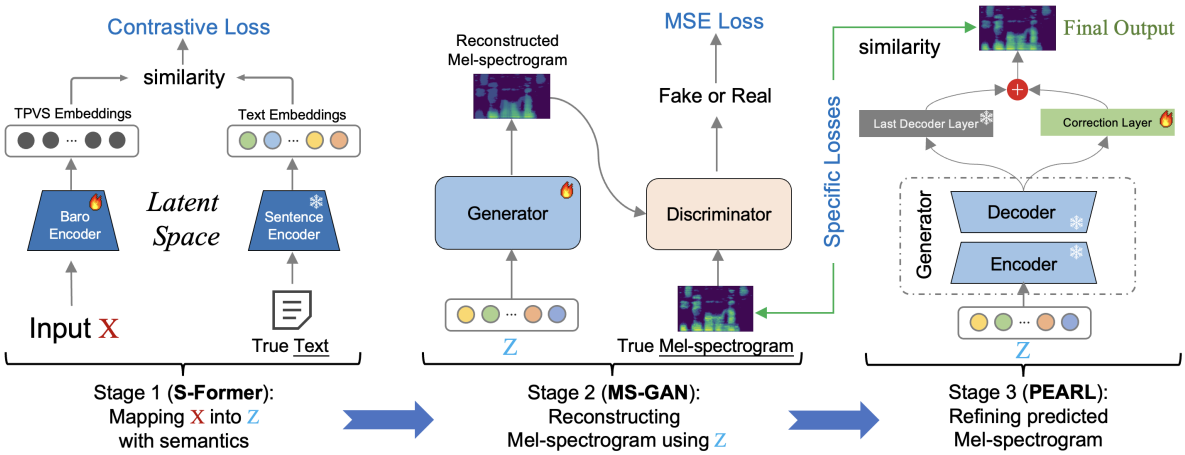


Fig. 9: **Overview of the three-stage Mel-spectrogram Reconstruction pipeline.** Stage 1 (**S-Former**) performs semantic encoding by aligning TPVS embeddings with text through contrastive learning. Stage 2 (**MS-GAN**) generates coarse Mel-spectrograms from semantic embeddings using a GAN framework. Stage 3 (**PEARL**) refines the initial spectrograms via residual correction guided by phonetic, spectral, and prosodic objectives.

Generator Architecture. G_m uses a U-Net-style encoder-decoder with multi-head self-attention. The encoder applies convolutional blocks with batch normalization, ReLU, and max-pooling for downsampling, while the decoder uses transposed convolutions for upsampling. Skip connections are disabled here. A multi-head attention module after the encoder captures global semantics before decoding.

Discriminator Architecture. D is a lightweight CNN that processes Mel-spectrogram patches through three convolutional layers with Leaky ReLU, followed by flattening and a fully connected layer for probability output. Optional pressure-derived features can be concatenated before the final layer to enhance discrimination.

C. *Phonetic Enhancement*

To refine phonetic details and enhance acoustic fidelity, we introduce a residual enhancement stage (As illustrated in Fig. 9). Specifically, we freeze G_m to retain its coarse structural modeling capability, and add a zero-initialized correction layer—architecturally identical to the final decoder layer—which receives the same input and learns to predict missing high-frequency components. This residual module, termed PEARL (Phonetic Enhancement via Adaptive Residual Learning), is optimized with a task-specific loss to capture fine-grained phonetic and prosodic cues. The final Mel-spectrogram is computed as:

$$\hat{\mathcal{S}}_{\mathbf{f}}^{*} = \hat{\mathcal{S}}_{\mathbf{f}} + \Delta\mathcal{S}_{\mathbf{f}}, \tag{10}$$

where $\hat{\mathcal{S}}_{\mathbf{f}}$ is the coarse spectrogram from Stage 2, and $\Delta\mathcal{S}_{\mathbf{f}}$ is the learned residual correction.

