Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Efficient Sensor Configuration for Accelerometer-based Silent Speech Interfaces: Regionally Distributed Sensor Placement Strategy Considering Articulatory Dynamics

Heejin Choi
Hanyang University https://orcid.org/0009-0008-8565-4463
Sungmin Jung
Hanyang University https://orcid.org/0009-0007-5478-3011
Jangjay Sohn
Hanyang University https://orcid.org/0000-0001-6095-0831
Chang-Hwan Im
ich@hanyang.ac.kr
Hanyang University https://orcid.org/0000-0003-3795-3318

Research Article

Keywords: Silent Speech Interface, Accelerometer, Sensor Configuration, Deep Learning, Articulatory Dynamics

Posted Date: July 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10442598/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Efficient Sensor Configuration for Accelerometer-based Silent Speech Interfaces: Regionally Distributed Sensor Placement Strategy Considering Articulatory Dynamics

Heejin Choi[1], Sungmin Jung[2], Jangjay Sohn[3], and Chang-Hwan Im[1, 3*]

[1]Department of Electronic Engineering, Hanyang University, Seoul, Republic of Korea
[2]Department of Artificial Intelligence, Hanyang University, Seoul, Republic of Korea
[3]Department of Biomedical Engineering, Hanyang University, Seoul, Republic of Korea
[*]*Corresponding author*

ORCID id
Heejin Choi
0009-0008-8565-4463

Sungmin Jung
0009-0007-5478-3011

Jangjay Sohn
0000-0001-6095-0831

Chang-Hwan Im
0000-0003-3795-3318

*Corresponding author: Chang-Hwan Im
Email: ich@hanyang.ac.kr

Word count: 7,171 words (Introduction–Conclusion)

Abstract

Silent speech interfaces (SSIs) decode articulatory activity without audible speech. Although accelerometers can capture facial and neck movements in compact wearable setups, accelerometer-specific sensor placement has not been systematically examined. This study evaluated all combinations of ten accelerometers across the lower face and neck in a 100 English word silent speech classification task with 20 participants. Subject-dependent classification was performed using a convolutional neural network followed by a bidirectional long short-term memory network. Accuracy improved as the number of sensors increased, reaching a plateau with five sensors. At this point, Sensors 1, 4, 5, 7, and 8 achieved the highest accuracy of 92.30 ± 4.64% and were selected as the efficient configuration. Regional coverage comparisons showed that configurations spanning broader functional articulatory regions outperformed those with more restricted coverage. Together, these results support regionally distributed placement as a design strategy for capturing complementary articulatory information in compact accelerometer-based SSI systems.

Key words –Silent Speech Interface, Accelerometer, Sensor Configuration, Deep Learning, Articulatory Dynamics

1 Introduction

Speech is the most fundamental and universal medium of human communication, enabling the instantaneous and natural exchange of information through speaking and listening (Alsaawi, 2019). However, for individuals with speech impairments, particularly those who have undergone laryngectomy, vocal communication may be severely limited or impossible. Nearly 200,000 new cases of laryngeal cancer are diagnosed worldwide each year (Hoffmann, 2021; Ritter et al., 2023). For patients with advanced laryngeal cancer, total laryngectomy may be required, after which many patients rely on written notes, gestures, or assistive devices for communication. These communication methods can limit real-time interaction and reduce social participation (Sharpe et al., 2019). Beyond clinical needs, spoken communication may also be difficult or inappropriate in noisy environments, settings where silence is required, or security contexts where audible speech can compromise intelligibility or confidentiality (Li et al., 2025; Yasmin et al., 2023). To address these limitations, silent speech interfaces (SSIs) have emerged as a promising approach that enables direct communication without relying on acoustic speech signals (Denby et al., 2010).

Research on SSIs has primarily focused on detecting articulatory activity during silent speech. To this end, a range of sensing modalities have been explored to capture non-vocalized signals, including computer vision for lip reading (Sun et al., 2018), ultrasound imaging of the tongue (Hueber et al., 2010), electromagnetic articulography (EMA) (Cao et al., 2023), and surface electromyography (sEMG) of facial muscles (Meltzner et al., 2018). For example, Sun et al. (2018) used a mobile camera and deep learning to classify lip movements into predefined commands. Hueber et al. (2010) reconstructed speech from ultrasound tongue images and optical lip videos using a segmental vocoder. Cao et al. (2023) tracked three-dimensional tongue kinematics in real time using a wearable eyeglasses frame for magnetic field generation and a tongue-mounted inertial measurement unit (IMU). Meltzner et al. (2018) recognized silently spoken words and phrases by measuring muscle activity in the jaw, cheek, and neck regions with sEMG electrodes. More recently, researchers have explored advanced sensing modalities to capture subtle articulatory activity. Kim et al. (2022), for instance, used ultra-thin silicon strain sensors to measure micro-deformations of the skin surface during speech, while Ferreira et al. (2022) employed frequency-modulated continuous-wave (FMCW) radar to monitor articulatory displacement without physical contact. Although valuable for measuring articulatory activity during silent speech, these modalities often face practical limitations, such as bulky hardware, setup complexity, limited durability, or sensitivity to environmental conditions.

A recent study by Kwon et al. (2023) introduced three-axis accelerometers as a novel modality for capturing speech-related facial dynamics. Specifically, sensors attached to multiple facial sites record acceleration signals that reflect facial surface movements during silent articulation. Prior work showed that only four accelerometers placed on the lower face could classify 40 silently spoken words with 95.58 ± 1.83% accuracy. In the same 40-word classification setting, the accelerometer-based system significantly outperformed an sEMG-based method, which achieved 89.68 ± 5.27% accuracy. Moreover, its classification accuracy was comparable to those reported for prior radar- and video-based SSI approaches, even though the classification task involved a relatively larger vocabulary. In addition to this competitive performance, accelerometers offer practical advantages such as a compact form factor, lightweight design, and high durability, making them well suited for SSI applications.

Prior accelerometer-based SSI studies have shown that sensor placement can substantially affect recognition performance. In the previous work (Kwon et al., 2023), five accelerometer positions were initially selected, but the highest recognition accuracy was achieved with a four-sensor configuration that excluded the zygomatic site (Channel 1 in the original work). Notably, these sensor locations were determined based on previous studies of sEMG-based SSIs, which measure the myoelectric activity of articulatory muscles. However, because sEMG and accelerometers rely on fundamentally different measurement principles—sEMG captures electrical potential, whereas accelerometers capture motion-induced acceleration signals (Gómez-Vilda et al., 2019)—sensor placements effective for one modality may not be optimal for the other. The exclusion of the zygomatic site illustrates this mismatch. Although the zygomaticus muscles are known to show substantial activation during speech (Adamov et al., 2025) and have therefore been frequently used as recording sites in sEMG-based SSIs (Salomons et al., 2025), speech-related surface motion in this region may be limited because it is relatively distant from the primary articulators. A subsequent accelerometer-based speech synthesis study showed that including an under-chin accelerometer in the previously optimized four-channel configuration significantly improved synthesis quality (Kwon et al., 2024). Together, these findings suggest that sensor placement should be carefully considered in accelerometer-based SSIs because it affects the quality of speech-related motion signals. Despite this importance, recent accelerometer-based SSI studies have mainly sought performance improvements through signal processing, data augmentation, and deep learning methods (Liu et al., 2024; Xie et al., 2026), leaving sensor placement optimization largely unexplored.

From a practical perspective, systematic evaluation of sensor configurations is also needed, as both sEMG-based and accelerometer-based SSIs require sensing elements to be attached to the face and neck. Increasing the number of attachment sites can increase user discomfort, prolong setup time, and require larger or more complex hardware for sensor fixation. A compact set of informative sensing sites is essential for practical SSI design because it can reduce the burden of attachment while preserving recognition performance. For these reasons, electrode configuration optimization has been extensively studied in sEMG-based SSIs to improve system efficiency. Zhu et al. (2021) used sequential forward selection (SFS) to identify optimal electrode configurations from 120 high-density electrodes placed over the facial and neck muscles for silent classification of ten digits in both English and Chinese. Their results showed that ten optimized electrodes achieved classification accuracy comparable to that of 40 non-optimized electrodes. Similarly, Zhu et al. (2024) developed a lightweight sEMG acquisition system with flexible array electrodes and optimized electrode placement using a genetic algorithm (GA), showing that 11 selected channels outperformed the full 32-channel configuration. Together, these studies demonstrate that identifying a compact subset of electrodes can maintain or even improve performance while enhancing practical efficiency. These sEMG studies highlight the value of sensor configuration optimization, yet comparable evidence for systematic sensor-configuration evaluation in accelerometer-based SSIs remains limited.

