

# Poster: Recognizing Hidden-in-the-Ear Private Key for Reliable Silent Speech Interface Using Multi-Task Learning

Xuefu Dong  
Liqiang Xu  
The University of Tokyo  
Tokyo, Japan

Lixing He  
The Chinese University of  
Hong Kong  
Hong Kong SAR, China

Zengyi Han\*  
Dalian Maritime University  
Dalian, China  
zyhan@dlnu.edu.cn

Ken Christofferson  
University of Toronto  
Toronto, Ontario, Canada

Yifei Chen  
Tsinghua University  
Beijing, China

Akihito Taya  
The University of Tokyo  
Tokyo, Japan

Yuuki Nishiyama  
The University of Tokyo  
Chiba, Japan

Kaoru Sezaki  
The University of Tokyo  
Chiba, Japan

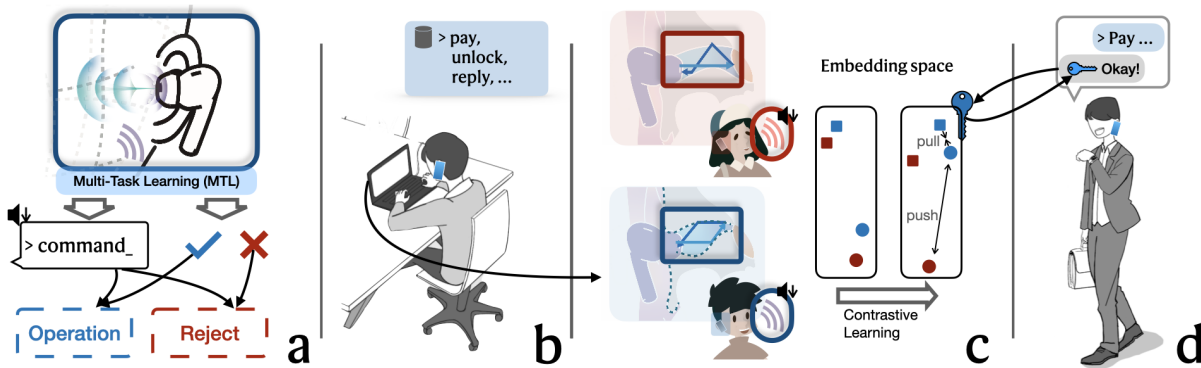


Figure 1: (a) HEar-ID uses multi-task learning (MTL) to authenticate user identity and reliably infer silent speech; (b) a new user registers in HEar-ID by recording few shots of silent commands; (c) HEar-ID leverages contrastive learning to pull cosine similarity between projected features of whisper and ECDM; (d) the user can be verified in a natural and safe style.

## Abstract

Silent speech interface (SSI) enables hands-free input without audible vocalization, but most SSI systems do not verify speaker identity. We present HEar-ID, which uses consumer active noise-canceling earbuds to capture low-frequency “whisper” audio and high-frequency ultrasonic reflections. Features from both streams pass through a shared encoder, producing embeddings that feed a contrastive branch for user authentication and an SSI head for silent spelling recognition. This design supports decoding of 50 words while reliably rejecting impostors, all on commodity earbuds with a single model. Experiments demonstrate that HEar-ID achieves strong spelling accuracy and robust authentication.

## CCS Concepts

• Security and privacy → Biometrics; • Human-centered computing → Text input; Sound-based input / output.

## Keywords

silent speech interface, user authentication, contrastive learning, multi-task learning, earable computing, acoustic sensing

## ACM Reference Format:

Xuefu Dong, Liqiang Xu, Lixing He, Zengyi Han, Ken Christofferson, Yifei Chen, Akihito Taya, Yuuki Nishiyama, and Kaoru Sezaki. 2025. Poster: Recognizing Hidden-in-the-Ear Private Key for Reliable Silent Speech Interface Using Multi-Task Learning. In *Companion of the 2025 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp Companion '25)*, October 12–16, 2025, Espoo, Finland. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3714394.3754429>