To supervise this refinement, we introduce the following three complementary losses, each targeting a distinct perceptual or structural dimension of the spectrogram.

a) Phoneme-Level Loss ($\mathcal{L}_{\mathrm{PL}}$): To enforce phonetic alignment between the generated spectrogram and the underlying text, we adopt the Connectionist Temporal Classification (CTC) loss to compare predicted phoneme sequences against the ground truth:

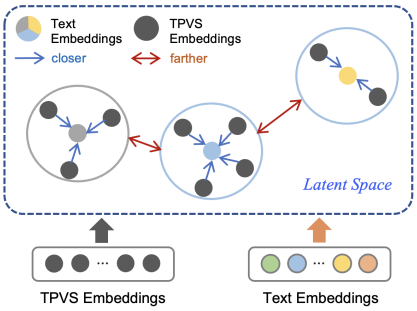


Fig. 10: Semantic encoding via contrastive learning.

$$\mathcal{L}_{\mathrm{PL}} = -\sum_{t=1}^{T} \log p(\pi_t \mid \mathbf{y}_t), \tag{11}$$

Here, π_t denotes the ground-truth phoneme at time t, and $\mathbf{y}_t$ is the predicted phoneme distribution. This loss enables sequence-level supervision without requiring explicit frame-wise alignment.

b) Spectral Convergence Loss ($\mathcal{L}_{\mathrm{SC}}$): To improve spectral structure, we apply a frequency-weighted mean squared error between the generated and target spectrograms:

$$\mathcal{L}_{\mathrm{SC}} = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} W(f) \left(\hat{S}_{\mathbf{f}}^{*}(f,t) - \mathrm{GT}_m(f,t) \right)^2, \tag{12}$$

where f and t represent the frequency and time dimensions, respectively. The weights $W(f)$ are assigned to each frequency bin to emphasize the importance of lower frequency ranges, which often contain more critical speech information [25].

c) Envelope Correlation Loss ($\mathcal{L}_{\mathrm{Corr}}$): To capture prosodic cues such as syllabic rhythm and amplitude dynamics, we introduce an envelope correlation loss computed from Hilbert-transformed waveforms reconstructed from the spectrogram:

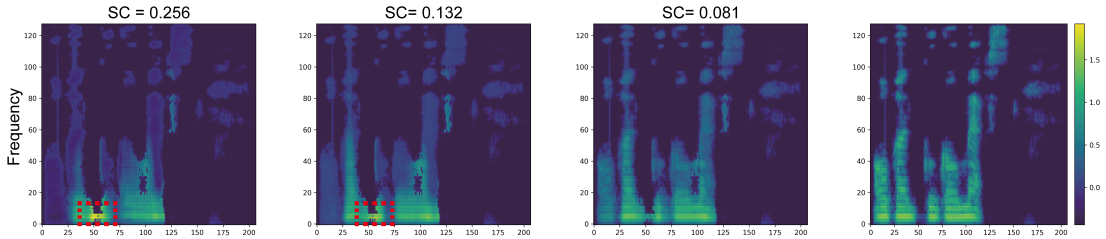


Fig. 11: Mel-spectrogram reconstruction with progressively enhanced training: (a) L1-only training yields blurred harmonics and noise (SC = 0.256); (b) MS-GAN improves sharpness and reduces artifacts (SC = 0.132); (c) Adding PEARL further enhances fidelity and continuity (SC = 0.081); (d) Ground-truth reference. Lower SC indicates better alignment.

$$\mathcal{L}_{\text{Corr}} = 1 - \frac{\sum_{t=1}^{T}(x(t)-\bar{x})(\hat{x}(t)-\bar{\hat{x}})}{\sqrt{\sum_{t=1}^{T}(x(t)-\bar{x})^2}\sqrt{\sum_{t=1}^{T}(\hat{x}(t)-\bar{\hat{x}})^2}}, \quad (13)$$

Here, $x(t)$ and $\hat{x}(t)$ denote the ground-truth and predicted temporal envelopes. Maximizing their correlation promotes accurate reproduction of prosodic structure.