In this study, we systematically evaluated sensor configurations for accelerometer-based SSIs in a 100-word silent speech classification task. To capture a broader range of speech-induced motion, ten three-axis accelerometers were distributed across the lower face and neck, grouped into three functional articulatory regions—the labial, buccal-mandibular, and submental regions—according to the dominant articulatory dynamics they captured. We first investigated how many accelerometers are needed to maintain reliable

classification performance while reducing sensor burden, by exhaustively evaluating all possible sensor combinations. We then examined how the regional coverage of sensor configurations affected classification performance by comparing configurations that covered different numbers of functional articulatory regions. To the best of our knowledge, this is the first study to systematically investigate both sensor count and placement strategy in accelerometer-based SSIs, where such design guidance has been lacking. Our findings provide accelerometer-specific evidence on sensor placement and practical guidance for efficient sensor configuration, serving as a design basis for future wearable accelerometer-based SSI systems.

2 Methods

2.1 Data Acquisition System

2.1.1 Hardware System

A total of ten IMUs were positioned across the lower face and neck to capture articulatory dynamics. Each IMU was an MPU-9250 sensor (TDK InvenSense, CA, USA), a 9-axis module integrating a 3-axis accelerometer, gyroscope, and magnetometer. Following (Kwon et al., 2023), only the 3-axis acceleration signals were analysed in this study. The ten sensors were connected to two shoulder-mounted acquisition units, with Sensors 1–5 connected to one unit and Sensors 6–10 connected to the other (Figure 1a and 1b). To account for possible facial asymmetry, the side of the unit connected to Sensors 1–5 was randomized across participants, resulting in a left–right reversal of the entire sensor layout. In ten participants, Sensors 1–5 were placed on the left side and Sensors 6–10 on the right side, whereas this arrangement was reversed in the remaining ten participants. Each device aggregated data from five IMUs through an I2C multiplexer (TCA9548A, Texas Instruments, Dallas, TX, USA) and transmitted the signals to an ESP32 microcontroller (Espressif Systems, Shanghai, China) via I2C communication. The ESP32 then transmitted the collected data to a PC via a Wi-Fi-based protocol, where the signals were monitored and recorded using a custom MATLAB (MathWorks, Natick, MA, USA) graphical user interface (GUI). At the same time, the PC received synchronization triggers from Unity (Unity Technologies, San Francisco, CA, USA) to mark the onset and offset of each silent utterance as well as the corresponding word labels. The overall data acquisition and communication architecture is illustrated in Figure 1c.

2.1.2 Locations of Sensors and Functional Articulatory Regions

Sensor locations were selected to maximize the capture of speech-related facial movements based on the kinematic characteristics of accelerometers, rather than the conventions of sEMG-based research. Unlike sEMG, which detects electrical activity in specific muscles, accelerometers measure physical displacement. Facial and neck surface movements during speech arise not only from muscle activation but also from rotational and translational movements of the articulators. Accordingly, sensor placement was guided by prior studies on visual speech animation, facial motion analysis, and neck motion tracking. Visual speech animation studies

have shown that lower-face motion, particularly around the nose, lips, and chin, is strongly related to speech production (Xie et al., 2016). Nakamura et al. (2022) further demonstrated that incorporating chin movement features in addition to lip movement features significantly improved speech estimation performance. In addition, motion capture studies of the anterior neck have highlighted the importance of vertical laryngeal movement between the chin and clavicle, as well as activity in the lateral extrinsic laryngeal muscles near the base of the tongue (Hogue et al., 2023). Based on these findings, we selected ten sensor locations to cover major articulatory areas, including the lip and jawline regions, the cheek area, and the neck. Figure 2a shows the locations of the ten accelerometers attached to the face and neck, and the anatomical placement of each sensor is summarized in Table 1.

The ten accelerometer locations were further grouped into three functional articulatory regions—the labial, buccal-mandibular, and submental regions—as shown in Figure 2b. Because accelerometers measure physical displacement, this categorization provided a systematic framework for analysing sensor placement in terms of articulatory dynamics. This grouping also extended the analysis from individual sensors to the regional level, allowing us to examine how different articulatory regions contributed, both individually and in combination, to recognition performance.

- Labial Region: Sensors 1, 2, and 7 were assigned to this region because they were positioned near the lips and were intended to capture direct labial movements (Nakamura et al., 2022; Xie et al., 2016).

- Buccal-Mandibular Region: Sensors 3, 4, 6, and 8 were grouped into this region because they were located along the jawline (Nakamura et al., 2022; Xie et al., 2016) or in the buccal area, where surface movements may reflect mandibular motion. Although Sensor 6 was positioned on the cheek, it was included in this region because mandibular opening and closing have been associated with significant soft-tissue displacements on the facial surface (Chen et al., 2011).

- Submental Region: Sensors 5, 9, and 10 were classified into this region. These sensors were positioned on the neck and upper throat to capture motion associated with tongue activity and structural changes in the vocal tract (Hogue et al., 2023).

2.2 Data Collection

2.2.1 Participants

Twenty native Korean-speaking adults familiar with English participated in the experiment (10 males and 10 females; age: 24 ± 2.62 years). None reported a history of neurological, psychiatric, or other serious conditions that could affect the results. Individuals with a history of facial surgery or orthodontic treatment that could alter facial skin displacement were excluded. Prior to the experiment, all participants were provided with detailed information about the procedure, and written informed consent was obtained from each participant. The study protocol

was approved by the Institutional Review Board (IRB) of Hanyang University, Republic of Korea (IRB No. HYU-202408-007-3), and all procedures were conducted in accordance with the Declaration of Helsinki.

2.2.2 Experimental Procedure

Participants silently spoke each of the 100 English words once per session (Appendix 1), with the words presented in a randomized order within each session. The same procedure was repeated across five sessions. The word list was composed of high-frequency terms commonly used in user-friendly interfaces. To ensure phonetic diversity, words were selected to minimize vowel overlap. When vowels were shared, words with clearly different consonants were chosen to maximize articulatory contrast.

During the experiment, participants sat in a stationary chair facing a monitor and viewed target words through a Unity-based interface (Figure 3). Each trial consisted of a 5 s recording period, during which participants silently articulated the displayed word within a 4 s articulation window. This window was preceded and followed by 0.5 s pre- and post-buffer periods, and its onset and offset were indicated by auditory beeps. Participants were instructed to avoid non-speech movements, such as swallowing or nodding, during recording. If they completed articulation before the end beep, they were asked to return to a neutral facial expression and remain still until the recording ended to maintain a stable baseline.

2.3 Data Processing and Classification

2.3.1 Data Preprocessing

Raw data were collected for 5 s per trial at a sampling rate of 100 Hz, including the 4 s articulatory window and 0.5 s buffers before the start beep and after the end beep. Only the three-axis acceleration signals (x, y, and z) were retained, and, following the previous study (Kwon et al., 2023), the data were downsampled to 50 Hz. For subsequent analysis, only the 4 s articulatory segment (200 samples) was used after the 0.5 s pre- and post-buffers (25 samples each) were removed. After trial-level preprocessing, each trial was represented as a 30 × 200 matrix, comprising 30 channels from ten sensors with three-axis acceleration signals and 200 time samples. Z-score normalization was then applied independently to each channel along the time axis within each trial before the data were input to the model.

2.3.2 Model Architecture, Training, and Evaluation

We adopted the deep learning architecture proposed in the previous study (Kwon et al., 2023) to classify 100 silently spoken words using three-axis accelerometer signals. As shown in Figure 4, the model consisted of a Convolutional Neural Network (CNN) followed by a Bidirectional Long Short-Term Memory (BiLSTM) network. The normalized acceleration signals were first passed through four CNN blocks to extract local temporal features. The extracted features were then passed to a BiLSTM layer, which modeled temporal dependencies in both forward and backward directions. Finally, global average pooling and a dense layer with softmax activation

were used to produce probabilities for the 100 word classes. Dropout with a rate of 0.5 was applied to reduce overfitting.

The model was trained to minimize the categorical cross-entropy loss using the Adam optimizer (Kingma & Ba, 2015). The random seed was fixed at 0, and the hyperparameters were empirically determined according to the previous study (Kwon et al., 2023), as summarized in Table 2. Here, β_1 and β_2 represent the exponential decay rates for the first- and second-moment estimates, respectively. All implementation and training procedures were carried out in Keras with TensorFlow as the backend. Training was accelerated using an NVIDIA RTX 4070 Ti GPU.

For evaluation, subject-specific models were trained and tested using the five recorded sessions for each participant. In each fold, three sessions were used for training, and the remaining two sessions were used for testing. All possible combinations of three training sessions selected from five sessions were considered, resulting in 10 folds. The final accuracy for each participant was computed as the average test accuracy across the 10 folds.

2.4 Sensor Configuration Analysis

2.4.1 Selection of the Efficient Sensor Configuration

An efficient sensor configuration should balance recognition accuracy and hardware complexity, achieving strong performance with as few sensors as possible. Accordingly, the primary goal was to identify the minimum number of sensors required to maintain reliable classification performance. To determine the appropriate number of sensors, models were trained for all possible combinations of the ten sensors (2^{10}- 1 = 1,023 combinations). Throughout this paper, sensor configurations are denoted in the text by the included sensor numbers in braces; for example, {7, 8} indicates a configuration using Sensors 7 and 8. For each sensor combination, classification accuracy was computed for each participant and then averaged across participants. To summarize the overall effect of sensor count, configurations with the same number of sensors were grouped together, and the mean accuracy of all configurations within each sensor-count group was calculated. For statistical comparison, the highest-performing configuration was identified within each sensor-count group. The subject-wise accuracies of these configurations were compared across sensor-count groups using the Wilcoxon signed-rank test with FDR correction. The saturation point was defined as the smallest number of sensors beyond which adding more sensors no longer produced a statistically significant increase in accuracy. The configuration with the highest accuracy at this saturation point was selected as the efficient sensor configuration.

