POLITECNICO DI TORINO
Repository ISTITUZIONALE

SilentWear: an Ultra-Low Power Wearable System for EMG-based Silent Speech Recognition

Original
SilentWear: an Ultra-Low Power Wearable System for EMG-based Silent Speech Recognition / Spacone, G., Frey, S., Pollo, G., Burrello, A., Pagliari, D.J., Kartsch, V., Cossettini, A., Benini, L.. - In: IEEE SENSORS JOURNAL. - ISSN 1530-437X. - (2026), pp. 1-1. [10.1109/jsen.2026.3742328]

Availability:
This version is available at: 11583/3016807 since: 2026-10-20T07:39:30Z

Publisher:
IEEE

Published
DOI:10.1109/jsen.2026.3742328

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright
GENERICO preprint/submitted version accettata

This article has been accepted for publication in IEEE SENSORS JOURNAL, published by IEEE.

(Article begins on next page)

22 October 2026

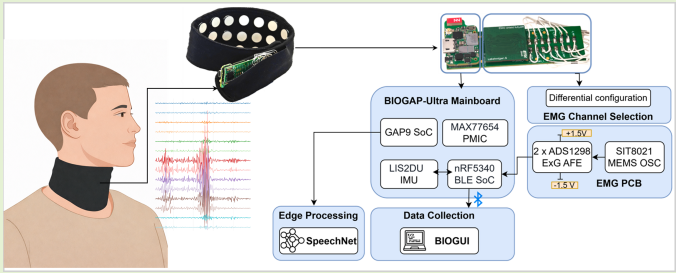
SilentWear: an Ultra-Low Power Wearable System for EMG-based Silent Speech Recognition

Giusy Spacone, *Graduate Student Member, IEEE*, Sebastian Frey, *Graduate Student Member, IEEE*, Giovanni Pollo, *Graduate Student Member, IEEE*, Alessio Burrello, *Member, IEEE*, Daniele Jahier Pagliari, *Member, IEEE*, Victor Kartsch, *Member, IEEE*, Andrea Cossettini *Senior Member, IEEE* and Luca Benini, *Fellow, IEEE*

arXiv:2603.02847v2 [eess.SP] 4 Mar 2026

***Abstract*— Detecting speech from biosignals is gaining increasing attention due to the potential to develop human-computer interfaces that are noise-robust, privacy-preserving, and scalable for both clinical applications and daily use. However, most existing approaches remain limited by insufficient wearability and the lack of edge-processing capabilities, which are essential for minimally obtrusive, responsive, and private assistive technologies.**

In this work, we present *SilentWear*, a fully wearable, textile-based neck interface for EMG signal acquisition and processing. Powered by BioGAP-Ultra, the system enables end-to-end data acquisition from 14 differential channels and on-device speech recognition. *SilentWear* is coupled with *SpeechNet*, a lightweight 15k-parameter CNN architecture specifically tailored for EMG-based speech decoding, achieving an average cross-validated accuracy of 84.8±4.6% and 77.5±6.6% for vocalized and silent speech, respectively, over eight representative human–machine interaction commands collected over multiple days. We evaluate robustness to repositioning induced by the multi-day use. In an inter-session setting, the system achieves average accuracies of 71.1±8.3% and 59.3±2.2% for vocalized and silent speech, respectively. To mitigate performance degradation due to repositioning, we propose an incremental fine-tuning strategy, demonstrating more than 10% accuracy recovery with less than 10 minutes of additional user data. Finally, we demonstrate end-to-end real-time on-device speech recognition on a commercial multi-core microcontroller unit (MCU), achieving an energy consumption of 63.9 µJ per inference with a latency of 2.47 ms. With a total power consumption of 20.5 mW for acquisition, inference, and wireless transmission of results, SilentWear enables continuous operation for more than 27 hours.



***Index Terms*— edge-AI, EMG, HMI, on-device, online, speech, silent speech, ultra-low power, wearables**

I. Introduction

Silent speech interfaces (SSIs) are assistive technologies designed to decode intended speech without overt vocalization, aiming to restore communication for individuals with speech impairments and to enable silent interaction in everyday contexts, such as noisy environments or privacy-sensitive scenarios [1]. Their development is further driven by the increasing demand for efficient, robust, and socially acceptable assistive solutions, fostering innovation toward comfortable, miniaturized, and energy-efficient wearable systems [2]. As these technologies mature and scale, they are also becoming more cost-effective, accelerating the transition of SSIs from research prototypes to practical real-world applications.

SSIs rely on various sensing modalities, among which surface electromyography (EMG) is particularly attractive due to its ability to capture speech-related activity directly at the neuromuscular source, providing low-latency information [1], [3]. At the same time, EMG is well suited for non-invasive, minimally obtrusive wearable systems with low energy requirements and able to operate over multiple days [4].

Advances in Deep Learning (DL) models have become key enablers for practical SSIs, allowing the extraction of complex non-linear representations from EMGs signals and improving robustness to the non-stationarities encountered in daily use. However, realizing truly wearable SSIs requires shifting the computational burden from the cloud to the edge, demanding efficient on-device inference to ensure real-time operation and responsiveness while minimizing wireless communication and preserving data privacy [1], [5]. Recent developments

This work was submitted on 20 February 2026.

G. Spacone, S.Frey, V.Kartsch, A. Cossettini, and L. Benini are with the Integrated Systems Laboratory of ETH Zürich, Zürich, Switzerland (correspondence to: gspacone@iis.ee.ethz.ch).

L. Benini is also with the Department of Electrical, Electronic and Information Engineering (DEI), University of Bologna, Bologna, Italy.

G. Pollo, A.Burrello, D.J. Pagliari are with the Department of Control and Computer Engineering, Politecnico of Turin, Italy.

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

in energy-efficient edge platforms, such as VersaSens [6] or BioGAP-Ultra [7], address these constraints by providing efficient biosignal acquisition together with the computational capability required to execute DL models under strict latency and battery limitations.

Despite steady progress and partial advancements, current EMG-based SSIs still fall short in several key areas essential to reliable, socially acceptable everyday use [8]. The first bottleneck is the sensing interface: most of existing systems still rely on wet electrodes, facial placements, or bulky form factors, which are cumbersome, stigmatizing, and impractical for continuous wear. Second, while some approaches report strong results on individual subjects, systematic characterization across diverse users and, critically, performance stability across sessions and multiple days remain limited. Third, introducing SSIs into daily life requires embedded deployment, which guarantees low latency, long battery life and inherently safeguards user privacy by keeping sensitive biosignals local [1].

To cope with these challenges, we introduce *SilentWear*, a novel, fully dry, textile-based speech interface in a neckband form factor, powered by the BioGAP-Ultra platform [7] for end-to-end on-device EMG acquisition and processing. Extending our previous conference work [9], we present the following additonal novel contributions:

- We present a new dataset[1] from four subjects, consisting of vocalized and silent EMG recordings of 8 commands for human-machine interfaces (HMIs), collected over multiple days, under realistic sensor repositioning scenarios.
- We propose *SpeechNet*, a compact convolutional neural network (CNN) architecture, showcasing the feasibility of detecting HMI-related commands on-board, with an average cross-validated accuracy of $84.8 \pm 4.6\%$ for vocalized and $77.5 \pm 6.6\%$ for silent speech.
- We assess the cross-day generalization by evaluating on an unseen recording day (excluded from training), achieving an average inter-session accuracy across subjects of $71.1 \pm 8.3\%$ and $59.3 \pm 2.2\%$ (vocalized/non-vocalized).
- To overcome the challenges arising from repositioning, we propose an incremental fine-tuning strategy, achieving more than 10% accuracy recovery with less than 10minutes of data re-collection.
- We deploy the network architecture on the GAP9 Micro-Controller Unit (MCU) embedded in BioGAP-Ultra, showcasing an inference time of 2.47ms, with an energy/inference of 63.9 μJ. Including signal acquisition, on-device inference, and wireless transmission of classification results, the system consumes $20.5\,\mathrm{mW}$, yielding $27.1\,\mathrm{h}$ of battery life with a $150\,\mathrm{mAh}$ Li-Po battery.