To balance the loss terms, we normalize each component and set weights via grid search. The final loss, as a weighted sum, guides the residual module to enhance phonetic, spectral, and prosodic features. Fig. 11 illustrates improvements from each enhancement stage.

VII. Implementation and Setup

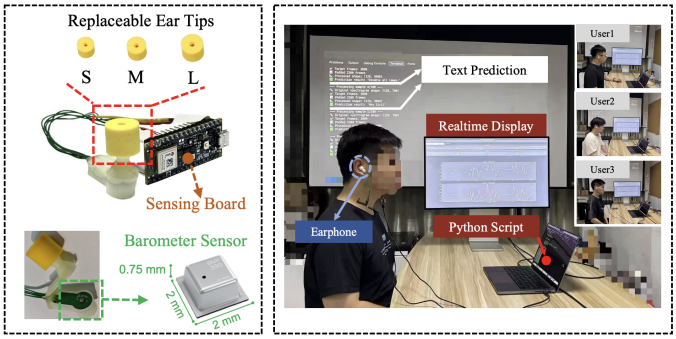
A. Experimental Setup

Hardware. As shown in Fig. 12a, Baro2Talk is implemented on *BaroBud*, our designed earbud prototype that embeds a miniature Bosch BMP390 barometric sensor (3 mm × 3 mm × 0.75 mm) inside a 3D-printed casing adapted from OpenEarable [26]. As illustrated in Fig. 12b, *BaroBud* is connected to an Arduino Nano 33 BLE board for real-time sampling at 108 Hz and transmits data via a serial interface (115,200 bps). A host PC (Intel i7-8750H CPU, 16 GB RAM, NVIDIA RTX 1060 GPU) collects synchronized in-ear pressure and ambient audio (44,100 Hz) using a Python interface, with software-layer timestamp-based synchronization. To ensure sealing and comfort, participants select from three sizes of foam ear tips (S/M/L).

Training. Baro2Talk adopts a three-stage training pipeline, each implemented in PyTorch and optimized with Adam (initial LR = 0.001, StepLR decay = 0.95 per epoch, batch size = 64). Dropout (0.1) and batch normalization are applied where appropriate. Stage 1 is trained for up to 30 epochs, while Stage 2 and Stage 3 run for up to 200 epochs. Models use 5-fold cross-validation and early stopping (patience = 10). For generalization evaluation, 20% of commands and 5 users are excluded during training and reserved for final testing.

B. Data Collection

We collect data from 25 participants (7 female, aged 21–50) using BaroBud while seated or standing with minimal head motion. Each participant selects 10 commands from our command list and articulates them in both *voiced* and *silent* modes. To introduce articulatory variability, commands are



(a) Earbud prototype (b) Experiment setup

Fig. 12: Experimental setting for Baro2Talk.

spoken at three speeds and two levels of mouth opening. The final 15 days of the campaign include mobile scenarios such as walking, stair climbing, and elevator rides to evaluate robustness under daily activities. We construct two datasets: 1) **Trace A:** Voiced & silent samples collected over 30 consecutive days under mixed conditions, totaling over 150,000 labeled instances with articulation metadata; 2) **Trace B:** Silent-only samples collected six months later across one week, used to evaluate long-term drift, contributing 17,500 additional samples. All sessions are video-recorded to support alignment and validation. During inference, only the in-ear pressure signal is used.

VIII. Evaluation

A. Baselines and Metrics

Baselines: We compare Baro2Talk with representative silent speech recognition systems that follow the widely adopted architecture of combining a CNN-based feature extractor with a sequential modeling component (e.g., BiLSTM [8], TCN [10], or Transformer [27]) trained using a CTC loss to enable alignment-free decoding. In addition, we include *Unvoiced* [13] as a reconstruction-based baseline that attempts to predict semantical text via generated Mel-spectrogram.

Metric: We evaluate performance on two tasks: 1) *Text Prediction*. We report **Word Error Rate (WER)**, **Character Error Rate (CER)**, and **Phoneme Error Rate (PER)**. WER and CER are computed as the normalized edit distance between predicted and reference sequences at the word and character levels, respectively. PER converts both sentences into phoneme sequences via a G2P model (CMUdict) and computes their

TABLE I: Overall Performance Comparison with Baselines.