2.4.2 Evaluation of Regional Coverage in Sensor Configurations

To evaluate the effect of regional coverage, sensor configurations were grouped by the number of functional articulatory regions they covered. Three regional coverage levels were defined: single-region, two-region, and three-region placement. These levels indicated whether a configuration included sensors from one, two, or all three functional articulatory regions,

respectively. This grouping was used to compare classification performance between configurations limited to a single articulatory region and those covering multiple articulatory regions.

First, 3-sensor configurations were used to compare single-region, two-region, and three-region placement under the same fixed sensor-count condition. Three sensors were chosen because this was the minimum sensor count required to include sensors from all three functional articulatory regions. At the same time, because each region contained either three or four sensor locations, the three-sensor condition was also the maximum fixed sensor count that allowed single-region configurations to be formed in every region. Among the 120 possible 3-sensor configurations, the placement groups were single-region (n = 6), two-region (n = 78), and three-region (n = 36). For each regional coverage level, the highest-performing configuration was identified using the mean accuracy across participants. The participant-wise accuracies of these configurations were then compared using the Wilcoxon signed-rank test with FDR correction.

Second, following the fixed 3-sensor comparison, regional coverage was examined at the saturation-point sensor count identified in the preceding sensor-count analysis. At this sensor count, single-region configurations were not possible because the sensor count exceeded the maximum number of sensor locations within any single functional articulatory region. Therefore, the comparison was made between full-coverage configurations, which included sensors from all three functional articulatory regions, and partial-coverage configurations, which included sensors from two regions. For the full- and partial-coverage groups, the highest-performing configuration was selected using the mean accuracy across participants, and the participant-wise accuracies of the selected configurations were compared using the Wilcoxon signed-rank test.

3 Results

3.1 Selection of the Efficient Sensor Configuration

Sensor configuration {1, 2, 4, 5, 7, 8, 9, 10} achieved the overall highest accuracy of 92.56 ± 4.77%. Figure 5 shows the average classification accuracy for each sensor count, computed across all possible configurations. As the number of sensors increased from 1 to 10, the mean classification accuracy steadily rose from 75.33% to 92.46%. An accuracy of over 80% was achieved with only two sensors, and performance exceeded 90% with five sensors. Although the average accuracy continued to improve as more sensors were used, the increase in accuracy became smaller at higher sensor counts. In addition, performance variability across configurations decreased as the number of sensors increased. This reduction in variability suggests that compact sensor sets may be more sensitive to which sensors are selected, whereas larger sensor sets tend to yield more consistent performance across configurations.

The saturation point was statistically identified using the best-performing configuration at each sensor count. Figure 6a shows the highest accuracy at each sensor count, and Table 3 lists the corresponding sensor combinations grouped by functional articulatory region. Pairwise Wilcoxon signed-rank tests with FDR correction were then performed across sensor counts, as summarized in

Figure 6b. Significant differences compared with higher sensor counts were still found up to the four-sensor configuration. In contrast, the five-sensor configuration did not differ significantly from any best-performing configuration with a higher sensor count ($p > 0.05$), indicating a performance plateau from five sensors onward. Consistent with this plateau, the highest accuracy at each sensor count did not always increase as more sensors were added and, in some cases, decreased. Accordingly, the saturation point was determined to be five sensors, beyond which additional sensors provided no statistically significant improvement in accuracy.

At the five-sensor saturation point, the configuration {1, 4, 5, 7, 8} achieved the highest accuracy (92.30 ± 4.64%) and was selected as the most efficient configuration. This configuration showed no statistically significant difference from the overall highest-accuracy configuration, {1, 2, 4, 5, 7, 8, 9, 10} ($p = 0.263$). Appendix 2 illustrates the average classification accuracies for each participant when using the efficient configuration. Classification accuracy was above 80% for all participants and exceeded 90% for 13 of the 20 participants. Inter-subject variability was relatively low for the efficient configuration, with the standard deviation of classification accuracy across participants ranking within the lowest 1% among all possible sensor configurations. The sensor locations of this configuration are shown in Figure 7, with colors indicating the corresponding functional articulatory regions. Notably, the efficient configuration included sensors from all three articulatory regions—labial (Sensors 1 and 7), buccal-mandibular (Sensors 4 and 8), and submental (Sensor 5).

Furthermore, to illustrate its practical applicability, we also recorded a video demonstrating real-time classification of all 100 silently spoken words using the efficient configuration. The video demonstration is available online at https://youtube/LWrmnXyDMaY. Three sessions of accelerometer data for the 100 words were first collected to train the model. Accelerometer signals for each word were fed into the trained model, which predicted the silently spoken word. The system then displayed the predicted label and cumulative classification accuracy in real time. During real-time operation, the system showed stable performance, supporting the feasibility of maintaining high classification accuracy with a minimal sensor set in a practical SSI setting.

3.2 Effect of Regional Coverage on Classification Performance

In the fixed 3-sensor comparison, the highest-performing configurations were {1, 2, 7} for the single-region group (88.24 ± 6.08%), {1, 5, 7} for the two-region group (90.95 ± 5.33%), and {1, 4, 5} for the three-region group (90.98 ± 4.99%). Both multi-region configurations significantly outperformed the single-region configuration, with significant differences between the single-region and two-region configurations ($p = 0.0003$) and between the single-region and three-region configurations ($p = 0.0004$). The two multi-region configurations did not differ significantly from each other ($p = 0.75$). Overall, configurations spanning multiple functional articulatory regions achieved higher classification performance than configurations restricted to a single region.

At the five-sensor saturation point, the highest-accuracy full-coverage configuration was the efficient configuration {1, 4, 5, 7, 8} (92.30 ± 4.64%). This configuration included Sensors 1 and 7 in the labial region, Sensors 4 and 8 in the buccal-mandibular region, and Sensor 5 in the submental region.

The highest-accuracy partial-coverage configuration was {1, 2, 5, 7, 9} (91.29 ± 5.08%). This configuration included Sensors 1, 2, and 7 in the labial region and Sensors 5 and 9 in the submental region. The full-coverage configuration showed significantly higher accuracy than the partial-coverage configuration (p = 0.0009).

4 Discussion

Based on these results, the five-sensor configuration {1, 4, 5, 7, 8} was selected as the efficient sensor configuration for accelerometer-based SSIs. This configuration used the minimum number of sensors beyond which adding more sensors no longer yielded significant accuracy gains, while achieving accuracy comparable to that of the overall highest-performing configuration. The selected configuration also included sensors from the labial, buccal-mandibular, and submental regions. Considered alongside the regional coverage analyses, this sensor composition suggests that classification performance was related not only to sensor count but also to the distribution of sensors across functional articulatory regions. To our knowledge, this study is the first to systematically examine accelerometer-specific sensor placement for SSIs. In the following sections, we first interpret the selected sensor locations from a sensor-level perspective. The effect of regional coverage on classification performance is then examined using region-level performance and ablation analyses. Finally, these findings are integrated into a regionally distributed placement strategy for compact and practical accelerometer-based SSI design.

To begin the sensor-level analysis, we evaluated the standalone classification accuracy of each sensor location. Figure 8a shows the results obtained when each sensor was used independently. Sensor 7, located below the lower lip, achieved the highest individual performance (85.69 ± 7.67%), followed by Sensor 4 (81.47 ± 8.45%), Sensor 8 (79.44 ± 10.32%), Sensor 5 (77.32 ± 9.38%), and Sensor 1 (76.76 ± 9.88%). Notably, all five sensors in the efficient configuration ranked among the top five in individual performance. This indicates that these sensor locations contained relatively more discriminative information for distinguishing the 100 silently articulated words. In particular, the strong performance of Sensor 7 may be attributed to its proximity to both the lower lip and the chin, allowing it to capture movement patterns associated with both labial and mandibular articulation.

We next assessed each sensor's contribution within the full sensor set by estimating how strongly the trained model depended on information from each location. For this purpose, perturbation-based explainability (Ivanovs et al., 2021) was applied (Figure 8b). The model was first trained using data from all ten sensors, and perturbation was applied only during inference. Specifically, one sensor input was perturbed by randomly shuffling its data along the time axis, thereby disrupting the temporal structure of the signal. The perturbed input was then fed into the model, and the resulting drop in accuracy was calculated relative to that obtained with the original input (Sood & Craven, 2022). This performance degradation reflects the model's sensitivity to each sensor, indicating the extent to which the model relied on information from that location. A large perturbation effect may suggest two related possibilities:

1) The sensor provides discriminative information that is particularly important for word recognition.

2) The remaining sensors may not fully preserve the model's performance when the signal from that location is disrupted.