II. Related Works

Over the last decade, significant progress has been made towards the development of wearable SSIs, leveraging diverse sensing modalities such as Inertial Measurement Unit (IMU), strain sensing, Electroencephalogram (EEG), and Electrooculography (EOG). Among these, EMG is particularly promising due to its ability to capture neuromuscular activations at the source while remaining cost-effective, easy to use, and minimally obtrusive [1].

On the algorithmic side, open-source datasets such as that of Gaddy *et al.* [10] have enabled the development of robust deep learning architectures tailored to speech recognition. However, such datasets have been collected under controlled laboratory conditions that do not reflect realistic daily usage (e.g., obtrusive electrode placements, wet electrodes, fixed sensor positioning).

Bridging the gap between algorithmic demonstrations under ideal data collection settings and truly wearable systems requires addressing major challenges: i) the development of unobtrusive sensing interfaces robust to daily variability induced by device removal, repositioning, and physiological changes; ii) the adoption of energy-efficient hardware platforms supporting on-device processing for low latency, extended battery life, and privacy preservation; iii) the design of lightweight algorithms compatible with embedded deployment constraints.

Several wearable SSIs solutions have been proposed in the state-of-the-art (SoA) and are discussed in the following sections; a comparative overview is provided in Table IV.

Kapur *et al.* [11] presented one of the first approaches toward wearable SSIs with AlterEgo, a head-worn system employing face-contact electrodes. They demonstrated strong recognition performance (92% accuracy) on a dataset collected from 10 subjects. Despite these promising results, the system relies on a relatively bulky form factor, which may limit user acceptance in everyday scenarios. To address such limits, subsequent works, including Liu *et al.* [12] and Wang *et al.* [13], focused on the development of novel electrode technologies (e.g., tattoo-like electrodes) designed for conformal facial and jaw placement. These approaches reduced thickness and improved flexibility and comfort. Nevertheless, face-based electrode placement raises stigmatization concerns and may hinder long-term usability and social acceptance.

More recently, alternative sensing locations have been explored to improve wearability. Wu *et al.* [14] proposed a textile-based neckband integrating gold-plated electrodes and an energy-efficient acquisition system (WANDmini [15]) with Bluetooth Low Energy (BLE) connectivity for multi-day operation ($46\,\mathrm{mW}$). They demonstrated the feasibility of recognizing 11 phonemes from two subjects, achieving 92.7% accuracy with a Random Forest classifier. However, the results are based on an extremely limited dataset (110 samples in total), a vocabulary with limited real-world applicability, and the work remains narrow in its algorithmic exploration. Tang *et al.* [16] investigated the ear region, embedding four textile-based electrodes into commercial headphone earmuffs. The system, which relies on custom hardware (a SeedStudio EMG readout combined with a WiFi-enabled ESP32-S3 MCU), achieves 96% recognition accuracy on 10 daily-life commands (e.g., start, stop) using a Squeeze-Excitation ResNet over a dataset comprising 10 subjects.

Despite significant progress, several challenges remain before EMG-based wearable SSIs can achieve widespread real-

[1]Dataset: `https://huggingface.co/datasets/PulpBio/SilentWear`
Code: `https://github.com/pulp-bio/SilentWear`

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

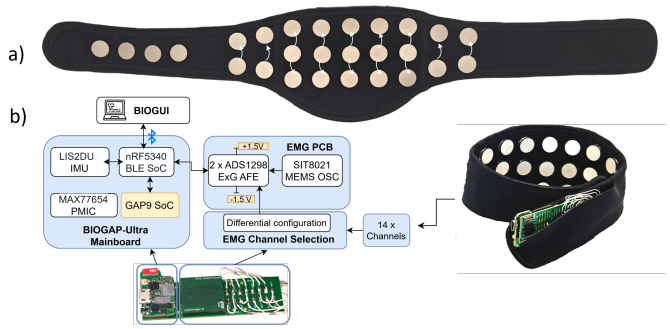


Fig. 1: a) SilentWear neckband. The interface features 14 EMG differential acquisition channels; 4 channels provide the ground reference.; b) Overview of the acquisition hardware, based on BioGAP-Ultra. The system features a baseboard, for BLE communication and on-device processing, and an EMG acquisition board, with a total size of $26 \times 65 \times 13\,mm^3$.

world adoption. First, most algorithmic studies conducted on data collected with wearable SSIs rely on small datasets collected under highly controlled conditions, typically excluding realistic factors such as sensor repositioning and multi-day acquisitions. These factors, however, are intrinsic to daily use and critically affect system robustness. Second, the lack of standardized data collection protocols and evaluation methodologies makes it difficult to conduct fair algorithmic comparisons across studies. Third, the majority of evaluations are performed on external computing platforms (laptops). For truly deployable systems, low-latency and privacy-preserving operation, on-device inference under embedded hardware constraints is needed.

Our work targets the identified gap at the intersection of wearability, robustness, and embedded intelligence, extending the results presented in our previous work. We explicitly address the challenges arising from multi-day system use and enable fully end-to-end, low-latency, and privacy-preserving on-device speech detection.

III. Materials and Methods

A. Silent Wear Interface

In this study, we employ the hardware presented in our previous work [9]. A description of the system design is presented in Fig. 1.

The interface is a soft-fabric neckband with fully dry electrodes, adjustable in size via velcro straps. The design follows the principles of simplicity and everyday usability. Since no wet electrodes or conductive gel are required, it can be worn immediately, supporting daily use without being intrusive. The inner side of the band contains 27 sewn-in snap fasteners that support connection to 27 fully dry electrodes (Datwyler, SoftPulse [17]). The central portion of the neckband contains 15 electrodes arranged in three rows and connected as overlapping differential channels (top-middle and middle-bottom), such that adjacent channels share the middle electrode, resulting in 10 effective differential channels. This design choice ensures higher coverage of the infrahyoid muscles, which are primary active during phonation and articulation [18]. In addition, four fully differential channels are placed laterally (two per side of the neck), to ensure coverage of the sternocleidomastoid muscle [19]. Hence, the system features a total of 14 acquisition channels; four electrically shorted electrodes, placed at the back of the neck, provide the ground reference.

The acquisition and processing hardware is based on BioGAP-Ultra [7]. BioGAP-Ultra comprises a baseboard that integrates GAP9, a SoA ultra-low-power (ULP) System-on-Chip (SoC) with AI capabilities, able to execute Deep Neural Networks (DNNs) efficiently on-board and the Nordic nRF5340 MCU with BLE capabilities. The baseboard interfaces with a biopotential acquisition board equipped with two Texas Instruments ADS1298 Analog-Front-Ends (AFEs), which enable simultaneous acquisition from 16 differential EMG channels. A complete description of the system is provided in [7]. The acquisition hardware has a total dimension of $26 \times 65 \times 13\,mm^3$ and is powered by a standard 150 mAh Li-Po battery.