Method	Text Prediction			Spectrogram Reconstruction	
	CER	WER ↓	PER	SC ↓	MSE ↓
TCN + CTC	14.72%	18.47%	14.21%	–	–
BiLSTM + CTC	12.11%	15.96%	11.32%	–	–
Transformer + CTC	12.65%	15.48%	12.13%	–	–
Unvoiced-like	10.71%	12.66%	8.11%	0.363	0.022
Baro2Talk	**6.50%**	**9.90%**	**2.05%**	**0.081**	**0.014**

TABLE II: Ablation Study on Key Components.

Setting	PT	SF	GAN	PL	CER	WER	PER	SC	MSE
Full Version	✓	✓	✓	✓	**6.50%**	**9.90%**	**2.05%**	**0.081**	**0.014**
w/o PEARL	✓	✓	✓	✗	8.15%	13.88%	8.03%	0.132	0.018
w/o MS-GAN	✓	✓	✗	✗	11.39%	20.31%	10.74%	0.256	0.139
w/o S-Former	✓	✗	✓	✓	9.22%	14.67%	8.94%	0.193	0.023
w/o S-Former (un)	✓	✗	✓	✓	15.62%	21.61%	11.83%	0.282	0.147
w/o GRL	✗	✓	✓	✓	8.92%	14.22%	4.53%	0.162	0.021
w/o Augmentation	✓	✓	✓	✓	8.20%	14.61%	3.47%	0.121	0.029

Levenshtein distance; 2) *Mel-spectrogram Reconstruction.* We evaluate Mel-spectrogram quality using **Mean Squared Error (MSE)** and **Spectral Convergence (SC)**. MSE quantifies the average point-wise squared error between predicted $\hat{\mathcal{S}}$ and reference $\mathcal{S}$ spectrograms. SC measures the relative spectral mismatch based on magnitude deviation:

$$\mathrm{SC} = \frac{\left\| |\hat{\mathcal{S}}| - |\mathcal{S}| \right\|_F}{\| |\mathcal{S}| \|_F}, \tag{14}$$

where $|\cdot|$ denotes magnitude and $\|\cdot\|_F$ is the Frobenius norm. While MSE captures absolute numerical accuracy, SC reflects structural alignment in the frequency domain.

B. Overall Performance

To comprehensively evaluate Baro2Talk, we re-implement representative SSI methods from the past five years, categorized into two groups: 1) CTC-based sequence-to-text models [8], [10], [27], and 2) spectrogram-to-text pipeline [13] using Whisper. Since none of the baselines are open-sourced, we reconstruct their architectures based on the original papers and retrain them on our dataset for fair comparison.

As shown in Table I, Baro2Talk significantly outperforms all baselines in both spectrogram reconstruction and text prediction. Compared with Unvoiced, our method reduces CER from 10.71% to 6.50%, WER from 12.66% to 9.90%, and PER from 8.11% to 2.05%.

Notably, CTC-based methods cannot generate intermediate features, limiting open-vocabulary support. While Unvoiced reconstructs spectrograms, it concatenates text embeddings before the *Decoder*, which is infeasible during inference without spoken input. In contrast, Baro2Talk uses S-Former to explicitly learn a semantic latent space, enabling inference without textual priors. For fair comparison, we use purely supervised training for spectrogram reconstruction, excluding adversarial loss. Despite this, Baro2Talk significantly improves SC (0.363 → 0.081) and MSE (0.022 → 0.014), validating our semantic and generative approach.

C. Ablation Study

To assess the contribution of each component of Baro2Talk, we conduct a systematic ablation study (As shown in Table II).

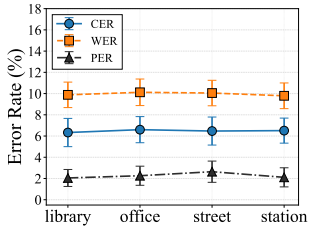


Fig. 13: Noisy conditions.