When each sensor input was perturbed, Sensor 1 yielded the largest decrease in performance (12.25 ± 4.69%p), followed by Sensors 7 (8.50 ± 3.67%p), 8 (6.82 ± 2.81%p), 4 (6.64 ± 3.82%p), and 5 (6.20 ± 3.19%p). The five sensors selected for the efficient configuration also ranked among the top five in perturbation effect. The signals from these locations appear to have played particularly influential roles in word classification within the full sensor set. Sensor 1, located at the philtrum, may have reflected subtle upper-lip deformations involved in shaping labial articulation, which could explain its high perturbation sensitivity.

An ablation study was then conducted to examine the uniqueness or redundancy of the information provided by each sensor (Table 4). Unlike the perturbation-based analysis, a separate model was trained from scratch for each nine-sensor subset, with one sensor excluded at a time. This approach allowed us to assess whether the remaining nine sensors could compensate for the excluded sensor, or whether the excluded sensor provided information that could not be fully recovered from the others. The values in Table 4 represent the difference in mean classification accuracy, calculated by subtracting the accuracy of each 9-sensor subset from that of the full 10-sensor configuration. Sensors included in the efficient configuration are indicated in bold. The largest performance decrease was observed for Sensor 1 (0.93%p), followed by Sensors 7, 4, 5, and 8. The five sensors in the efficient configuration again ranked among the top five in terms of performance degradation after sensor removal. This indicates that information from these locations was less readily substituted by the remaining sensors. However, the accuracy drops themselves were small overall, with all drops below 1%p. This likely reflects the fact that the model was trained without the excluded sensor from the beginning, allowing the remaining sensors to adaptively support classification. These small differences are also consistent with Figure 5, where performance variability across configurations decreased as the number of sensors increased. Nevertheless, Sensor 1 showed the largest accuracy drop among all locations, suggesting that the philtrum region may capture specific upper-lip movement characteristics that were not fully represented by the other sensor locations. In contrast, removing Sensors 3 and 6 resulted in a slight performance improvement, as these locations may have introduced redundant or noisy information into the model. Taken together, the three sensor-level analyses show that Sensors 1, 4, 5, 7, and 8 consistently ranked among the most informative and influential locations, helping to explain why these locations formed the efficient configuration.

Across both the three- and five-sensor comparisons, multi-region configurations consistently outperformed those with more restricted regional coverage. At the fixed three-sensor count, both two-region and three-region configurations significantly outperformed the single-region configuration, although they did not differ significantly from each other. This may be partly explained by the position of Sensor 7 near both the lower lip and chin, which may have allowed the two-region configuration to capture some jaw-related movement information despite the absence of a sensor from the buccal-mandibular region. In contrast, at the five-sensor saturation point, the full-coverage configuration significantly outperformed the best partial-coverage configuration, revealing a clearer advantage of full regional coverage. This advantage can be understood in relation to the articulatory diversity involved in word classification. In a large and

phonetically diverse classification task such as the present 100-word set, the words are likely to differ along multiple articulatory dimensions. Some distinctions may be expressed mainly through lip-related movements, whereas others may require jaw motion or tongue- and vocal-tract-related activity. A single movement source may not provide sufficient information across all word contrasts. Combining information from multiple sources is consequently important for covering a broader range of articulatory patterns in a large word set. The three functional articulatory regions defined in this study—labial, buccal-mandibular, and submental—each correspond to a distinct movement source, reflecting lip-, jaw-, and tongue-related activity, respectively. Distributing sensors across these regions therefore enables complementary articulatory information to be captured jointly.

This interpretation assumes that the three regions provide distinct and only partially overlapping articulatory information. To evaluate this assumption, we considered regional performance and region-level ablation results. Table 5 summarizes the classification performance of the three articulatory regions. The values represent classification accuracy obtained using models trained only with sensors from the corresponding region. Because the buccal-mandibular region contained four sensors, the average accuracy across all possible three-sensor subsets was also calculated for comparison. The labial region showed the highest classification accuracy (88.24 ± 6.08%), followed by the buccal-mandibular region (86.87 ± 7.13%) and the submental region (81.92 ± 8.96%). The same ranking was observed when the three-sensor average of the buccal-mandibular region was used. This implies that the three regions differed in both the amount and discriminability of the articulatory information they provided, with lip-related movements likely offering richer and more distinctive information for word recognition. However, standalone regional performance does not reveal whether information from one region is redundant with or complementary to information from other regions. Table 6 presents the accuracy drop when all sensors from each region were removed from either the full sensor set or the efficient configuration. The accuracy drops were generally larger in the efficient configuration, possibly because its reduced sensor count left fewer remaining sensors to compensate for the removed region. Across both configurations, removing the labial region caused the largest decrease in accuracy, consistent with its high standalone regional performance. This suggests that lip-related movements provided discriminative information that was not fully captured by the other regions. The buccal-mandibular region showed a slightly smaller ablation effect than the submental region, despite its higher standalone regional performance. This may be related to the characteristics of mandibular movement. Jaw dynamics directly reflect mouth opening, which is particularly relevant to vowel articulation, and can therefore provide substantial discriminative information for word recognition (Maeda et al., 2002; Vatikiotis-Bateson et al., 1996). At the same time, because mandibular movements are relatively large, part of this information may also be captured by other regions, making it more readily substitutable (Vatikiotis-Bateson et al., 1996). In contrast, within the efficient configuration, removing the submental region involved excluding only Sensor 5, yet it produced a larger accuracy drop than removing the two buccal-mandibular sensors. This indicates that submental information may have been less readily replaced by the remaining regions, as submental movements may reflect region-specific information related to tongue activity or vocal-tract changes.

The sensor-level analyses identified the five sensors in the efficient configuration—Sensors 1, 4, 5, 7, and 8—as important locations across standalone performance, perturbation, and ablation analyses. However,

configuration-level performance could not be explained solely by the characteristics of individual sensors. Simply combining sensors that ranked highly in sensor-level analyses did not necessarily produce the highest-performing configuration. For example, although {7, 8} was the highest-accuracy two-sensor configuration, adding another high-performing sensor to this pair did not yield the best three-sensor configuration. Instead, the highest-accuracy three-sensor configuration was {1, 4, 5}, which covered all three functional articulatory regions: Sensor 1 in the labial region, Sensor 4 in the buccal-mandibular region, and Sensor 5 in the submental region. Thus, configuration performance may depend not only on the quality of individual sensors but also on how their information is combined. The regional coverage and region-level analyses provide a possible explanation for this configuration-level pattern. Configurations spanning multiple articulatory regions achieved higher classification performance because they could combine movement information from different articulatory sources. The region-level analyses further showed that the three regions differed in both performance and the extent to which their information overlapped. From this perspective, efficient placement should prioritize distributed coverage that captures complementary articulatory information, rather than concentrating sensors within a limited region. Accordingly, we propose a regionally distributed placement strategy that places sensors across regions associated with different articulatory movement sources, such as labial, buccal-mandibular, and submental movements, to capture complementary articulatory information. Following this design principle, the efficient configuration {1, 4, 5, 7, 8} combined individually important sensor locations with broad regional coverage, achieving high classification performance with a reduced number of sensors.

The regionally distributed placement strategy proposed above was derived from a specific 100-word vocabulary. However, silent speech recognition performance can depend strongly on lexical composition, because the articulatory movements required to produce a given vocabulary may not be evenly represented. Certain articulatory patterns may be overrepresented, underrepresented, or absent in a given vocabulary, potentially influencing which sensor locations appear most informative. Findings derived from a single vocabulary may not generalize to word sets with different articulatory compositions. We therefore examined whether the proposed strategy remained valid across different vocabularies. Another 100 English words (Appendix 3) were selected based on the New General Service List 0–2000 high-frequency word list (Browne et al., 2013) and their semantic suitability for frequent use in user interfaces. Vowel overlap was minimized as much as possible, and when it could not be avoided, consonantal differences were maximized. To examine the composition of the two word sets from the perspective of articulatory movements, visemes were introduced as units representing visually distinguishable speech movements (Lee & Yook, 2002). Using the viseme indices in Appendix 4, we compared the original and new word sets in terms of their viseme distributions. As shown in Appendix 5, both word sets covered all viseme indices and exhibited broadly similar distributions. The new word set was not substantially biased toward specific articulatory patterns. The validation experiments were therefore conducted with five participants using the same paradigm and device, with all conditions kept identical except for the word set. For all possible combinations of the ten sensors, performance was evaluated using 10-fold cross-validation and averaged across subjects. Figure 9 summarizes the classification performance for the new word set, including the average accuracy across all possible configurations (Figure 9a) and the highest-performing configuration at each sensor count (Figure 9b). A similar saturation pattern was observed: as the number of sensors increased, performance improved,

but the gains gradually diminished. Despite the similar saturation pattern, the highest-performing sensor configurations differed across the two word sets (Table 7). In particular, for the five-sensor configurations, four sensors were retained, whereas Sensor 4 in the original word set was replaced by Sensor 10 in the new word set, yielding {1, 5, 7, 8, 10} with an accuracy of 92.18 ± 3.24%. Although the specific sensor identities differed between the two word sets, the highest-performing configurations with three or more sensors consistently included sensors from all three articulatory regions in both word sets. These results provide additional evidence that broad regional coverage contributes to robust performance across different word sets. Thus, regionally distributed placement may serve as a generalizable design strategy for accelerometer-based SSI systems.