## 1 Introduction

Although speech-based interactions offer a natural alternative to manual or visual controls, they are not always feasible in public settings: speaking aloud in quiet environments (such as a library) may disturb others or expose private information (e.g., dictating a text message). Silent speech interface (SSI) addresses these issues by relying on mouth and facial movements rather than audible vocalization, allowing use in both noisy and quiet spaces without disrupting bystanders. Consequently, users tend to prefer SSI over traditional speech recognition in public contexts [26]. However, prior SSI-related works [4, 30, 37, 39] hardly consider the safety issues. While normal voice-based authentication mitigates these risks by verifying the speaker’s identity before granting access, it can be vulnerable to replay and injection attacks (e.g., triggering Siri via a loudspeaker), leaving an imperative need for developing a reliable SSI system.

\*Zengyi Han is the contact author.

Posted for personal use only; not for redistribution. The definitive version will appear in *UbiComp Companion '25*. <https://doi.org/10.1145/3714394.3754429>

We analyzed and found that the silent speech recognition task and the speaker authentication task are correlated rather than independent. The inherent structure uniqueness of each individual's ear canal creates distinct acoustic propagation paths, so that subtle ear canal deformations encode both the utterance content and speaker identity. Fortunately, recent years have witnessed an essential boost in earable sensing ranging from activity recognition [17, 23, 28], pose or mesh reconstruction [6, 10, 21, 25], speech enhancement [16, 18, 19], and user authentication [7, 8, 11, 31]. Among them, several works use commodity earbuds with an in-ear microphone to detect ear canal dynamic motion (ECDM) using ultrasonic sensing, where the ear canal serves as the reflection chamber. Therefore, we employ a multi-task learning (MTL) approach upon the earable platform, leveraging this correlation to perform silent spelling recognition and speaker authentication simultaneously.

In this work, we enable reliable SSI by proposing HEar-ID, which only leverages a commodity active noise-canceling earbud to emit an inaudible OFDM signal and record both ultrasonic reflections and whisper audio to enable silent spelling input (e.g. /i: eɪ ər/) for the word "ear") and user verification with a single machine learning model. Here we treat the silent speech as equivalent to whispering, since we found most experiment participants still make a subtle voice due to a lack of training. As shown in Figure 1, HEar-ID extracts features from ultrasonic waveforms (17.5–23kHz) and whispers (0–11kHz), then feeds the twin features through a shared encoder (TCN→Bi-GRU→MLP) to produce  $d$ -dimensional embeddings. A contrastive head aligns genuine-user whisper-ultrasonic pairs and repels attacker samples, while an SSI head on concatenated embeddings enables word-level decoding of silent spellings. In the preliminary experiments, HEar-ID consistently delivered promising results: for 11 participants, the system reliably rejected impostors with a false positive rate (FPR) of 3.2% and a true positive rate (TPR), accompanying 90.25% Top-1 word recognition accuracy for eight of them.

Our contributions are: (1) a multi-task SSI framework that fuses ultrasonic features with whisper ones for joint spelling and authentication; (2) a contrastive learning formulation (CLWUM) that embeds whisper and ultrasonic signals into a “private-key” space for robust verification; (3) an evaluation demonstrating accuracy spelling on 50 words and strong authentication.

## 2 Related Work

### 2.1 Ultrasonic Silent Speech Interfaces

Ultrasonic sensing interpreted by doppler effect [36] or other machine learning techniques [2, 3, 24] detects movements similar to those captured by radio-frequency- [22, 35], optics- [29, 32] and motion-based sensors [14, 15] to enable silent speech interface (SSI) techniques. Wearable approaches include Jin et al. [20]'s EarCommand which uses a custom earbud emitting an FMCW chirp, with a microphone in the ear canal recording reflections, achieving 89.9% accuracy on 32 silently spoken words across 12 participants.

### 2.2 Earable-based User Authentication

Wearable authentication can use passwords—e.g., voice or gesture-based schemes on earbuds [1, 33]—but spoken passwords risk leaking personal data and are vulnerable to replay attacks, motivating biometric alternatives.

However, almost all existing systems require normal speech or detectable motion, limiting usability in public or quiet settings. To our knowledge, HEar-ID is the first earable solution to authenticate via silent or whispered speech. Unlike EarDynamic [31], which relies solely on ultrasonic reflections, we learn a joint mapping between ultrasonic and whisper signals to capture user identity even when no audible speech is present.