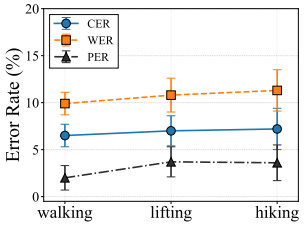


Fig. 14: Mobility.

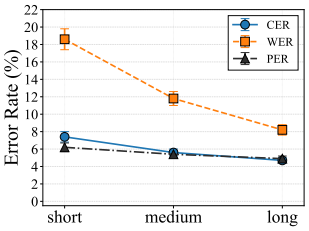


Fig. 15: Command length.

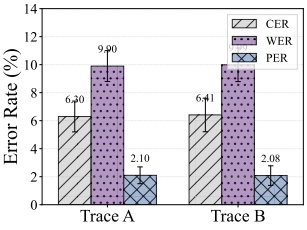


Fig. 16: Long term.

1) Rhythm-based Augmentation: To assess the contribution of rhythm-aware augmentation, we remove this module while keeping the rest of the pipeline intact. Performance declines notably without it (CER ↑ 6.50% → 8.20%, WER ↑ 9.90% → 14.61%, SC ↑ 0.081 → 0.121, MSE ↑ 0.014 → 0.029), demonstrating its role in improving temporal robustness and spectrogram reconstruction fidelity.

2) GRL: Disabling the GRL in DANN pre-training results in degraded generalization (CER ↑ 6.50% → 8.92%, WER ↑ 9.90% → 14.22%, PER ↑ 2.05% → 4.53%), highlighting the importance of adversarial training in promoting user-invariant representations of TPVSs during Baro-Encoder pre-training for consistent silent speech modeling.

3) S-Former: We assess the impact of semantic encoding through ablation on both seen and unseen commands. Removing S-Former (`w/o S-Former`) leads to notable degradation on seen inputs (CER ↑ 6.50% → 9.22%, PER ↑ 2.05% → 8.94%, SC ↑ 0.081 → 0.193), indicating that contrastive alignment enhances semantic representation beyond mere classification. On unseen commands (`w/o S-Former (un)`), performance drops further (CER ↑ 15.62%, PER ↑ 11.83%, SC ↑ 0.282), confirming that S-Former is critical for generalizing to open-vocabulary silent speech by embedding TPVSs into a structured semantic space.

4) MS-GAN: Removing MS-GAN leads to notable performance degradation (CER ↑ 6.50% → 11.39%, WER ↑ 9.90% → 20.31%, SC ↑ 0.081 → 0.256), underscoring its essential role in generating semantically coherent and acoustically realistic Mel-spectrograms. The combination of hybrid loss and adversarial training enhances the generator's ability to model global spectral structures, providing a strong foundation for subsequent phonetic refinement.

5) PEARL: Removing PEARL results in a marked decline in phonetic precision and structural coherence (PER ↑ 2.05% → 8.03%, SC ↑ 0.081 → 0.132), underscoring the role of residual refinement. As a lightweight correction module, PEARL learns high-frequency phonetic and prosodic details atop the coarse spectrogram without disrupting its global structure, thereby complementing MS-GAN and enhancing overall acoustic realism and expressiveness.

D. Robustness to Real-world Conditions

1) Noisy Environments: We evaluate Baro2Talk across four typical noisy environments: library, office, street, and railway station. As shown in Fig. 13, Baro2Talk maintains stable performance across these conditions, with CER, WER, and PER fluctuations remaining within 0.5%. This demonstrates the noise-resilient nature of our pressure-based modality, which is inherently immune to ambient sound interference.

2) Mobile Scenarios: We evaluate Baro2Talk under three mobility scenarios: walking, lifting, and hiking, each with varying vertical motion. As shown in Fig. 14, the system maintains stable performance throughout. Walking yields WER below 10%, with CER and PER under 7% and 3%. Lifting shows slight increases (WER 10.8%, PER 3.7%) due to effective drift removal. Hiking results in similar performance (WER 11.3%, CER 7.2%, PER 3.6%), demonstrating robustness to diverse mobility conditions and potential for handling more complex vertical dynamics.