B. Data Collection Protocol

We adopt a data collection protocol inspired by our previous work [9]. An overview of the protocol is shown in Fig. 2.

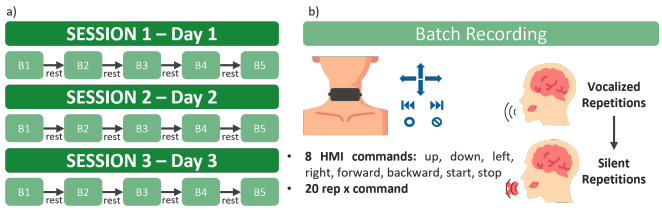


Fig. 2: Data collection protocol. (a) Recordings are organized into sessions conducted over multiple days. Each session consists of 10 batches (5 vocalized and 5 silent), executed alternately, with 2-minute rest intervals between batches. (b) In each batch, subjects repeat 8 HMI-related commands, each performed 20 times in randomized order.

Data acquisition and user guidance are managed through the open-source BioGUI software [20][2]. Specifically, the BioGUI is responsible for setting the acquisition configuration (sampling rate of 500 Hz and programmable-gain amplifier (PGA) gain of 6) for the two ADS1298 AFEs, displaying real-time EMG signals, displaying user instructions, labeling and storing data.

During data collection, participants sit comfortably in a chair. The recordings are organized into 3 *sessions* per subject, conducted over multiple days. Between sessions, the neckband is repositioned to reproduce realistic usage conditions and to account for variability in sensor placement. Each session consists of multiple *batches*. In each batch, subjects are instructed to repeat eight HMI-related commands (*up, down, left, right, forward, backward, start, stop*) in randomized order. Each command is repeated 20 times. For every repetition, subjects are given 2 s for speech production (vocalized or silent), followed by a 1.5 s rest interval. During the rest period, the GUI displays the next command to minimize reaction time.

[2]Data collection software: `https://github.com/pulp-bio/biogui/tree/sensors_speech`

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

Each session comprises ten batches, alternating between five vocalized and five silent batches. A 2 min break is provided between consecutive batches to reduce fatigue.

With 8 commands, 20 repetitions per command, and 3.5 s per repetition (2 s production + 1.5 s rest), each batch lasts approximately 560 s ($\approx$ 9 min 20 s). Including the 2 min breaks between the ten batches, the total duration of each session is approximately 110 min. Overall, the dataset comprises 2400 vocalized and 2400 silent utterances per subject (8 commands $\times$ 20 repetitions $\times$ 5 batches $\times$ 3 sessions). For each condition, an additional 2400 samples correspond to rest segments.

We collected data from four subjects (three male, one female, average age 30 years old). Written informed consent was obtained from all participants prior to data acquisition. All procedures complied with the Declaration of Helsinki.

C. Data Processing and Dataset Creation

Raw EMG signals are preprocessed using a fourth-order zero-phase Butterworth high-pass filter with a cutoff frequency of 20 Hz, followed by a 50 Hz notch filter, consistent with our previous work.

Labels are derived from the trigger signal generated by the BioGUI and recorded together with the EMG data. Each word and each rest period is assigned a unique identifier. The continuous EMG signal is segmented into windows starting at the corresponding trigger onset and ending at the subsequent trigger (rest between words). Each window inherits the label of the associated word or rest period. As a result, for each subject and condition (silent or vocalized), the dataset contains 4800 samples: 300 samples per command (8 commands) and 2400 rest samples. During model training, the rest class is randomly downsampled to ensure class balance.

D. Evaluation Settings

To assess the quality of the acquired signals, we tested three evaluation protocols to discriminate between the 9 classes (8 commands plus rest) collected:

I) Global Evaluation Setting: Data from all sessions of a subject are pooled and evaluated using a *leave-one-batch-out* cross-validation scheme. In each fold, four batches from each session are used for training, and the remaining three batches (one for each session) are used for testing. This protocol establishes a baseline and serves as a means of comparison with prior works, which do not include sensor repositioning. Since the dataset spans multiple recording days, both training and test sets draw from the same overall data distribution. Testing is, however, always performed on unseen batches, thereby explicitly accounting for batch-to-batch variability.

II) Inter-Session Evaluation Setting: This setting adopts a *leave-one-session-out* cross-validation scheme. In each fold, two sessions (total of 10 batches) are used for training, and the remaining session for that subject is used for testing. Therefore, the held-out test fold contains completely unseen data, collected on a different day and after sensor repositioning. This protocol assesses model robustness to sensor repositioning and day-to-day variability.

III) Incremental Fine-Tuning Setting: This protocol assesses the amount of data needed to calibrate the system after sensor repositioning by an incremental training strategy, which is studied over two scenarios:

- *a) Fine-tuning from a pre-trained model*. We initialize the model with weights pre-trained on two sessions (as in Setting *II*) and evaluate its zero-shot accuracy on the first batch of a new acquisition session. The model is then incrementally fine-tuned by incorporating data from the new batches. Specifically, 70% of the batch data (126 utterances, corresponding to 14 samples per class) are used for fine-tuning, while the remaining 30% are reserved for validation. The updated model is then evaluated zero-shot on the subsequent batch, followed by another fine-tuning step. This procedure is repeated sequentially for the four batches. We repeat this experiment three times, using 2 sessions each time for model pre-training (as in *II*) and fine-tuning on the remaining one.
- *b) Training from scratch*. We initialize the model's weights randomly. At each training round, the model is trained using 70% of a new batch (126 utterances, 14 samples per class), while the remaining 30% are used for validation. Performance is evaluated in a zero-shot manner on the next unseen batch (from batch 2 to batch 5). We repeat this experiment for the three acquisition sessions.

We train subject-specific and condition-specific (vocalized or silent) models. For all evaluation settings, we report the mean and standard deviation of the balanced accuracy across batches or sessions. In addition, we report results averaged across subjects by computing the mean of the per-subject scores and their corresponding standard deviation.

E. Network Architecture: SpeechNet

We design a network architecture inspired by EpiDeNet [21], developed for seizure detection and already proven to be suitable for HMI-related tasks [22].

As in [21], the first three blocks learn temporal patterns using kernels that operate independently on each sensor channel. For the first three layers, zero-padding is applied to make the architecture flexible with respect to any temporal window size. The last two blocks learn cross-channel spatial representations, integrating information across multiple channels. The features learned in the convolutional blocks are then pooled and passed to a single fully connected layer (output size 9 classes). Further details on the network architecture are displayed in Table I. The total number of parameters is 15489.

The network is trained using the Cross-Entropy Loss and the Adam optimizer. For all evaluation settings, we use an initial learning rate of 10^{-3} and the *Reduction on Plateau* scheduler, applied to the validation loss (patience of 2 epochs). Weight decay is set to 10^{-4}. In evaluation settings I, II, III-b, we train for a maximum of 100 epochs. For evaluation setting III-a, we apply fine-tuning for 50 epochs on models pre-trained on 100 epochs (as in II). In both cases, early stopping with a patience of 10 epochs is applied if validation loss stops improving. Models are initially trained with an EMG input size of 1400ms, as in our previous work [9], and to enure that

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

TABLE I: SpeechNet Architecture

Type	Layer	# Filters	Kernel	Output
ϕ^1	Conv2D	8	(1×4)	$(8, 14, T)$
	MaxPool	–	(1×8)	$(8, 14, T/8)$
ϕ^2	Conv2D	16	(1×16)	$(16, 14, T/8)$
	MaxPool	–	(1×4)	$(16, 14, T/32)$
ϕ^3	Conv2D	16	(1×8)	$(16, 14, T/32)$
	MaxPool	–	(1×4)	$(16, 14, T/128)$
ϕ^4	Conv2D	32	(7×1)	$(32, 8, T/128)$
	MaxPool	–	(1×1)	$(32, 8, T/128)$
ϕ^5	Conv2D	32	(7×1)	$(32, 2, T/128)$
	MaxPool	–	(1×1)	$(32, 2, T/128)$
ϕ^6	AdaptiveAvgPool	–	–	$(32, 1, 1)$
ϕ^7	Dense	–	–	9

T = number of time samples; for ϕ^1, ϕ^2, ϕ^3 zero-padding is applied on the time dimension.

the same amount of samples is available both for words and the rest data (duration of 1500ms between words, as described in Sec. III-B).