## 3 Contrastive Multi-Task Learning (CMTL)

In this section, we describe HEar-ID's approach for word inference and user verification, as illustrated in Figure 2.

### 3.1 Signal Processing and Feature Extraction

HEar-ID leverages an OFDM signal of 2046 samples over a sampling rate of 48000Hz, lasting for about 42ms. Upon receiving the reflection from the ear canal, HEar-ID first isolates the *ultrasonic band* (17.5–23 kHz) with a high-pass filter, then aligns each received segment to an exact OFDM cycle. A coarse alignment finds the lag  $t_{coarse}$  that maximizes cross-correlation between one transmitted symbol and two received symbols. A fine alignment then corrects sub-sample delays ( $\pm 25$  samples) by minimizing phase differences with the transmitted signal. The original audio is aligned using the result index.

We extract features from the data after slicing it into 426-ms (10 frames) sliding windows with a stride of 85ms (2 frames). The *ultrasonic band* of each window, containing information of ear canal dynamic motions (ECDM), is further boosted by a second-order differentiation for compensation of rapid attenuation above 16 kHz in commercial earbuds, and later represented by 200-lag autoregressive (AR) coefficients. On the other hand, the whispering audio is procured by applying another 12kHz lowpass filter to the *aligned original audio*. To reduce computational complexity, we downsample the audio by a factor of 2. Then we fed it into a widely adopted mel-spectrogram extractor ( $n\_fft = hop\_length \approx 42ms$ ). The paired features are then fed into our neural network.

### 3.2 Contrastive Learning for Whisper-Ultrasonic Mapping (CLWUM)

When a user articulates silently, facial movements induce correlated perturbations in both the low-frequency whisper signal (captured by an earbud's microphone) and the high-frequency ultrasonic reflections inside the ear canal. CLWUM exploits this intrinsic link by mapping paired whisper and ultrasonic signals into a shared embedding space, effectively acting as a “private key” that aligns only genuine-user embeddings.

As illustrated in Figure 2, each batch provides  $2N$  segments per modality: the first  $N$  from the *genuine* user and the next  $N$  from *attackers* ( $N$  is set to 50 words in experiments). After modality-specific encoders  $f_w(\cdot)$  (whisper) and  $f_u(\cdot)$  (ultrasonic) and small

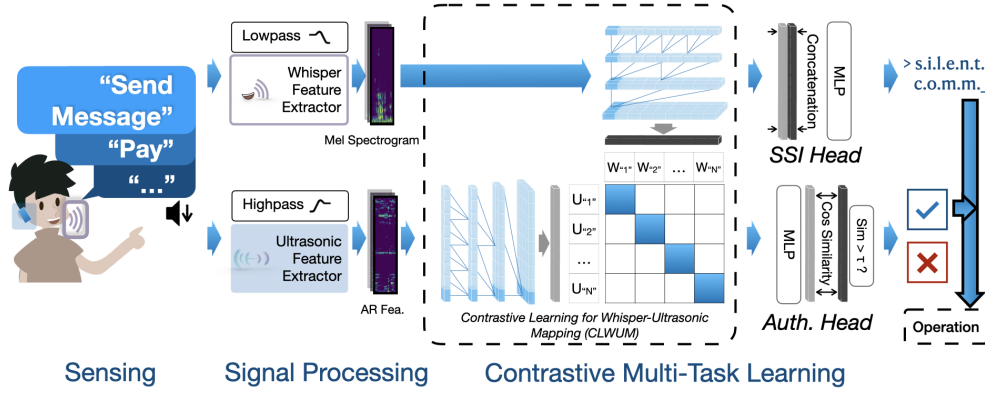


Figure 2: The three-fold workflow of HEar-ID.