3) Command Lengths: We evaluate how Baro2Talk performs across commands of different lengths, categorized as *short* ($\leq$3 words), *medium* (4–6 words), and *long* ($\geq$7 words). As shown in Fig. 15, the system yields lower error rates for longer commands, with WER decreasing from 19.7% (short) to 10.3% (long). This trend suggests that longer phrases provide richer semantic context and more stable articulatory patterns, enabling more reliable decoding. In contrast, shorter commands are more prone to ambiguity due to limited content and articulation.

4) Long Term: To assess the temporal stability of Baro2Talk, we evaluate its performance on *Trace B*, a silent-only dataset collected six months after *Trace A*. As shown in Fig. 16, Baro2Talk achieves consistently high performance over time, with only a marginal increase in CER (+0.11%) and WER (+0.09%), and a slight decrease in PER. These minimal variations indicate that Baro2Talk maintains robust performance over extended periods without retraining, underscoring its suitability for long-term, real-world deployment.

IX. Related Work

A. SSI Techniques

Vision-based Methods. Lip-interact [2], Lip-learner [3], and LipType [4] rely on smartphone cameras to capture lip dynamics under varying conditions. Although effective, video-based methods are sensitive to lighting and posture changes, and demand high-resolution hardware.

Wireless-based Methods. 1) *Acoustic:* SilentWhisper [27] utilizes built-in microphones to capture weak vocalizations. Ultrasonic sensing further enhances privacy by reconstructing speech from echo patterns of articulators. SVoice [11] reconstructs audible speech by analyzing the subtle disturbances of articulatory gestures on reflected ultrasound signals captured by smart phones. This requires the phone to be placed near mouth and the active signal modulation consumes energy. Sound-Lip [5], EchoWhisper [6], SilentTalk [7], EchoSpeech [28], and SottoVoce [29] employ ultrasonic transceivers embedded in mobile or wearable devices. EarCommand [8], EarSSR [9], and ReHEarSSE [10] embed ultrasonic modules in earphones to detect articulation-induced ear canal deformation. These solutions provide robustness and privacy but require continuous signal emission. 2) *mmWave:* WaveEar [30] and Msilent [31] sense throat and skin vibrations using millimeter-wave radar. Although precise, their front-facing deployment limits mobility.

Inertial-based Methods. IMU-based systems such as JawSense [32], Mutelt [12], and Unvoiced [13] embed sensors in earphones to monitor jaw motion. Derma [33] deploys multiple IMUs on the skin surface of the chin and upper neck to capture subtle skin deformations caused by articulatory muscle movements. However, these require careful attachment and may hinder comfort.

In contrast, our Baro2Talk system departs from active-sensing paradigms by leveraging passive barometric sensing embedded in a custom earbud prototype. Built upon a commercial-style form factor, our design enables low-power, privacy-preserving, and seamless silent speech interaction without the need for active signal emission or skin attachment.

B. Mel-Spectrogram Reconstruction Schemes

Recent efforts have investigated Mel-spectrogram reconstruction from non-acoustic physical signals to support speech-related tasks. Some works [15]–[17] leverage motion sensors to detect vibrations caused by audio playback. Others [34]–[37] utilize mmWave or RF reflections to capture sound-induced surface vibrations, while EM backscatter techniques [38], [39] monitor device oscillations. Although these signals do not directly record speech, they often retain pitch and harmonic patterns that implicitly encode semantic information, easing spectrogram reconstruction through frequency-domain regularities. In contrast, Baro2Talk reconstructs spectrograms from silent, TMJ-driven pressure signals, which lack harmonic structure and explicit linguistic content, making the task significantly more challenging and demanding robust modeling of articulatory dynamics.