For practical use, further reducing the number of sensors may improve user comfort and ease of attachment. We therefore considered a sensor configuration that was more compact than the efficient five-sensor configuration while preserving the proposed regionally distributed placement strategy. To implement broad articulatory coverage with the fewest possible sensors, one representative sensor was selected from each functional articulatory region based on practical wearability and word recognition performance. In the labial region, Sensor 1, located at the philtrum, was excluded because it may cause discomfort during extended wear and is more prone to detachment. Among the remaining labial sensors, Sensor 7 was selected for its superior individual performance. In the buccal-mandibular region, Sensor 8 was chosen to ensure greater spatial diversity, as Sensor 4 is located close to Sensor 7 and was therefore likely to capture similar movements. In the submental region, Sensor 5 was selected because it was the only submental sensor included in the efficient configuration and also showed better performance than the other sensors in the same region. Taken together, the configuration {5, 7, 8} was identified as a practical reduced-sensor candidate. This combination achieved nearly 90% word recognition accuracy in both the original word set (90.14 ± 6.15%) and the new word set (89.82 ± 3.50%). In addition, as shown in Figure 10, the selected sensor locations were clustered around the chin and upper neck, suggesting the feasibility of developing a mask-type wearable SSI device.

Two questions remain to be addressed. First, given that the efficient configuration was identified from subject-averaged performance, its relationship to subject-specific optimal configurations should be examined. Second, because the same model parameters were applied across sensor counts despite changes in input dimensionality, it is important to determine whether the observed performance trends remain consistent across different parameter settings. Figure 11 compares, for each subject, the performance difference between the subject-specific optimal configuration and the efficient configuration {1, 4, 5, 7, 8}. For more than half of the participants (12 of 20), no significant difference was observed between their optimal configuration and the efficient configuration, showing that the efficient configuration achieved comparable performance for many participants. In addition, Sensors 1, 4, 5, 7, and 8 were frequently included in the subject-specific optimal configurations. Such subject-specific differences may be expected due to individual variability in sensor attachment and articulatory movement. However, practical wearable deployment requires a unified sensor configuration that can be applied across users rather than individually personalized. From this perspective, identifying an efficient and generalizable sensor configuration is meaningful for real-world use. Next, the effect of parameter settings, particularly the number of filters, was assessed (Appendix 6). Following the setting used in the previous study (Kwon et al., 2023), 250 filters were

initially employed in the 1D convolution layer, yielding 250 feature maps regardless of the input dimensionality determined by the number of sensors (Figure 4). To investigate whether this fixed-parameter setting affected the main findings, models were additionally trained with different numbers of filters. Because of computational cost, three configurations were randomly selected for each sensor count. For each sensor count, performance differences across filter settings were first evaluated using a Friedman test. No significant effect of filter number was observed for the single-sensor condition (p = 0.0743), whereas all other sensor counts showed significant differences. Post hoc comparisons were then conducted using Wilcoxon signed-rank tests with FDR correction. These analyses showed significant performance differences for almost all pairwise comparisons, except between 375 and 500 filters in the 3-sensor condition. Despite these differences in absolute accuracy, the same overall trend was observed across filter settings: accuracy increased with the number of sensors and then gradually reached saturation. This indicates that, although the number of filters affected classification accuracy, the relationship between sensor count and performance remained stable. Therefore, the main findings of the present study appear robust to variations in the number of filters.

While the present findings were obtained in subject-dependent settings, several directions remain for future work. Because subject-independent modelling remains challenging due to inter-subject variability (Jung et al., 2026), as also reflected in Figure 11, extending this placement strategy to subject-independent settings will be an important next step. In addition, future work should investigate implementation in a wearable device together with the development of a lightweight model for on-device inference. Furthermore, this placement strategy may support future accelerometer-based SSI applications beyond word recognition, including real-time continuous classification and speech synthesis.

5 Conclusion

This study systematically investigated sensor placement for accelerometer-based SSIs by evaluating all possible combinations of ten accelerometers placed across the lower face and neck. We identified the sensor count at which performance saturated and selected the highest-performing configuration at that point as the efficient configuration. Among the five-sensor configurations at the saturation point, {1, 4, 5, 7, 8} achieved the highest accuracy, with performance comparable to the highest observed accuracy across all sensor counts. In addition, regional coverage analyses consistently showed that configurations covering broader functional articulatory regions outperformed those with more restricted coverage.

The main contribution of this study lies in shifting the focus of accelerometer-based SSI research from algorithmic improvement to sensor placement design. Previous studies have advanced accelerometer-based SSIs through recognition models, speech synthesis, and data augmentation, but sensor locations were largely inherited from prior SSI modalities or selected without systematic evaluation. In contrast, this study treated sensor placement as a core design factor and examined it in relation to the motion-related signals captured by accelerometers. This accelerometer-specific approach showed that efficient placement cannot be determined solely by the performance of individual sensors but also depends on how articulatory information across sensors is combined.

Based on these findings, we proposed a regionally distributed placement strategy for compact accelerometer-based SSIs. This strategy emphasizes sensor placement across functional articulatory regions to capture complementary movement information from different articulatory sources. Its robustness was further validated using a different word set, in which high-performing configurations preserved broad regional coverage despite changes in specific sensor locations. A reduced practical configuration also demonstrated how this principle can be applied while balancing articulatory coverage and usability. Overall, these findings provide an empirical basis for designing practical and effective accelerometer-based SSI systems.

6 Funding

This work was supported by the National Research Foundation of Korea (NRF), funded by the Korean government (MSIT), under Grant RS-2024-00348497.

7 Author Contributions

Heejin Choi contributed conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Visualization, Writing – original draft.

Sungmin Jung contributed conceptualization, Methodology, Validation, Writing – review & editing.

Jangjay Sohn contributed Methodology, Resources, Supervision, Validation, Writing – review & editing.

Chang-Hwan Im contributed conceptualization, Supervision, Project administration, Funding acquisition, Writing – review & editing.

8 Declaration of Interest Statement

The authors declare no conflicts of interest related to the research, authorship, and/or publication of this article.

9 Declaration of Generative AI Use

The first author used GPT-5 to improve the language and readability of this work. After using this tool, the authors thoroughly reviewed and edited the content as necessary to take full responsibility for the final content of the article.

10 Data availability statement

The data that support the findings of this study are not publicly available due to participant privacy and ethical restrictions but are available from the corresponding author upon reasonable request and with permission of the Institutional Review Board of Hanyang University.

11 References

Adamov, L., Petrović, B., Milić, L., Štrbac, V., Kojić, S., Joseph, K., & Stojanović, G. M. (2025). Comparative analysis of electrical signals in facial expression muscles. *Biomedical Engineering Online*, *24*(1), 17. https://doi.org/10.1186/s12938-025-01350-3

Alsaawi, A. (2019). Spoken and written language as medium of communication: A self-reflection. *International Journal of Applied Linguistics & English Literature*, *8*(2), 194–198. https://doi.org/10.7575/aiac.ijalel.v.8n.2p.194

Browne, C., Culligan, B., & Phillips, J. (2013). *The New General Service List*. New General Service List Project. https://www.newgeneralservicelist.com

Cao, B., Ravi, S., Sebkhi, N., Bhavsar, A., Inan, O. T., Xu, W., & Wang, J. (2023). MagTrack: a wearable tongue motion tracking system for silent speech interfaces. *Journal of Speech, Language, and Hearing Research*, *66*(8S), 3206–3221. https://doi.org/10.1044/2023_JSLHR-22-00319

Chen, C.-C., Chen, Y.-J., Chen, S.-C., Lin, H.-S., & Lu, T.-W. (2011). Evaluation of soft-tissue artifacts when using anatomical and technical markers to measure mandibular motion. *Journal of Dental Sciences*, *6*(2), 95–101. https://doi.org/10.1016/j.jds.2011.03.010

Denby, B., Schultz, T., Honda, K., Hueber, T., Gilbert, J. M., & Brumberg, J. S. (2010). Silent speech interfaces. *Speech Communication*, *52*(4), 270–287. https://doi.org/10.1016/j.specom.2009.08.002

Ferreira, D., Silva, S., Curado, F., & Teixeira, A. (2022). Exploring silent speech interfaces based on frequency-modulated continuous-wave radar. *Sensors*, *22*(2), 649. https://doi.org/10.3390/s22020649

Gómez-Vilda, P., Gómez-Rodellar, A., Ferrández Vicente, J. M., Mekyska, J., Palacios-Alonso, D., Rodellar-Biarge, V., Álvarez-Marquina, A., Eliasova, I., Kostalova, M., & Rektorova, I. (2019). Neuromechanical modelling of articulatory movements from surface electromyography and speech formants. *International Journal of Neural Systems*, *29*(2), 1850039. https://doi.org/10.1142/S0129065718500399

Hoffmann, T. K. (2021). Total laryngectomy—still cutting-edge? *Cancers*, *13*(6), 1405. https://doi.org/10.3390/cancers13061405