projection heads  $h_w(\cdot)$ ,  $h_u(\cdot)$ , we obtain embeddings

$$w_i = h_w(f_w(x_i^{\text{whisper}})), \quad u_j = h_u(f_u(x_j^{\text{ultra}})), \quad i, j = 1, \dots, 2N,$$

which we collect into  $W(d \times 2N)$ , and  $U(d \times 2N)$ . Since CLWUM focuses on the genuine-user mapping, we extract only the first  $N$  embeddings from the genuine user:  $W^g(d \times N)$ ,  $U^g(d \times N)$ . We then compute the cosine-similarity array between whisper and ultrasonic embeddings from  $N$  genuine user samples  $S_{ij} = \text{sim}(w_i, u_j)$ . Here, the diagonal entries  $S_{ii}$  represent true whisper-ultrasonic pairs. Contrastive learning fosters a robust cross-modal alignment between ultrasonic and whisper embeddings for silent spelling recognition, while simultaneously extracting the shared features that underpin reliable user authentication. Accordingly, we define the contrastive loss:

$$L_{\text{CL}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S_{i,i}/\tau)}{\sum_{j=1}^N \exp(S_{i,j}/\tau)},$$

where  $\tau > 0$  is a temperature hyperparameter and is set to 0.7 based on empirical experience. Minimizing  $L_{\text{CL}}$  maximizes similarity of genuine pairs and minimizes similarity of mismatched pairs, embedding whisper and ultrasonic signals from the genuine user into a coherent shared space.

### 3.3 MTL-based Reliable Silent Spelling Interface

To simultaneously verify user identity and recognize spelled words in silent speech, we adopt a multi-task learning (MTL) paradigm. Multi-task learning aims to enhance the overall performance on multiple tasks by leveraging shared representation and useful information across related tasks [38]. Our architecture branches into two tasks—User Authentication and Silent Spelling Recognition—sharing the CLWUM encoders as a foundation.

**3.3.1 User Authentication.** As shown in Figure 2, each whisper or ultrasonic segment first passes through the CLWUM encoder—three stacked TCN blocks followed by two Bi-GRU layers—and then through an embedding MLP to produce a  $d$ -dimensional representation. This shared embedding is further projected by lightweight modality-specific heads to yield authentication vectors. For each batch, let  $h_i^{\text{whisper}}$  and  $h_i^{\text{ultra}}$  be the shared  $d$ -dimensional embeddings (after CLWUM encoder and embedding MLP) corresponding

to the  $i$ -th genuine-user silent spelling, and let  $h_j^{\text{ultra}}$  be the embedding for the  $j$ -th attacker segment (where  $i, j = 1, \dots, N$  and there are  $N$  unique words per batch). We then compute the authentication vectors by applying two separate MLP for ultrasonic and whisper data in the authentication head. The system then calculates the cosine similarity of the pair, and makes the decision of authentication if similarity is greater than a threshold  $thr$ . During training, we form pairs in dataset from genuine-user and attackers:

$$w_i = \text{MLP}_w(h_i^{\text{whisper}}), \quad u_i^+ = \text{MLP}_u(h_i^{\text{ultra}}), \quad u_j^- = \text{MLP}_u(h_j^{\text{ultra}}),$$

where  $w_i, u_i^+, u_j^- \in \mathbb{R}^d$ . Here,  $(w_i, u_i^+)$  form genuine-user pairs (both embedding the same word), while each  $u_j^-$  represents an attacker embedding for that same word.

Then, we compute cosine similarities  $s_i^+ = \text{sim}(w_i, u_i^+)$  measures similarity for a genuine pair, and  $s_i^- = \text{sim}(w_i, u_i^-)$  for a genuine-attacker pair. Training minimizes the angular triplet loss with margin  $m$  (where  $m$  is set to 11.45):

$$L_{\text{auth}} = \frac{1}{N} \sum_{i=1}^N \max\left(0, m + \arccos(s_i^+) - \arccos(s_i^-)\right).$$

This loss pulls the angle between genuine pairs smaller (higher similarity) while pushing the hardest negative (attacker) angle larger, yielding a clear decision boundary. We calibrate the  $thr$  with Youden's  $J$  statistics [34], and adopt the empirical value of  $m$  as 30 degrees.