X. Conclusion

We present Baro2Talk, an SSI that reconstructs articulatory representations from in-ear pressure signals without acoustic input. The design of Baro2Talk is inherently non-intrusive, privacy-friendly, and requires zero deployment effort. Moreover, ear canal pressure is minimally affected by motion or environmental changes, offering strong robustness to both. Extensive experiments validate the effectiveness of Baro2Talk across varied conditions.Baro2Talk indicates high phonetic fidelity in reconstructed speech, laying a foundation for future integration with LLM-based post-processing to further improve semantic restoration in challenging scenarios.

Acknowledgment

This research was supported in part by the National Natural Science Foundation of China (Grant No. 62472083, 62432008), the AI-Enhanced Research Program of Shanghai Municipal Education Commission (Grant No. SMEC-AI-DHUZ-01), the Fundamental Research Funds for the Central Universities (Grant No. CUSF-DH-D-2025031), and the Shanghai Science and Technology Plan Project (Grant No. 25DX1400200).

REFERENCES

[1] Coherent Market Insights, "Voice recognition market size and share analysis – growth trends and forecasts (2025-2032)." https://www.coherentmarketinsights.com/industry-reports/voice-recognition-market, 2025. Accessed: Jul. 27, 2025.

[2] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-interact: Improving mobile device interaction with silent speech commands," in *Proceedings of the 31st annual ACM symposium on user interface software and technology*, pp. 581–593, 2018.

[3] Z. Su, S. Fang, and J. Rekimoto, "Liplearner: Customizable silent speech interactions on mobile devices," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–21, 2023.

[4] L. Pandey and A. S. Arif, "Liptype: a silent speech recognizer augmented with an independent repair model," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19, 2021.

[5] Q. Zhang, D. Wang, R. Zhao, and Y. Yu, "Soundlip: Enabling word and sentence-level lip interaction for smart devices," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 5, no. 1, pp. 1–28, 2021.

[6] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin, "Echowhisper: Exploring an acoustic-based silent speech interface for smartphone users," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 4, no. 3, pp. 1–27, 2020.

[7] J. Tan, C.-T. Nguyen, and X. Wang, "Silenttalk: Lip reading through ultrasonic sensing on mobile phones," in *IEEE INFOCOM 2017-IEEE Conference on Computer Communications*, pp. 1–9, IEEE, 2017.

[8] 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," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 2, pp. 1–28, 2022.

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

[10] X. Dong, Y. Chen, Y. Nishiyama, K. Sezaki, Y. Wang, K. Christofferson, and A. Mariakakis, "Rehearsse: Recognizing hidden-in-the-ear silently spelled expressions," in *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024.

[11] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang, "Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, pp. 622–636, 2022.

[12] T. Srivastava, P. Khanna, S. Pan, P. Nguyen, and S. Jain, "Muteit: Jaw motion based unvoiced command recognition using earable," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1–26, 2022.

[13] T. Srivastava, P. Khanna, S. Pan, P. Nguyen, and S. Jain, "Unvoiced: Designing an llm-assisted unvoiced user interface using earables," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, pp. 784–798, 2024.

[14] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*, pp. 28492–28518, PMLR, 2023.

[15] P. Hu, H. Zhuang, P. S. Santhalingam, R. Spolaor, P. Pathak, G. Zhang, and X. Cheng, "Accear: Accelerometer acoustic eavesdropping with unconstrained vocabulary," in *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1757–1773, IEEE, 2022.

[16] S. Zhang, Y. Liu, and M. Gowda, "I spy you: Eavesdropping continuous speech on smartphones via motion sensors," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 4, pp. 1–31, 2023.

[17] D. Cayir, R. Mohamed, R. Lazzeretti, M. Angelini, A. Acar, M. Conti, Z. B. Celik, and S. Uluagac, "Speak up, i'm listening: Extracting speech from zero-permission vr sensors," in *NDSS*, 2025.

[18] C. Pirzanski and B. Berge, "Ear canal dynamics: Facts versus perception," *The Hearing Journal*, vol. 58, no. 10, pp. 50–52, 2005.