In addition, to enable comparison with [9] and assess the performance of the new architecture, we also train a Random Forest Classifier (maximum 100 estimators), adopting manual feature extraction. The 1400ms of data are segmented into seven non-overlapping windows; from each window, we extract time and frequency-domain features and use them as training data. Additional details are available in [9] (Section II-D).

We use PyTorch for the CNN models; the Scikit-learn library is used to train the Random Forest models and for data preparation. We run all experiments on a single GPU (Nvidia GEFORCE GTX 1080).

F. Ablation on window size and information transfer rate

We conduct an ablation on the EMG window size to assess its impact on two scores: accuracy and information transfer rate (ITR), a widely adopted metric in brain-computer interfaces [23]. The ITR (bit/min) is computed as:

$$\mathrm{ITR} = \frac{60}{T}\left[\log_2(C) + P\log_2(P) + (1-P)\log_2\left(\frac{1-P}{C-1}\right)\right]$$

where C is the number of classes, T the window lenght (s), P the classification accuracy (0-1 range). We ablate over window sizes spanning from 400ms to 1400ms, with 200ms steps.

G. Model Deployment

We deploy the model on the GAP9 MCU integrated in BioGAP-Ultra. GAP9 combines a single RISC-V controller core with a programmable 9-core RISC-V compute cluster and a dedicated Neural Network (NN) hardware accelerator (NE16). GAP9 provides a hierarchical on-chip memory system with 128 kB of L1 memory and 1.6 MB of L2 RAM, while BioGAP-Ultra augments the platform with an additional 64 MB off-chip PSRAM used as L3 memory.

For model deployment, we use the proprietary NNTool framework [24]. First, we export the trained network to the Open Neural Network Exchange (ONNX) format and import it into NNTool to map the computational graph to the GAP9 processor and generate optimized C code. NNTool also performs 8-bit integer post-training quantization for efficient execution. Model constants (weights, biases, and quantization parameters) are stored in off-chip Flash (overall footprint $\approx$15.13 kB) and promoted to L2 at initialization. Since the complete network fits in on-chip memory, inference proceeds using only on-chip L2 storage, with per-layer tiles moved to L1 by the DMA during kernel execution. We evaluate two operating points that reflect common deployment objectives:

- *Maximum inference speed*: NE16 runs at its highest frequency 370 MHz and 0.8 V to minimize latency.
- *Highest energy efficiency:* NE16 runs at 240 MHz and 0.65 V, its most efficient operating point.

IV. Results

A. EMG acquisitions

Figure 3 shows a representative example of the signals acquired from one of the subjects (S04) during the repetition of each command included in the dataset, for both vocalized and silent speech. Clear differences can be visually observed between the rest (no speech) and articulation phases in both vocalized and silent conditions, confirming the system's ability to detect speech-related muscle activity reliably.

Furthermore, across all commands, the central channels exhibit a higher signal-to-noise ratio (SNR) than the lateral channels. Based on the neckband design, these central electrodes are positioned over the infrahyoid muscles (sternohyoid, omohyoid, sternothyroid, and thyrohyoid), which are involved in laryngeal stabilization and vertical movement during phonation and articulation [18]. The stronger activation observed on these channels is therefore consistent with the physiological role of these muscles in modulating vocal fold positioning and supporting speech production.

B. SpeechNet Performance

To analyze the performance gains of SpeechNet over our previous work, we first compare it against the previously used Random Forest classifier. To ensure a direct comparison, both models are trained using a window length of 1400 ms. Table II summarizes the performance for the two pipelines under the *Global* and *Inter-Session* evaluation settings, where SpeechNet consistently outperforms the Random Forest baseline in both settings and speaking conditions. Specifically, in the *Global* setting, SpeechNet improves accuracy by approximately 15% for vocalized speech and 15% for silent speech. In the *Inter-Session* setting, the corresponding gains are approximately 13% for vocalized and 11% for silent speech. Given its consistently superior performance, SpeechNet is adopted as the reference model for subsequent analyses.

Figure 4 reports the per-subject confusion matrices for the 1400 ms window under both the *Global* and *Inter-Session* evaluation settings. Several observations emerge. In the *Global* setting, diagonal dominance is consistently strong across subjects. In the vocalized condition, misclassifications are mainly confined to the *start* and *stop* pair. In the silent condition, errors predominantly affect short-duration commands, particularly for subjects S03 and S04, suggesting increased sensitivity to reduced articulatory dynamics.

Greater attention must be paid to the *Inter-Session* evaluation case. First, the *rest* class is classified with near-perfect accuracy across all subjects and speaking conditions, confirming

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

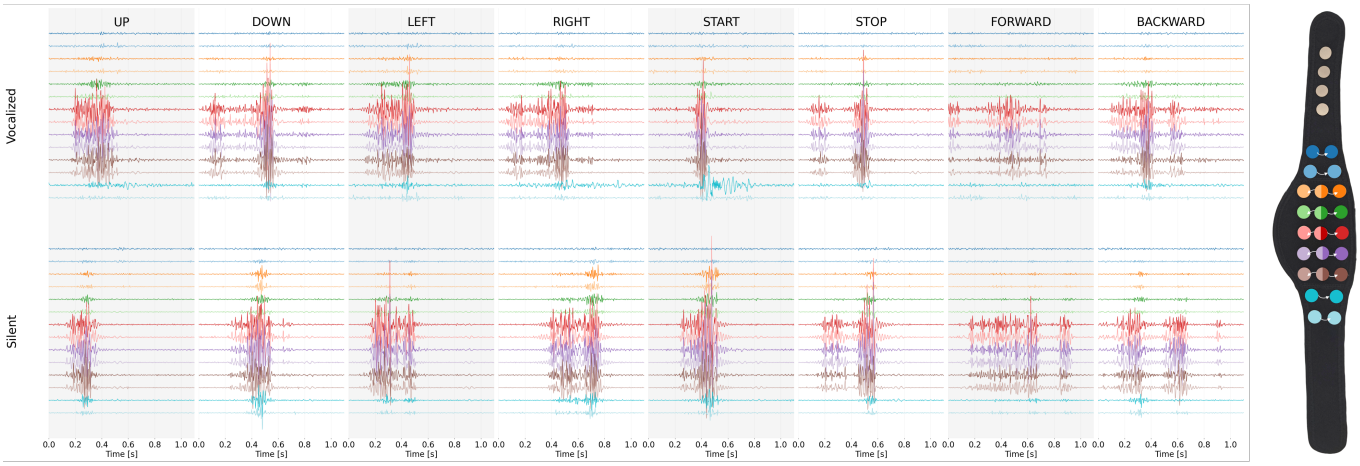


Fig. 3: Visualisation of the signals of the 14 EMG Channels for the 8 HMI commands, for vocalized (top) and silent (bottom) recordings.