**3.3.2 Silent Spelling Recognition.** As depicted in Figure 2, for silent spelling the genuine-user whisper and ultrasonic segments are first processed by the shared CLWUM encoder (TCN → Bi-GRU → embedding MLP), resulting in two sequences of  $d$ -dimensional vectors:  $h^{\text{whisper}} = [h_1^w, \dots, h_N^w]$ ,  $h^{\text{ultra}} = [h_1^u, \dots, h_N^u]$ . We then concatenate corresponding whisper and ultrasonic embeddings for each time step to  $c_i = [h_i^w \parallel h_i^u]$  and feed the concatenated sequence  $\{c_i\}$  into a small projection MLP followed by a softmax classifier to produce frame-level logits over the alphabet (26 letters + blank):  $\ell_i = \text{MLP}_{\text{spell}}(c_i)$ , where  $K = 27$ . Denote the entire logit sequence by  $\mathbf{L} = [\ell_1, \dots, \ell_N]$ .

We train with Connectionist Temporal Classification (CTC) loss [13] to align  $\mathbf{L}$  with the target letter sequence  $\mathbf{y}$  without frame-wise labels:  $L_{\text{CTC}} = -\log P(\mathbf{y} | \mathbf{L})$ . CTC allows the model to learn

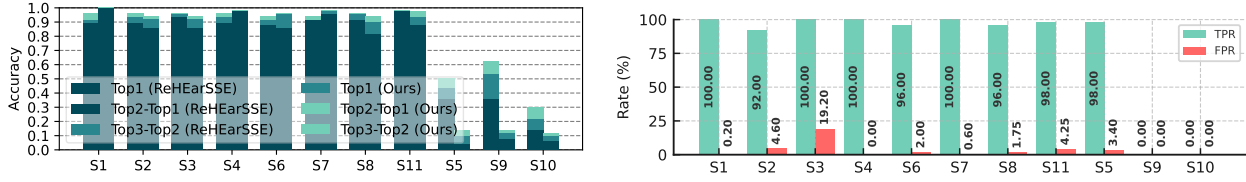


Figure 3: (a) Word inference accuracy, and (b) user authentication performance.

both letter identities and their temporal transitions implicitly, which is crucial for decoding silently mouthed words.

### 3.4 Authentication Pipeline

The authentication pipeline is divided into two phases:

**Registration Phase.** Each user begins by silently spelling a small set of word samples while wearing the earbuds. HEar-ID then trains the network using whisper-ultrasonic pair of samples with other individuals' data. We claim that these extra training data can be provided by the manufacturer along with the product. HEar-ID is trained end-to-end by minimizing the combined loss

$$L_{\text{total}} = \alpha L_{\text{CL}} + \beta L_{\text{auth}} + \gamma L_{\text{CTC}},$$

with task weights set to  $\alpha = 0.1$ ,  $\beta = 0.5$ , and  $\gamma = 0.3$  after hyperparameter tuning. A verification threshold is then calibrated.

**Verification Phase.** When a user silently spells a word, the system captures the corresponding whisper and ultrasonic segments, computes their embeddings, and compares these against the stored reference set. If the similarity to at least one reference embedding exceeds the threshold, the user is accepted; otherwise, access is denied. At the same time, the concatenated embeddings are sent to the head, which produces logits that are decoded (e.g., via beam search) to yield the final spelled word.

## 4 Preliminary Results

We used a subset of the dataset from ReHEarSSE [5] (10 participants) and recruited one additional participant, bringing our study to a total of 11 users (9 males, 2 females; mean age 22). All data is collected by off-the-shelf active noise-cancelling earbuds (Edifier W380NB) where we removed the in-ear mic from the earbud circuitry and wired it directly to a smartphone (Google Pixel 3a) via a 3.5 mm jack, ensuring unprocessed raw audio. All data collection procedures were conducted in accordance with Institutional Review Board (IRB) guidelines. All participants spelled 4 rounds of the same lexicon of 50 randomly selected words from the 1000 most used words from the Oxford English Dictionary [27] with subtle voice (e.g. /i: ei ar/) for the word "ear").