Hogue, S., Guo, X., Morrison, R. A., McDowell, S., & Shembel, A. C. (2023). Use of motion capture technology to study extrinsic laryngeal muscle tension and hyperfunction. *The Laryngoscope*, *133*(12), 3472–3481. https://doi.org/10.1002/lary.30829

Hueber, T., Benaroya, E.-L., Chollet, G., Denby, B., Dreyfus, G., & Stone, M. (2010). Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, *52*(4), 288–300. https://doi.org/10.1016/j.specom.2009.11.004

Ivanovs, M., Kadikis, R., & Ozols, K. (2021). Perturbation-based methods for explaining deep neural networks: A survey. *Pattern Recognition Letters*, *150*, 228–234. https://doi.org/10.1016/j.patrec.2021.06.030

Jung, S., Sohn, J., Kwon, J., & Im, C.-H. (2026). Subject-independent silent speech classification using three-axis accelerometers with Z-axis vector rotation-based data augmentation. *IEEE Transactions on Audio, Speech and Language Processing*, *34*, 1686–1697. https://doi.org/10.1109/TASLPRO.2026.3671660

Kim, T., Shin, Y., Kang, K., Kim, K., Kim, G., Byeon, Y., Kim, H., Gao, Y., Lee, J. R., Son, G., Kim, T., Jun, Y., Kim, J., Lee, J., Um, S., Kwon, Y., Son, B. G., Cho, M., Sang, M.,…Yu, K. J. (2022). Ultrathin crystalline-silicon-based strain gauges with deep learning algorithms for silent speech interfaces. *Nature Communications*, *13*(1), 5815. https://doi.org/10.1038/s41467-022-33457-9

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations (ICLR 2015),

Kwon, J., Hwang, J., Sung, J. E., & Im, C.-H. (2024). Speech synthesis from three-axis accelerometer signals using conformer-based deep neural network. *Computers in Biology and Medicine*, *182*, 109090. https://doi.org/10.1016/j.compbiomed.2024.109090

Kwon, J., Nam, H., Chae, Y., Lee, S., Kim, I. Y., & Im, C.-H. (2023). Novel three-axis accelerometer-based silent speech interface using deep neural network. *Engineering Applications of Artificial Intelligence*, *120*, 105909. https://doi.org/10.1016/j.engappai.2023.105909

Lee, S., & Yook, D. (2002). Audio-to-visual conversion using hidden Markov models. PRICAI 2002: Trends in Artificial Intelligence (7th Pacific Rim International Conference on Artificial Intelligence),

Li, Z., Shi, J., Si, J., Lv, L., Guan, L., Hao, B., Tie, Z., Wang, D., Xing, C., & Quek, T. Q. S. (2025). Intelligent covert communication: recent advances and future research trends. *Engineering*, *44*, 101–111. https://doi.org/10.1016/j.eng.2024.12.007

Liu, S., Fawden, T., Zhu, R., Malliaras, G. G., & Bance, M. (2024). A data-efficient and easy-to-use lip language interface based on wearable motion capture and speech movement reconstruction. *Science Advances*, *10*(26), eado9576. https://doi.org/10.1126/sciadv.ado9576

Maeda, S., Toda, M., Carlen, A. J., & Meftahi, L. (2002). Functional modeling of face movements during speech. 7th International Conference on Spoken Language Processing (ICSLP 2002),

Meltzner, G. S., Heaton, J. T., Deng, Y., De Luca, G., Roy, S. H., & Kline, J. C. (2018). Development of sEMG sensors and algorithms for silent speech recognition. *Journal of Neural Engineering*, *15*(4), 046031. https://doi.org/10.1088/1741-2552/aac965

Nakamura, E., Honda, Y., Kageyama, Y., & Hirose, S. (2022). Facial feature points for Japanese speech content estimation. 2022 IEEE 4th Global Conference on Life Sciences and Technologies (LifeTech),

Ritter, A. R., Yildiz, V. O., Koirala, N., Baliga, S., Gogineni, E., Konieczkowski, D. J., Grecula, J., Blakaj, D. M., Jhawar, S. R., VanKoevering, K. K., & Mitchell, D. (2023). Factors associated with total laryngectomy utilization in patients with cT4a laryngeal cancer. *Cancers*, *15*(22), 5447. https://doi.org/10.3390/cancers15225447

Salomons, I., del Blanco, E., Navas, E., & Hernáez, I. (2025). Electrode setup for electromyography-based silent speech interfaces: a pilot study. *Sensors*, *25*(3), 781. https://doi.org/10.3390/s25030781

Sharpe, G., Camoes Costa, V., Doubé, W., Sita, J., McCarthy, C., & Carding, P. (2019). Communication changes with laryngectomy and impact on quality of life: a review. *Quality of Life Research*, *28*(4), 863–877. https://doi.org/10.1007/s11136-018-2033-y

Sood, A., & Craven, M. (2022). Feature importance explanations for temporal black-box models. Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-22),

Sun, K., Yu, C., Shi, W., Liu, L., & Shi, Y. (2018). Lip-Interact: improving mobile device interaction with silent speech commands. Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology,

Vatikiotis-Bateson, E., Munhall, K. G., Hirayama, M., Lee, Y. V., & Terzopoulos, D. (1996). The dynamics of audiovisual behavior in speech. In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 221–232). Springer. https://doi.org/10.1007/978-3-662-13015-5_16

Xie, L., Wang, L., & Yang, S. (2016). Visual speech animation. In B. Müller & S. I. Wolf (Eds.), *Handbook of human motion* (pp. 1–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-30808-1_1-1

Xie, Y., Han, Z., Xiao, Q., Liang, L., Tao, L.-Q., & Ren, T.-L. (2026). Silent speech sentence recognition with six-axis accelerometers using Conformer and CTC algorithm. ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP),

Yasmin, S., Irsik, V. C., Johnsrude, I. S., & Herrmann, B. (2023). The effects of speech masking on neural tracking of acoustic and semantic features of natural speech. *Neuropsychologia*, *186*, 108584. https://doi.org/10.1016/j.neuropsychologia.2023.108584

Zhu, B., Zeng, X., Chen, Z., Zhang, D., & Xie, L. (2024). SSR using portable sEMG acquisition system with electrode layout optimization. *IEEE Sensors Journal*, *24*(15), 24997–25005. https://doi.org/10.1109/JSEN.2024.3410401

Zhu, M., Zhang, H., Wang, X., Wang, X., Yang, Z., Wang, C., Samuel, O. W., Chen, S., & Li, G. (2021). Towards optimizing electrode configurations for silent speech recognition based on high-density surface electromyography. *Journal of Neural Engineering*, *18*(1), 016005. https://doi.org/10.1088/1741-2552/abca14

Tables

Table 1. Anatomical locations of the ten accelerometer sensors

Sensor Number	Locations
1	philtrum, centered between the nose and upper lip
2	laterally to the oral commissure (the corner of the mouth)
3	masseter region (near the jaw angle)
4	apex of the chin, aligned vertically with Sensor 1
5	under the chin, specifically within the submandibular triangle
6	posterior buccal (cheek) region, along the line of the cheek dimple
7	cutaneous lower lip, directly below the mouth
8	mandibular body at the midpoint between the Menton (chin tip) and Gonion (jaw angle)
9	midpoint between the Menton and the hyoid bone
10	lower neck area, specifically at the carotid triangle

Table 2. Training hyperparameters used for the CNN-BiLSTM model

Hyperparameter	Value
Epoch	100
Batch size	60
Learning rate	0.0005
Optimizer	Adam
β_1	0.9
β_2	0.999
Decay	5e-8

Table 3. Highest-performing sensor configurations among all possible combinations for each sensor count. Sensor locations are grouped into columns according to functional articulatory region. Dashes indicate that no sensor from the corresponding region was included. Boldface indicates the efficient configuration selected at the five-sensor saturation point.

Number of Sensors	Sensor Configuration		
	Labial	Buccal-Mandibular	Submental
1	**7**	–	–
2	7	8	–
3	1	4	5
4	1, 7	4	5
5	**1, 7**	**4, 8**	**5**
6	1, 7	4, 8	5, 9
7	1, 7	3, 4, 8	5, 9
8	1, 2, 7	4, 8	5, 9, 10
9	1, 2, 7	4, 6, 8	5, 9, 10
10	1, 2, 7	3, 4, 6, 8	5, 9, 10

Table 4. Accuracy drops after single-sensor ablation. Values indicate the decrease in mean classification accuracy relative to the full 10-sensor configuration when the corresponding sensor was excluded. Positive and negative values indicate performance degradation and slight performance improvement after sensor removal, respectively. Boldface indicates sensors included in the efficient configuration {1, 4, 5, 7, 8}.

Sensor Number	**1**	2	3	**4**	**5**	6	**7**	**8**	9	10
Accuracy Drop [%p]	**0.930**	0.048	-0.050	**0.620**	**0.332**	-0.043	**0.660**	**0.248**	0.228	0.060

Table 5. Classification accuracy using sensors from each functional articulatory region. For the labial and submental regions, accuracy was calculated using all sensors in the corresponding region. For the buccal-mandibular region, both the full four-sensor set and the average across all possible three-sensor subsets are reported for comparison.