TABLE II: Accuracy Comparison for Global and Inter-Session Evaluations between Random Forest and SpeechNet (window size 1.4s).

		Global Evaluations					Inter-Session Evaluations				
Condition	Model	S01	S02	S03	S04	Avg.	S01	S02	S03	S04	Avg.
Vocalized	Random Forest	78.4 ± 3.0	60.1 ± 3.0	63.2 ± 4.8	75.9 ± 4.7	69.4 ± 7.9	70.7 ± 2.6	42.4 ± 14.1	53.1 ± 8.6	67.9 ± 7.0	58.5 ± 11.5
	SpeechNet	90.3 ± 1.6	82.6 ± 1.3	78.3 ± 6.2	87.8 ± 2.5	84.8 ± 4.6	77.0 ± 6.2	58.8 ± 19.0	68.3 ± 2.5	80.2 ± 7.4	71.1 ± 8.3
Silent	Random Forest	68.2 ± 3.1	64.8 ± 3.3	59.6 ± 3.0	58.7 ± 2.7	62.9 ± 3.9	56.2 ± 3.2	47.4 ± 8.6	45.6 ± 4.5	45.6 ± 1.8	48.7 ± 4.4
	SpeechNet	84.4 ± 3.2	83.2 ± 2.0	74.1 ± 2.9	68.5 ± 2.8	77.5 ± 6.6	62.9 ± 2.3	56.8 ± 17.7	59.2 ± 2.8	58.5 ± 3.2	59.3 ± 2.2

the robustness of speech activity detection. This remains true even in the presence of short-duration commands (e.g., *up*), which could potentially increase the risk of misclassification. Second, longer commands (e.g., *forward*, *backward*) exhibit, on average, higher recognition accuracy than shorter commands. When misclassifications occur, *forward* and *backward* are more likely to be confused with each other, plausibly due to similarities in their underlying neuromuscular activation patterns reflected in comparable EMG temporal dynamics and signal duration (see Fig. 3).

For the remaining commands, no consistent cross-subject misclassification trend is observed. Off-diagonal entries display relatively high variance across folds, indicating that errors are not systematic but rather session-dependent. Such variability is likely driven by inter-session factors, including sensor repositioning, day-to-day physiological changes, and natural variations in speech tempo and articulation strategy.

In both the vocalized and silent conditions, subject S02 exhibits consistently lower performance than the other participants under the *Inter-Session* setting, particularly in the vocalized condition. The reduced accuracy is primarily driven by the model trained on sessions 2-3 and evaluated on session 1. Visual inspection revealed differences in amplitude and waveform morphology in session 1 compared to subsequent sessions, which are plausibly attributable to differences in neckband fit (e.g., looser contact) or sensor placement. Importantly, this behavior reflects realistic usage conditions and highlights the sensitivity of inter-session performance to hardware fitting and positioning.

Overall, the performance gap between the *Global* and *Inter-Session* settings provides strong evidence that multi-day use induces a distribution shift in the EMG feature space. When training and testing samples are drawn from the same global distribution, the model can internalize session variability; conversely, when evaluated on a previously unseen session, performance degrades, underscoring the impact of sensor repositioning and day-to-day physiological variability, and highlighting the need for mitigation strategies, which will be further described in Section IV-D.

C. ITR Analysis

Figures 5 and 6 report the accuracy and ITR under the *Inter-Session* evaluation protocol for vocalized and silent speech, respectively. For both conditions, larger window sizes yield the best results.

In the vocalized condition, the highest average accuracy is obtained with a $1400\,\mathrm{ms}$ window (71.1%). The maximum ITR is achieved with an $800\,\mathrm{ms}$ window, reaching $92.7\,\mathrm{bit/min}$. In the silent condition, the highest average accuracy is obtained with a $1200\,\mathrm{ms}$ window (59.6%). The maximum ITR is achieved with an $800\,\mathrm{ms}$ window, with $69.2\,\mathrm{bit/min}$.

In both speaking conditions, larger window lengths generally lead to higher classification accuracy. However, a shorter window ($800\,\mathrm{ms}$) provides a practical trade-off between robustness (near-peak accuracy), throughput (high ITR), and reduced latency, at the cost of a 9% and 3.3% accuracy drop in vocalized and silent setups, respectively.

For this reason, in the next section, we compare results for

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

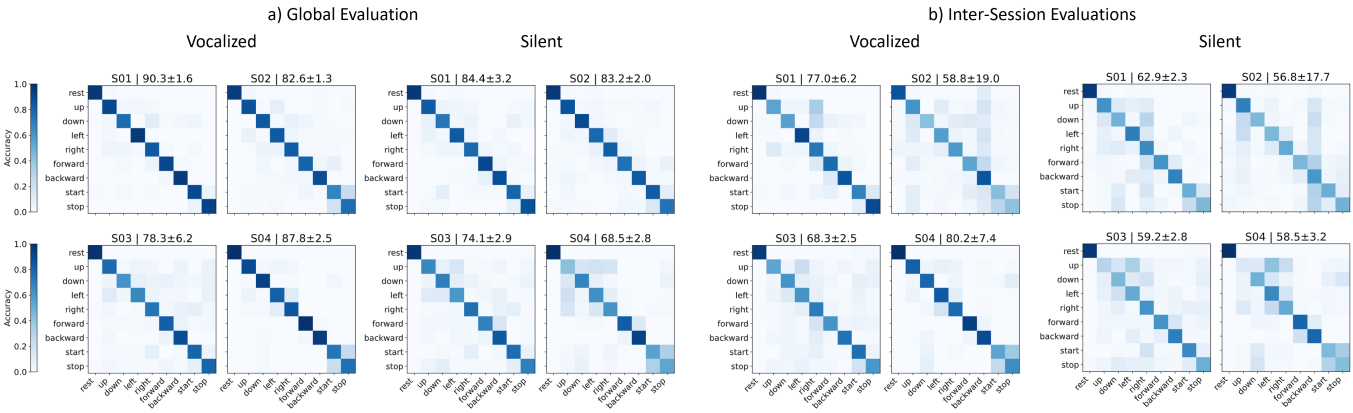


Fig. 4: Per Subjects Confusion Matrices for the a) Global Evaluation Setting and b) Inter-Session Evaluation Setting for vocalized and silent speech, for EMG window size of 1400ms.

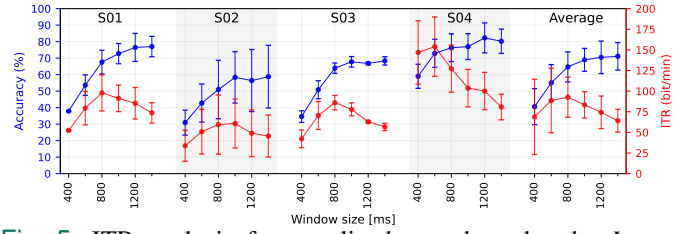


Fig. 5: ITR analysis for vocalized speech under the *Inter-Session* evaluation setting. Blue curves (left axis) show mean accuracy and red curves (right axis) show mean ITR as a function of window size. For S01–S04, markers with error bars represent the mean and standard deviation across sessions. In the average plot, markers with error bars represent the mean and standard deviation across subjects.