### 4.1 Overall Performance

**4.1.1 Experiment Design:** We take turns to regard one participant as the genuine user and others as impostors. For each participant, we form the testing set by randomly choosing a re-wearing session, and the training set by all other sessions except the left-out session. The authentication head and the CLWUM are trained and tested on all participants' data. That means each participant's model is attacked 500 times and accessed by the genuine user 50 times during testing. For the SSI head, we only use the genuine user's data,

resulting in a user-dependent model. We implement the pipeline of [5] serving as a baseline.

**4.1.2 Metrics.** For recognition, we report *Top-n word accuracy* ( $n=1, 2, 3$ ), i.e. the percentage where the true label appears in the top- $n$  predictions. For authentication, we adopt the standard **TPR** (True Positive Rate), **FPR** (False Positive Rate) and the resulting **EER**. Figure 3 illustrates the recognition accuracy of our **whisper+ultrasonic** Silent-Speech Interface (SSI) with the ultrasonic-only baseline ReHEarSSE [5], while Figure 3 summarises per-user authentication results (TPR/FPR).

**Silent-speech recognition:** Our *whisper + ultrasonic* SSI attains a mean Top-1 accuracy of **67.3 %**. However, we found that for 8 of 11 users without degraded performance in ultrasonic sensing, we achieved 90.25% Top-1 accuracy, which is comparable to the ultrasonic-only ReHEarSSE baseline (**91.4 %**). Three speakers (S1, S4, S7) exceed 90 %, confirming the benefit of the whisper stream. S5, S9, and S10 lag (<10 %)—likely owing to unclear articulation and idiosyncratic sensor placement, issues that the whisper-ultrasonic alignment amplifies.

**Authentication:** Average performance is **TPR = 81.76 %**, **FPR = 3.2 %**. Notably, S5 achieves **99.9 % / 3.4 %** despite poor recognition, illustrating that our identity embeddings remain robust when lexical decoding falters. Conversely, S9 and S10 record near-zero TPRs, mirroring their low recognition scores and indicating inconsistent headset placement across sessions.

Overall, the results confirm that coupling a whisper microphone with ultrasonic sensing brings recognition accuracy on par with state-of-the-art acoustic SSI while preserving high authentication robustness—meeting the dual goals of content retrieval and user verification in a single wearable and machine learning model.

## 5 Conclusion

We introduced HEar-ID, which combines AR ultrasonic and whisper features to enable silent spelling and user authentication on earbuds. A CLWUM module aligns genuine-user embeddings and repels impostors to simultaneously support two tasks. Tests show robust authentication under real-world conditions and promising spelling accuracy of 90.25% over 8 out of 11 participants for 50 words. Future work will expand lexicons and refine continuous verification, leveraging methods including a generative model to produce realistic synthetic data [9, 12].

## Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP24K20759, by the Fundamental Research Funds for the Central Universities (3132025240), China.