[19] M. J. Grenness, J. Osborn, and W. L. Weller, "Mapping ear canal movement using area-based surface matching," *The Journal of the Acoustical Society of America*, vol. 111, no. 2, pp. 960–971, 2002.

[20] F. Scholkmann, J. Boss, and M. Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, 2012.

[21] L. Erup, F. M. Gardner, and R. A. Harris, "Interpolation in digital modems. ii. implementation and performance," *IEEE Transactions on communications*, vol. 41, no. 6, pp. 998–1008, 2002.

[22] J. L. Flanagan and R. M. Golden, "Phase vocoder," *Bell system technical Journal*, vol. 45, no. 9, pp. 1493–1509, 1966.

[23] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*, pp. 1180–1189, PMLR, 2015.

[24] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[25] S. Rosen, "Temporal information in speech: acoustic, auditory and linguistic aspects," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 336, no. 1278, pp. 367–373, 1992.

[26] T. Röddiger, T. King, D. R. Roodt, C. Clarke, and M. Beigl, "Openearable: Open hardware earable sensing platform," in *ACM International Symposium on Wearable Computers*, 2022.

[27] H. Hiraki and J. Rekimoto, "Silentwhisper: inaudible faint whisper speech input for silent speech interaction," in *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–6, 2025.

[28] 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 *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–18, 2023.

[29] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proceedings of the 2019 CHI conference on human factors in computing systems*, pp. 1–11, 2019.

[30] C. Xu, Z. Li, H. Zhang, A. S. Rathore, H. Li, C. Song, K. Wang, and W. Xu, "Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface," in *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services*, pp. 14–26, 2019.

[31] S. Zeng, H. Wan, S. Shi, and W. Wang, "msilent: Towards general corpus silent speech recognition using cots mmwave radar," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 1, pp. 1–28, 2023.

[32] P. Khanna, T. Srivastava, S. Pan, S. Jain, and P. Nguyen, "Jawsense: recognizing unvoiced sound using a low-cost ear-worn system," in *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*, pp. 44–49, 2021.

[33] J. Rekimoto and Y. Nishimura, "Derma: silent speech interaction using transcutaneous motion sensing," in *Proceedings of the Augmented Humans International Conference 2021*, pp. 91–100, 2021.

[34] Y. Feng, K. Zhang, C. Wang, L. Xie, J. Ning, and S. Chen, "mmeavesdropper: Signal augmentation-based directional eavesdropping with mmwave radar," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*, pp. 1–10, IEEE, 2023.

[35] P. Hu, W. Li, Y. Ma, P. S. Santhalingam, P. Pathak, H. Li, H. Zhang, G. Zhang, X. Cheng, and P. Mohapatra, "Towards unconstrained vocabulary eavesdropping with mmwave radar using gan," *IEEE Transactions on Mobile Computing*, vol. 23, no. 1, pp. 941–954, 2022.

[36] C. Wang, F. Lin, H. Yan, T. Wu, W. Xu, and K. Ren, "{VibSpeech}: Exploring practical wideband eavesdropping via bandlimited signal of vibration-based side channel," in *33rd USENIX security symposium (USENIX Security 24)*, pp. 3997–4014, 2024.

[37] Y. Chen, J. Yu, L. Kong, H. Kong, Y. Zhu, and Y.-C. Chen, "Rf-mic: Live voice eavesdropping via capturing subtle facial speech dynamics leveraging rfid," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 7, no. 2, pp. 1–25, 2023.

[38] Y. Chen, J. Yu, Y. Chen, L. Kong, Y. Zhu, and Y.-C. Chen, "Rfspy: Eavesdropping on online conversations with out-of-vocabulary words by sensing metal coil vibration of headsets leveraging rfid," in *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, pp. 169–182, 2024.

[39] G. Wang, Z. Shi, Y. Yang, Z. An, G. Zhang, P. Hu, X. Cheng, and J. Cao, "Wireless eavesdropping on wired audio with radio-frequency retroreflector attack," *IEEE Transactions on Mobile Computing*, 2024.