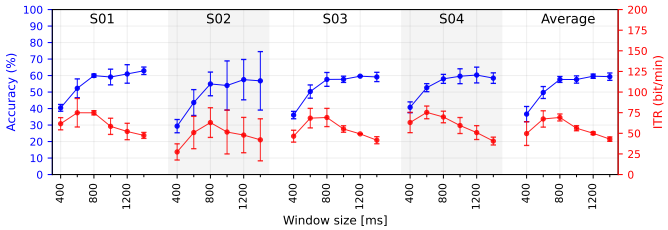


Fig. 6: ITR analysis for silent speech under the *Inter-Session* evaluation setting. Plot structure and statistical reporting follow the same conventions as in Fig. 5.

two representative configurations: 800 ms, prioritizing responsiveness and ITR, and 1400 ms, prioritizing accuracy.

D. Incremental Fine Tuning Results

Figures 7 and 8 report the incremental fine-tuning setting results for the vocalized and silent conditions, respectively, using a window length of 1400 ms. Similar trends are observed for the 800 ms configuration.

In both speaking conditions, fine-tuning consistently improves performance compared to evaluating the model without adaptation (i.e., using the zero-shot scenario with the model pre-trained on two acquisition sessions, without additional fine-tuning epochs, as explained in detail in Sec. III-D). Moreover, the fine-tuned models systematically outperform models trained from scratch, confirming the clear benefits of pre-training.

For a window size of 1400 ms, in the vocalized condition, the best average accuracy is achieved after four fine-tuning rounds (84%). Across the five batches, fine-tuning improves the average accuracy by approximately 8.5% compared to the non-fine-tuned baseline. In the silent condition, peak performance is achieved after five fine-tuning rounds (79% accuracy), with an average improvement of approximately 12% across batches compared to the non-fine-tuned model.

Similar behaviour is observed for the 800 ms window. For vocalized speech, performance peaks after four fine-tuning rounds (78.8%), reaching an average accuracy of 73.7%, corresponding to an improvement of approximately 8% over the non-adapted baseline. For silent speech, peak performance is achieved after five fine-tuning rounds (73% accuracy). The average accuracy across batches reaches 67.0%, corresponding to an improvement of approximately 10% compared to the non-fine-tuned model.

From a practical usability perspective, even a single fine-tuning round provides a substantial benefit. With a window size of 1400 ms, one round of adaptation increases accuracy from 70% to 80% in the vocalized condition and from 58% to 72% in the silent condition. For the 800 ms configuration, a single fine-tuning round yields an improvement of approximately 7% in the vocalized case and 12% in the silent case. In a real-world deployment scenario, this corresponds to recollecting 20 repetitions per word, corresponding to 10 min of data acquisition. However, this duration could be further reduced through protocol optimizations (e.g., shortening rest intervals between words) and by leveraging a continuous source of ground-truth labeling, such as a microphone, to streamline the annotation process.

E. Model Deployment

Table III reports the deployment results on the GAP9 microcontroller. We deploy the model trained with an EMG input window of 800 ms. The slight accuracy reduction associated with the shorter window is compensated by a higher ITR, as shown in Figs. 5 and 6. Moreover, inference latency and energy consumption are reduced nearly linearly, resulting in approximately a 50% reduction.

The 370 MHz/0.8 V operating point achieves the lowest latency, reaching 1.60 ms for a window size of 0.8 s, while

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

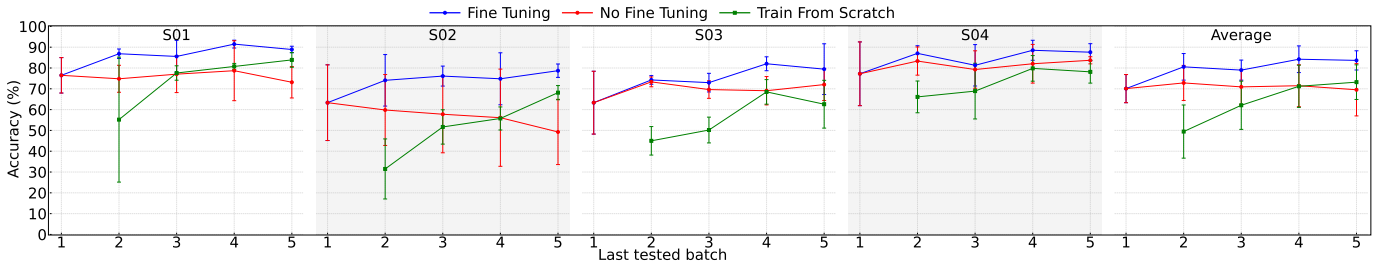


Fig. 7: Incremental training performance for vocalized speech (1400 ms input window). Results are shown per subject (S01–S04) and averaged across subjects (right). The blue curve represents fine-tuning from a model pre-trained on two sessions; the red curve shows the pre-trained model without fine-tuning; the green curve corresponds to training from scratch. For $x = 1$, performance is evaluated on the pre-trained model, thus blue and red overlap (no fine-tuning available at this step). Error bars denote standard deviation across sessions (n=3) for individual subjects and across subjects for the average.

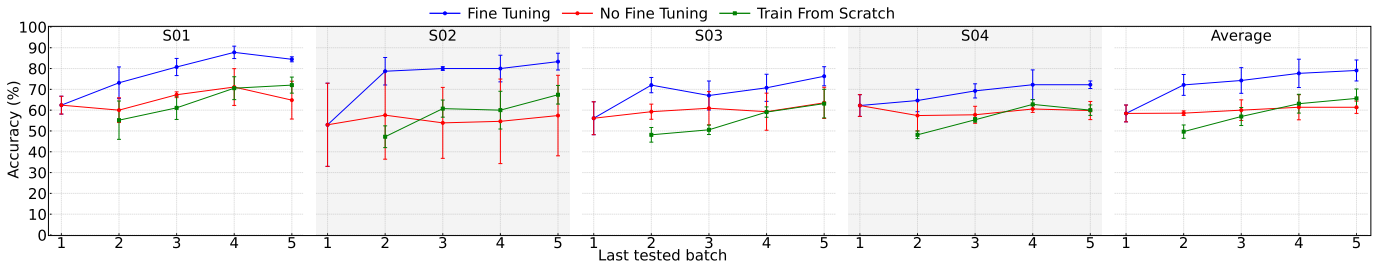


Fig. 8: Incremental training performance for vocalized speech (1400 ms window); methods is equivalent to the vocalized condition.

TABLE III: Model deployment on GAP9.

Input size	14 × 400	
MACs	2145984	
Frequency [MHz]	240	370
Voltage [V]	0.65	0.8
Time/inference [ms]	2.47	1.60
Cycles/inference	593497	593015
MACs/cycle	3.62	3.62
Energy/inference [μJ]	63.9	97.1

the 240 MHz/0.65 V configuration minimizes energy per inference, requiring 63.9 μJ. In both operating points, the accelerator sustains 3.62 MACs/cycle, indicating compute-bound execution on NE16. The network is deployed with 8-bit post-training quantization, resulting in a compact model size of approximately 15.13 kB.

For continuous operation, we adopt a sliding-window inference scheme with a step size of 100 ms, i.e., one inference every 100 ms. Focusing on the most energy-efficient operating point (240 MHz/0.65 V), the total system power is 20.5 mW, corresponding to an estimated 27.1 h of battery life with a 150 mAh battery. The power budget is dominated by EMG acquisition (15.3 mW), followed by data handling and result streaming (4.55 mW), while edge computation contributes only 0.639 mW.