## References

- [1] Tao Chen, Yongjie Yang, Chonghao Qiu, Xiaoran Fan, Xiuzhen Guo, and Longfei Shangquan. 2024. Enabling Hands-Free Voice Assistant Activation on Earphones. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 155–168.
- [2] Haiming Cheng and Wei Lou. 2022. PD-FMCW: Push the limit of device-free acoustic sensing using phase difference in FMCW. *IEEE Transactions on Mobile Computing* 22, 8 (2022), 4865–4880.
- [3] Haiming Cheng, Wei Lou, Yanni Yang, Yi-pu Chen, and Xinyu Zhang. 2023. TwinkleTwinkle: Interacting with Your Smart Devices by Eye Blink. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–30.
- [4] Debadatta Dash, Evan Kittle, Isabel Gerrard, Richard Csaky, Gabriel Gonzalez, David Taylor, Juan Pablo Llinas, Dominic Labanowski, Nishita Deka, and Richy Yun. 2025. Silent Speech Recognition with Wearable Magnetometers. *bioRxiv* (2025), 2025–08.
- [5] Xuefu Dong, Yifei Chen, Yuuki Nishiyama, Kaoru Sezaki, Yuntao Wang, Ken Christofferson, and Alex Mariakakis. 2024. ReHEarSSE: Recognizing Hidden-in-the-Ear Silently Spelled Expressions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [6] Di Duan, Shengzhe Lyu, Mu Yuan, Hongfei Xue, Tianxing Li, Weitao Xu, Kaishun Wu, and Guoliang Xing. 2025. Argus: Multi-view egocentric human mesh reconstruction based on stripped-down wearable mmwave add-on. In *Proceedings of the 23rd ACM Conference on Embedded Networked Sensor Systems*. 1–14.
- [7] Di Duan, Zehua Sun, Tao Ni, Shuaicheng Li, Xiaohua Jia, Weitao Xu, and Tianxing Li. 2024. F2key: Dynamically converting your face into a private key based on cots headphones for reliable voice interaction. In *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*. 127–140.
- [8] Yongjian Fu, Ke Sun, Ruyao Wang, Xinyi Li, Ju Ren, Yaoxue Zhang, and Xinyu Zhang. 2025. Enabling Cardiac Monitoring using In-ear Ballistocardiogram on COTS Wireless Earbuds. *arXiv preprint arXiv:2501.06744* (2025).
- [9] Yongjian Fu, Shuning Wang, Linghui Zhong, Lili Chen, Ju Ren, and Yaoxue Zhang. 2022. SVoice: Enabling Voice Communication in Silence via Acoustic Sensing on Commodity Devices. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. Association for Computing Machinery, New York, NY, USA, 622–636.
- [10] Tian Gao, Xuefu Dong, Akihito Taya, Yuuki Nishiyama, and Kaoru Sezaki. 2025. Expression Recognition Based on Ear Canal Shape Detection Using Earbud and Ultrasound. In *IEICE Conferences Archives*. The Institute of Electronics, Information and Communication Engineers.
- [11] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (2019), 1–24.
- [12] Chen Gong, Bo Liang, Wei Gao, and Chenren Xu. 2025. Data Can Speak for Itself: Quality-guided Utilization of Wireless Synthetic Data. *arXiv preprint arXiv:2506.23174* (2025).
- [13] Alex Graves. 2012. *Connectionist Temporal Classification*. Springer Berlin Heidelberg, Berlin, Heidelberg, 61–93. doi:10.1007/978-3-642-24797-2\_7
- [14] Zengyi Han, Xuefu Dong, Liqiang Xu, Zhen Zhu, En Wang, Yuuki Nishiyama, and Kaoru Sezaki. 2024. RideGuard: Micro-Mobility Steering Maneuver Prediction with Smartphones. In *2024 IEEE 44th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 1039–1049.
- [15] Lixing He, Yunqi Guo, Haozheng Hou, and Zhenyu Yan. 2025. VibOmni: Towards Scalable Bone-conduction Speech Enhancement on Earables. *arXiv preprint arXiv:2512.02515* (2025).
- [16] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*. Association for Computing Machinery, New York, NY, USA, 14–27.
- [17] Lixing He, Bufang Yang, Di Duan, Zhenyu Yan, and Guoliang Xing. 2025. EmbodiedSense: Understanding Embodied Activities with Earphones. *arXiv preprint arXiv:2504.02624* (2025).
- [18] Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto. 2024. Conductive Fabric Diaphragm for Noise-Suppressive Headset Microphone. In *Adjunct Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology (Pittsburgh, PA, USA) (UIST Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, Article 56, 3 pages. doi:10.1145/3672539.3686768
- [19] Hirotaka Hiraki, Shusuke Kanazawa, Takahiro Miura, Manabu Yoshida, Masaaki Mochimaru, and Jun Rekimoto. 2024. WhisperMask: a noise suppressive mask-type microphone for whisper speech. In *Proceedings of the Augmented Humans International Conference 2024 (Melbourne, VIC, Australia) (AHs '24)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3652920.3652925