Functional Articulatory Region	Classification Accuracy [%]
Labial	88.24 ± 6.08
Buccal-Mandibular	
4-Sensor Full Set	86.87 ± 7.13
3-Sensor Subset (Avg)	85.47 ± 7.64
Submental	81.92 ± 8.96

Table 6. Accuracy drops after region-level ablation. Values indicate the decrease in mean classification accuracy after removing all sensors from each functional articulatory region. Accuracy drops are reported relative to the corresponding reference configuration (all-sensor set or efficient configuration).

Functional Articulatory Region	Δ Accuracy [%p]	
	From All-Sensor Set	From Efficient Configuration
Labial	2.86	4.02
Buccal-Mandibular	0.98	1.35
Submental	1.47	1.49

Table 7. Highest-performing sensor configurations among all possible combinations for each sensor count using the new word set. Sensor locations are grouped into columns according to functional articulatory region. Dashes indicate that no sensor from the corresponding region was included. Boldface indicates the efficient configuration selected at the five-sensor saturation point.

Number of Sensors	Sensor Configuration		
	Labial	Buccal-Mandibular	Submental
1	–	8	–
2	7	–	5
3	1	8	5
4	1	8	5, 10
5	**1, 7**	**8**	**5, 10**
6	1, 7	6, 8	5, 10
7	1, 7	4, 6, 8	5, 10
8	1, 2, 7	4, 6, 8	5, 10
9	1, 2, 7	3, 6, 8	5, 9, 10
10	1, 2, 7	3, 4, 6, 8	5, 9, 10

Figures

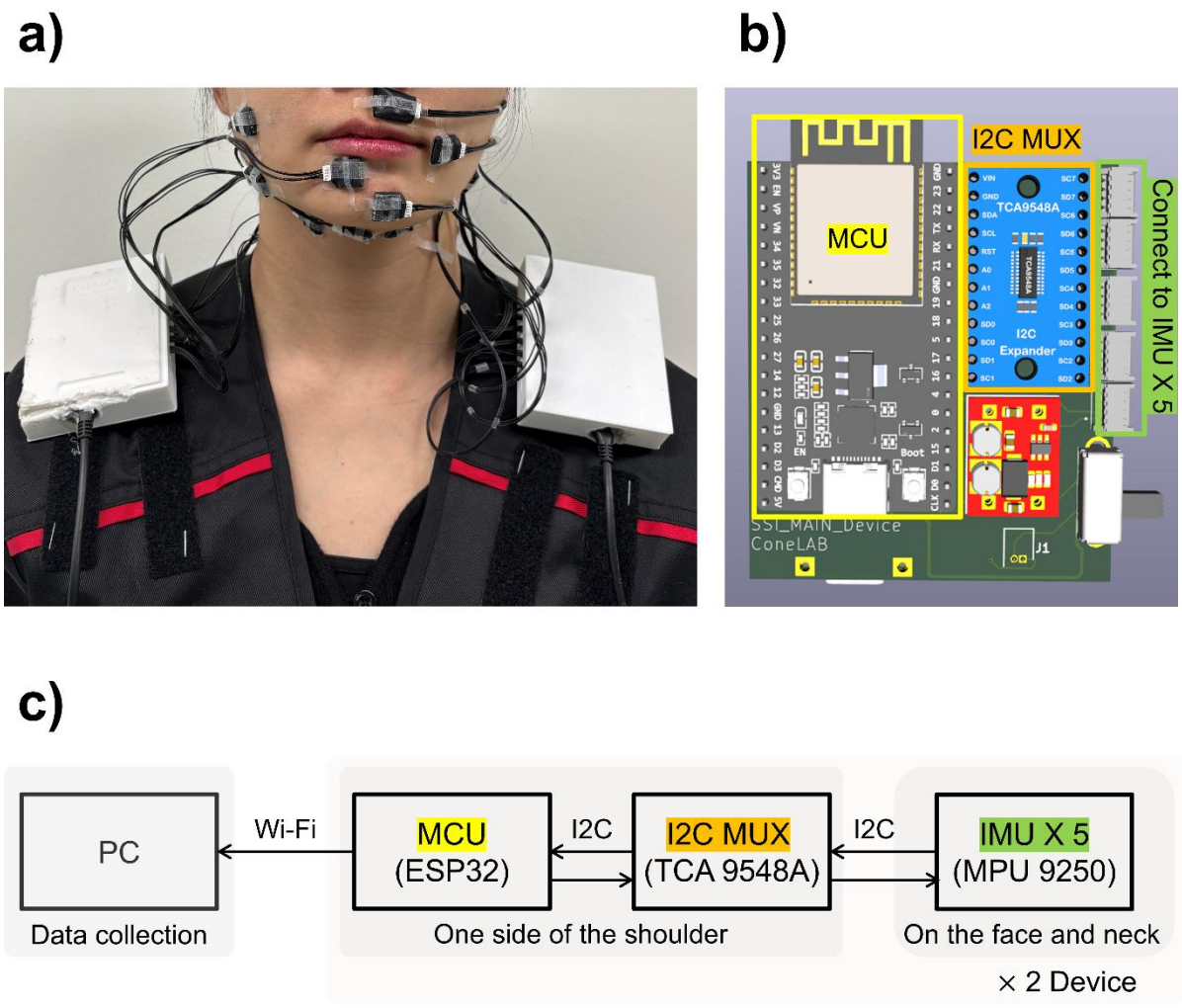


Figure 1. System implementation and data communication architecture.

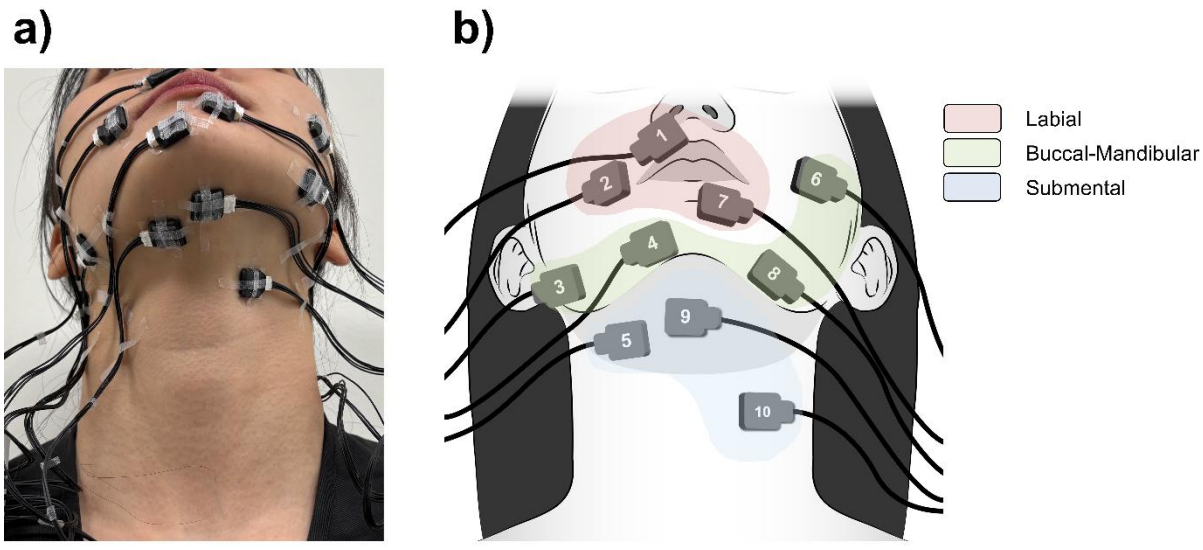


Figure 2. Sensor locations and functional articulatory regions.

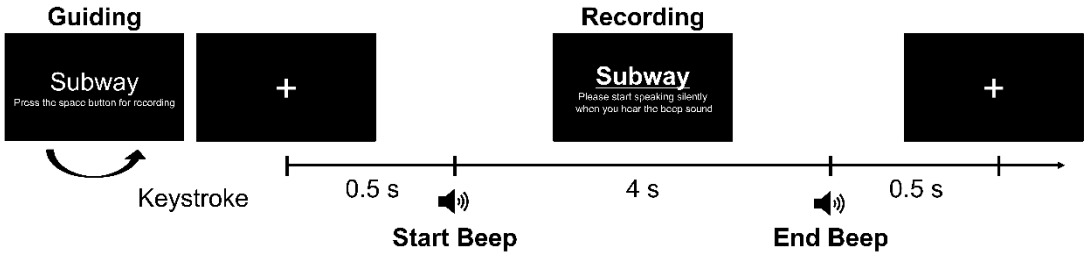


Figure 3. Experimental paradigm for silent speech data collection.