F. Comparison with previous works

Table IV compares our solution with previous work. From a system-level approach, our solution offers clear advantages in terms of wearability, comfort, and suitability for daily use. In particular, we propose a minimally obtrusive design based on electrodes embedded in a textile neckband, rather than face-mounted configurations such as those presented by Kapur [11] and Wang [13]. This design choice aims to improve user acceptance and long-term usability in real-world scenarios. Our design is inspired by the concept proposed by Wu et al. [14]. However we adopt a differential acquisition scheme to reject common-mode noise sources. This choice eliminates the need for wet reference electrodes, thereby enhancing wearability, ease of use, and long-term comfort.

Powered by the GAP9 platform, our system is, to the best of our knowledge, the first neck-based SSI solution to integrate on-device AI processing capabilities. This represents a key advantage, particularly in privacy-sensitive contexts, as signal processing and inference can be performed locally without transmitting raw biosignals.

During on-device inference, the power consumption is 20.5 mW, enabling continuous operation for 27.1 h using a 150 mAh Li-Po battery ($2.1\times2.8\times0.4$ cm^3, 4g).

From an application point of view, a quantitative comparison across prior SSI studies remains inherently difficult due to the lack of standardized data collection protocols and the absence of uniform evaluation methodologies. Compared to previous works, we intentionally acquired data over multiple days to better reflect realistic usage. This setting introduces additional sources of variability that are known to challenge model generalization: day-to-day physiological changes (e.g., in skin conductivity), domain shifts caused by sensor repositioning between sessions, and potential changes in speaking behavior across days (e.g., articulation pace and consistency). All of these factors can alter EMG amplitude, temporal structure, and overall signal morphology, ultimately increasing the difficulty

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

TABLE IV: Comparison of EMG-based SSI Systems (System and Application Level)

		Kapur 2018 [11]	Wang 2021 [13]	Wu 2025 [14]	Tang 2025 [16]	**Our Work**
System	Form Factor	Plastic Face-Neck Apparatus	Tattoos + Ear-mounted Electronics	Textile Neckband	Headphone Earmuff	Textile Neckband
	Sensing Location	Face	Face	Neck	Ears	Neck
	Electrode Type	Wet Gold-Plated + conductive paste[a]	Dry Au/Cr on PET	Dry Gold-Plated + Wet Ref.	Dry Textile-Based Graphene/PEDOT:PSS	Dry [17] Conductive Polymer-Elastomer + Ag/AgCl coating
	Channels	7–CR	4–CR	10–CR	4–CR	14–Diff[b]
	Acquisition Electronics	OpenBCI	Custom	WANDmini [15]	Custom	BioGAP-Ultra [7]
	Connectivity	BLE	Bluetooth	BLE	WiFi	BLE
	Power	N.A.	N.A.	46 mW	N.A.	20.5 mW
	Edge AI	No	No	No	No	GAP9
Application	Vocabulary	10 digits (0-9 numbers)	110 words	11 phonemes	10 commands	8 commands + rest
	Speaking Condition	Silent[c]	Silent	Vocalized	Silent	Vocalized + Silent
	#Subjects	10	1	2	10	4
	Dataset Size / Subject	750	1100	110	100[d]	Voc: 2400[e] ; Sil: 2400[e]
	Repositioning	No	No	No	No	Yes
	Model	CNN	LDA	Random Forest	1D SE-ResNet	SpeechNet (CNN)
	Accuracy (%)	92.0	92.6	92.7	96	Voc: 84.8±4.6 Sil: 77.5±6.6[f]

Diff: differential; *CR*: common reference; *Voc*: vocalized; *Sil*: silent.
[a] Results reported using gold-plated electrodes with conductive paste; [b] Four channels fully differential; 10 channels differential with electrode sharing; [c] Authors refer to the speaking condition as internal vocalization; [d] Data augmentation applied during training; [e] Effective dataset size is 4800 samples, including rest; [f] Reported under our *Global Evaluation* setting.

of the recognition task. In addition, our dataset substantially increases the number of recorded samples relative to most prior work, enabling a more robust assessment. These design choices deliberately raise the difficulty of the problem addressed in this work compared to previous studies.

V. Conclusion and Limitations

In this work, we extended the results of [9], where we introduced *Silent-Wear*, a truly wearable SSI based on a textile neckband form factor with on-board signal processing capabilities.

We presented a new dataset[1] collected from four subjects, over three different acquisition sessions spanning three different days, including vocalized and silent EMG recordings of 8 HMI-related commands.

By introducing *SpeechNet*, a tiny (15K parameters) CNN-based architecture for silent and vocalized speech detection, we achieved a cross-validated accuracy of $84.8 \pm 4.6\%$ for vocalized and $77.5 \pm 6.6\%$ for silent speech, showcasing superior performance compared to our prior solution [9]. Furthermore, we conducted an in-depth analysis of the generalization capabilities of our system across multiple days of use, demonstrating an inter-session accuracy of $71.1 \pm 8.3\%$ for vocalized and $59.3 \pm 2.2\%$ for silent speech. To overcome the challenges arising from repositioning, we proposed an incremental fine-tuning strategy, showing the feasibility of more than 10% accuracy recovery with less than 10min. of data recollection. Finally, we deployed our network architecture on the GAP9 MCUs embedded in BioGAP-Ultra, showcasing the feasibility of end-to-end vocalized and silent speech recognition at the edge, enabling low-latency and private speech detection. Signal acquisition, on-device inference, and wireless transmission of the classification results can be executed with as low as $20.5\,\mathrm{mWof}$ power, enabling more than $27.1\,\mathrm{h}$ of continuous use with a $150\,\mathrm{mAh}$ battery.

To ensure widespread use of our solution, promote standardization of data collection methodologies, and enable consistent algorithm performance comparison, we openly release the complete system[3] (hardware, firmware, software, algorithms and datasets).

Despite the promising results, some limitations remain and open directions for future research.

First, the current data collection protocol relies on visual trigger cues to prompt articulation. Consequently, subjective reaction time introduces temporal misalignment between trigger onset and the actual speech-related EMG activity. This delay may affect segmentation accuracy and partially explains why larger window sizes perform better, as they increase the likelihood of fully capturing the activation pattern. Incorporating a microphone as a continuous ground-truth source would allow automatic alignment of EMG signals with speech onset, reduce reaction-time variability, and streamline the acquisition

[3]Hardware and firmware: `https://github.com/pulp-bio/BioGAP`
Data collection software: `https://github.com/pulp-bio/biogui/tree/sensors_speech`
Dataset: `https://huggingface.co/datasets/PulpBio/SilentWear`
Code: `https://github.com/pulp-bio/SilentWear`

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.

process. Such an approach would shorten data collection time by removing unnecessary rest intervals and enable larger datasets to be collected within the same time budget. Faster acquisition would also reduce recalibration time during real-world deployment.

Second, although the proposed dataset includes multi-day recordings and realistic repositioning scenarios, it remains limited in terms of the number of subjects and vocabulary size. Expanding the dataset to include a larger and more diverse participant pool, as well as extending beyond isolated commands toward daily-life sentences and continuous speech [25], would improve validity and enable more comprehensive generalization studies.

Third, the relatively small dataset constrained the exploration of more advanced architectures. With larger-scale data, future work will investigate more advanced models, including pre-trained foundation models based on transformer backbones specifically designed for wearable execution [26], which have already demonstrated strong performance on speech tasks.