- [20] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiang Li, and Zhanpeng Jin. 2022. EarCommand: “Hearing” Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.
- [21] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. EarIO: A Low-power Acoustic Sensing Earable for Continuously Tracking Detailed Facial Movements. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 62 (jul 2022), 24 pages. doi:10.1145/3534621
- [22] Wenwei Li, Jiarun Zhou, Jie Xiong, Yuhui Xie, Leye Wang, Duo Zhang, and Daqing Zhang. 2025. Rethinking WiFi-based Angle Estimation for Robust Passive Indoor Localization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 4 (2025), 1–28.
- [23] Shengzhe Lyu, Yongliang Chen, Di Duan, Renqi Jia, and Weitao Xu. 2024. Earda: Towards accurate and data-efficient earable activity sensing. In *2024 IEEE Coupling of Sensing & Computing in AIoT Systems (CSCAIoT)*.
- [24] Wenguang Mao, Wei Sun, Mei Wang, and Lili Qiu. 2020. DeepRange: Acoustic ranging via deep learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 4, 4 (2020), 1–23.
- [25] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 529, 12 pages. doi:10.1145/3544548.3581392
- [26] Laxmi Pandey, Khalad Hasan, and Ahmed Sabbir Arif. 2021. Acceptability of speech and silent speech input methods in private and public. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–13.
- [27] Oxford University Press. 2023. *Oxford English Dictionary Online*. Oxford University Press, Oxford. <https://www.oed.com/search/advanced/Entries?textTermOpt0=WordPhrase&dateOfUseFirstUse=false&page=1&sortOption=Frequency>
- [28] Kazuki Shimomo, Zengyi Han, Yuuki Nishiyama, and Kaoru Sezaki. 2023. A Preliminary Study for Detecting Visual Search Behaviors During Street Walking Using Earable Device. In *Adjunct Proceedings of the 2022 ACM International Joint Conference on Pervasive and Ubiquitous Computing and the 2022 ACM International Symposium on Wearable Computers (Cambridge, United Kingdom) (UbiComp/ISWC '22 Adjunct)*. Association for Computing Machinery, New York, NY, USA, 254–257.
- [29] Zixiong Su, Shitao Fang, and Jun Rekimoto. 2023. LipLearner: Customizable Silent Speech Interactions on Mobile Devices. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (Hamburg, Germany) (CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 696, 21 pages. doi:10.1145/3544548.3581465
- [30] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. EarSSR: Silent Speech Recognition via Earphones. *IEEE Transactions on Mobile Computing* Early Access (2024), 1–17. doi:10.1109/TMC.2024.3356719
- [31] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. Eardynamic: An ear canal deformation based continuous user authentication using in-ear wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27.
- [32] Jingwen Wei, Lupeng Zhang, Jingchi Zhang, Bin Han, and Lei Wang. 2024. UniQR: A Secure QR Code Payment Scheme Using Device Pose and Environmental Matching. In *2024 21st Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.
- [33] Xuhai Xu, Haitian Shi, Xin Yi, Wenjia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K. Dey. 2020. EarBuddy: Enabling On-Face Interaction via Wireless Earbuds. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376836
- [34] WJ Youden. 1969. Statistical Techniques. *NBS Special Publication* 300-301 (1969), 421.
- [35] Shang Zeng, Haoran Wan, Shuyu Shi, and Wei Wang. 2023. mSilent: Towards General Corpus Silent Speech Recognition Using COTS mmWave Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 1 (2023), 1–28.
- [36] Chuanxin Zhang, Xue Jiang, and Dean Ta. 2024. Revealing the Incidence-Angle-Independent Frequency Shift in the Acoustic Rotational Doppler Effect. *Phys. Rev. Lett.* 132 (Mar 2024), 114001. Issue 11. doi:10.1103/PhysRevLett.132.114001
- [37] Ruidong Zhang, Ke Li, Yihong Hao, Yufan Wang, Zhengnan Lai, François Guimbretière, and Cheng Zhang. 2023. EchoSpeech: Continuous Silent Speech Recognition on Minimally-obtrusive Eyewear Powered by Acoustic Sensing. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–18.
- [38] Yu Zhang and Qiang Yang. 2018. An overview of multi-task learning. *National Science Review* 5, 1 (2018), 30–43.
- [39] Juntao Zhou, Dian Ding, Yijie Li, Yu Lu, Yida Wang, Yongzhao Zhang, Yi-Chao Chen, and Guangtao Xue. 2025. M2SILENT: Enabling Multi-user Silent Speech Interactions via Multi-directional Speakers in Shared Spaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.