Fourth, multimodal sensor fusion represents another promising direction. Integrating complementary sensing modalities such as IMU and EEG, which have been explored for speech decoding [1], may improve robustness to sensor repositioning and physiological variability, ultimately increasing system reliability.

Finally, online system validation remains to be investigated. Evaluating the system under real-time, closed-loop conditions with user interaction and adaptive recalibration will be essential to assess long-term robustness, usability, and readiness for widespread use.

VI. Acknowledgment

The authors thank Mattia Orlandi (University of Bologna) for technical support.

The authors used ChatGPT (OpenAI, GPT-4) for language editing, formatting assistance, code proofreading, and the generation of part of the abstract illustrative figure. All scientific concepts, experimental results, and conclusions were independently developed and validated by the authors.

References

[1] C. Tang, L. Qi, S. Gao, Z. Zhang, W. Yi, M. Xu, E. Occhipinti, Y. Pan, and L. G. Occhipinti, "Sensing technologies for silent speech interfaces," *Nature Sensors*, vol. 1, no. 1, pp. 16–26, 2026.

[2] W. Sun, Z. Guo, Z. Yang, Y. Wu, W. Lan, Y. Liao, X. Wu, and Y. Liu, "A review of recent advances in vital signals monitoring of sports and health via flexible wearable sensors," *Sensors*, vol. 22, no. 20, 2022.

[3] A. T. Chowdhury, M. Newaz, P. Saha, M. N. AbuHaweeleh, S. Mohsen, D. Bushnaq, M. Chabbouh, R. Aljindi, S. Pedersen, and M. E. Chowdhury, "Decoding silent speech: a machine learning perspective on data, methods, and frameworks," *Neural Computing and Applications*, vol. 37, no. 10, pp. 6995–7013, 2025.

[4] P. Kaifosh and T. R. Reardon, "A generic non-invasive neuromotor interface for human-computer interaction," *Nature*, vol. 645, no. 8081, pp. 702–711, 2025.

[5] J. F. Castruita-López, M. Aviles, D. C. Toledo-Pérez, I. Macías-Socarrás, and J. Rodríguez-Reséndiz, "Electromyography signals in embedded systems: A review of processing and classification techniques," *Biomimetics*, vol. 10, no. 3, 2025.

[6] T. A. Najafi, J. A. M. Calero, J. Thevenot, B. Duc, S. Albini, A. Amirshahi, H. Taji, M. J. B. Beneyto, A. Affanni, and D. Atienza, "Versasens: An extendable multimodal platform for next-generation edge-ai wearables," *IEEE Transactions on Circuits and Systems for Artificial Intelligence*, vol. 1, no. 1, pp. 83–96, 2024.

[7] S. Frey, G. Spacone, A. Cossettini, M. Guermandi, P. Schilk, L. Benini, and V. Kartsch, "Biogap-ultra: A modular edge-ai platform for wearable multimodal biosignal acquisition and processing," *IEEE Transactions on Biomedical Circuits and Systems*, pp. 1–17, 2026.

[8] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martín Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE Access*, vol. 8, pp. 177995–178021, 2020.

[9] F. Meier, G. Spacone, S. Frey, L. Benini, and A. Cossettini, "A parallel ultra-low power silent speech interface based on a wearable, fully-dry emg neckband," in *2025 IEEE SENSORS*, pp. 1–4, 2025.

[10] D. Gaddy and D. Klein, "Digital voicing of silent speech," *arXiv preprint arXiv:2010.02960*, 2020.

[11] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, pp. 43–53, 2018.

[12] H. Liu, W. Dong, Y. Li, F. Li, J. Geng, M. Zhu, T. Chen, H. Zhang, L. Sun, and C. Lee, "An epidermal semg tattoo-like patch as a new human–machine interface for patients with loss of voice," *Microsystems & nanoengineering*, vol. 6, no. 1, p. 16, 2020.

[13] Y. Wang, T. Tang, Y. Xu, Y. Bai, L. Yin, G. Li, H. Zhang, H. Liu, and Y. Huang, "All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics," *npj Flexible Electronics*, vol. 5, no. 1, p. 20, 2021.

[14] P. Wu, R. Kaveh, R. Nautiyal, C. Zhang, A. Guo, A. Kachinthaya, T. Mishra, B. Yu, A. W. Black, R. Muller, and G. K. Anumanchipalli, "Towards EMG-to-speech with a necklace form factor."

[15] A. Zhou, S. R. Santacruz, B. C. Johnson, G. Alexandrov, A. Moin, F. L. Burghardt, J. M. Rabaey, J. M. Carmena, and R. Muller, "A wireless and artefact-free 128-channel neuromodulation device for closed-loop stimulation and recording in non-human primates," *Nature biomedical engineering*, vol. 3, no. 1, pp. 15–26, 2019.

[16] 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 multi-channel textile emg sensors integrated into headphones," 2025.

[17] D. S. Inc., "Datwyler softpulse." `https://datwyler.com/company/innovation/softpulse/products`.

[18] I. M. Lang and R. Shaker, "Anatomy and physiology of the upper esophageal sphincter," *The American Journal of Medicine*, vol. 103, no. 5, Supplement 1, pp. 50S–55S, 1997.

[19] C. Puig-Herreros, J. L. Sanz, L. Barona-Lleó, L. Forner, and V. Rosell-Clari, "The characterization of normal male and female voice from surface electromyographic parameters," *Journal of Personalized Medicine*, vol. 14, no. 6, p. 592, 2024.

[20] M. Orlandi, P. M. Rapa, M. Zanghieri, S. Frey, V. Kartsch, L. Benini, and S. Benatti, "Real-time motor unit tracking from semg signals with adaptive ica on a parallel ultra-low power processor," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 18, no. 4, pp. 771–782, 2024.

[21] T. M. Ingolfsson, U. Chakraborty, X. Wang, S. Beniczky, P. Ducouret, S. Benatti, P. Ryvlin, A. Cossettini, and L. Benini, "Epidenet: An energy-efficient approach to seizure detection for embedded systems," in *2023 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–5, IEEE, 2023.

[22] S. Frey, M. A. Lucchini, V. Kartsch, T. M. Ingolfsson, A. H. Bernardi, M. Segessenmann, J. Osieleniec, S. Benatti, L. Benini, and A. Cossettini, "Gapses: Versatile smart glasses for comfortable and fully-dry acquisition and parallel ultra-low-power processing of eeg and eog," *IEEE Transactions on Biomedical Circuits and Systems*, 2024.

[23] B. Obermaier, C. Neuper, C. Guger, and G. Pfurtscheller, "Information transfer rate in a five-classes brain-computer interface," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 9, no. 3, pp. 283–288, 2001.

[24] GreenWaves Technologies, "NN Menu: Neural network menu." `https://github.com/GreenWaves-Technologies/nn_menu_gap9`. Accessed: 2026-02-15.

[25] D. Zhou, Y. Zhang, J. Wu, X. Zhang, L. Xie, and E. Yin, "Ave speech: A comprehensive multimodal dataset for speech recognition integrating audio, visual, and electromyographic signals," *IEEE Transactions on Human-Machine Systems*, 2025.

[26] M. Fasulo, G. Spacone, T. M. Ingolfsson, Y. Li, L. Benini, and A. Cossettini, "Tinymyo: a tiny foundation model for flexible emg signal processing at the edge," *arXiv preprint arXiv:2512.15729*, 2025.

© This work has been submitted to the IEEE for possible publication. Copyright may be transferred without notice, after which this version may no longer be accessible.