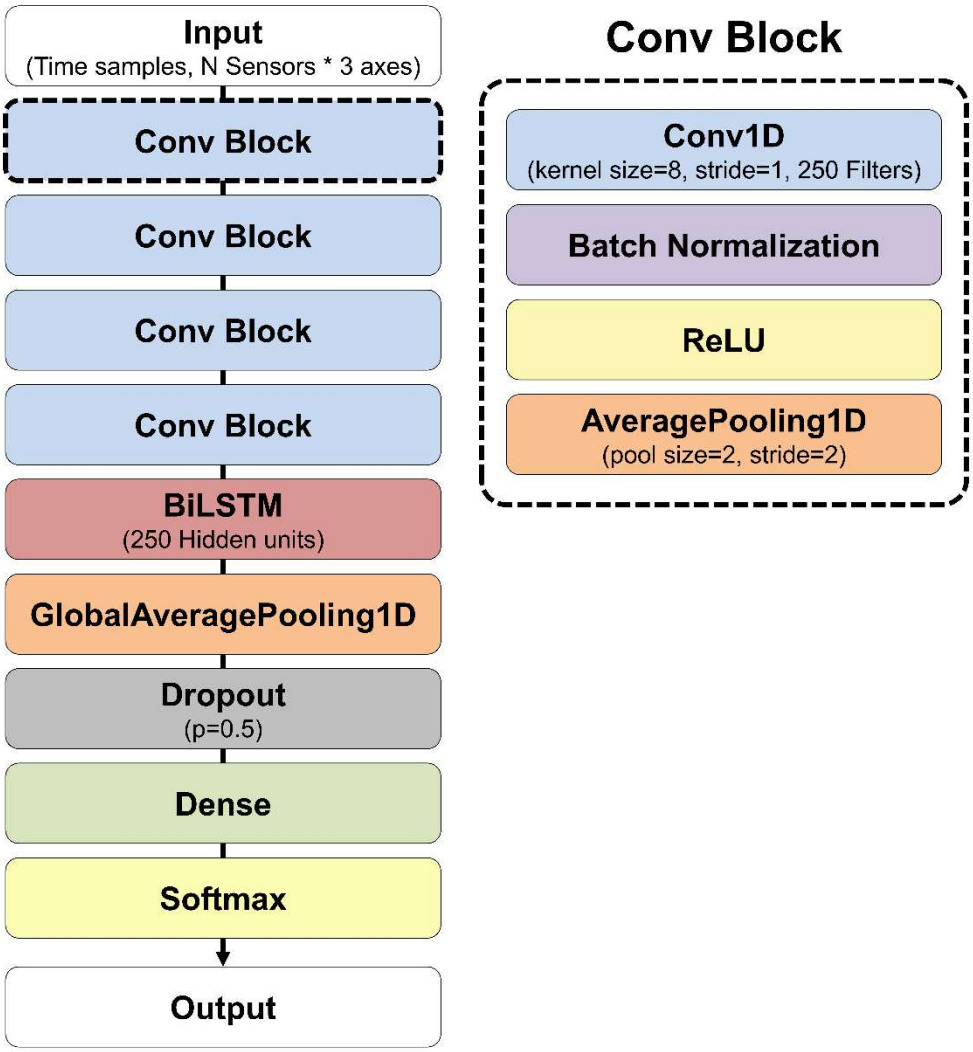


Figure 4. CNN-BiLSTM model architecture for 100-word classification.

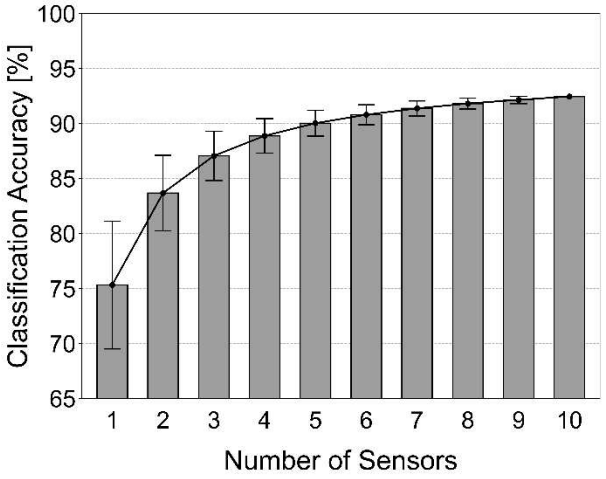


Figure 5. Average classification accuracy across all possible sensor configurations for each sensor count.

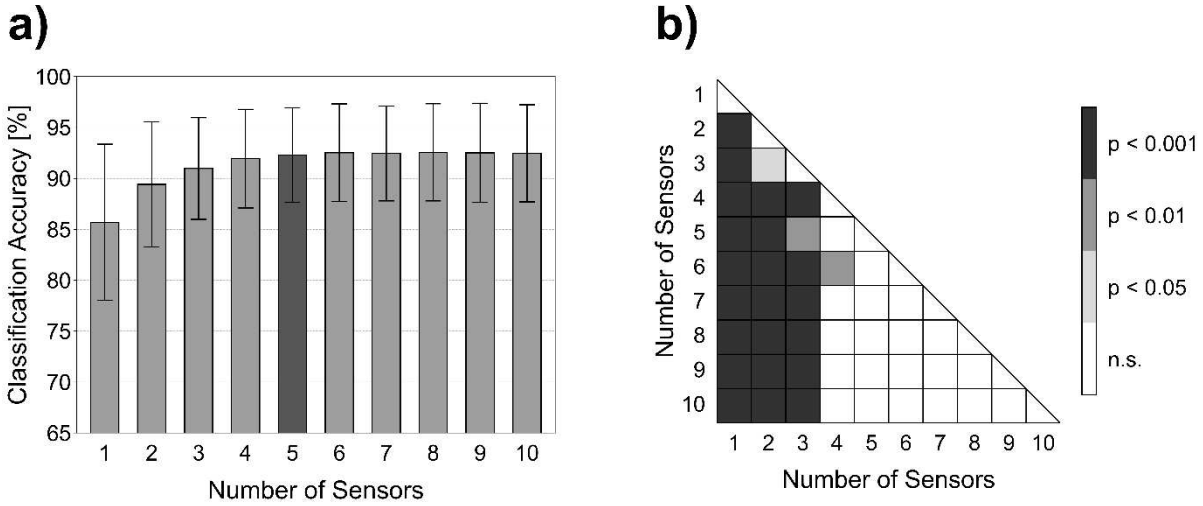


Figure 6. Statistical identification of the saturation point by sensor count.

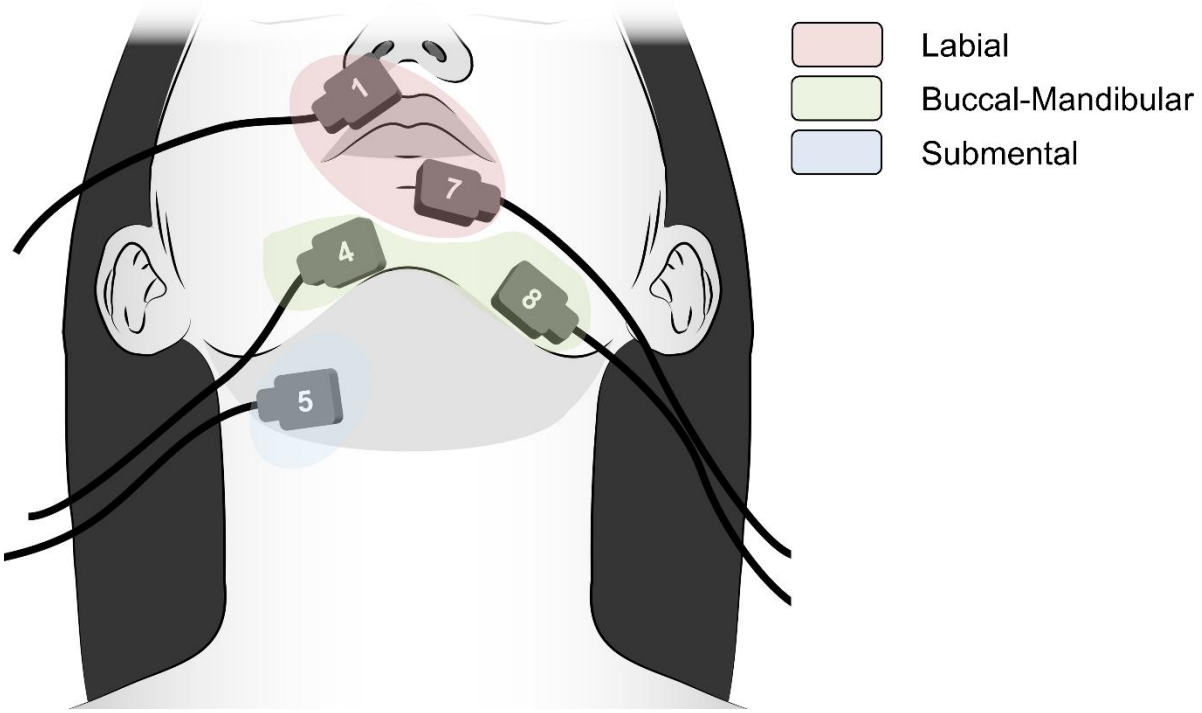


Figure 7. Sensor locations of the efficient configuration {1, 4, 5, 7, 8}.

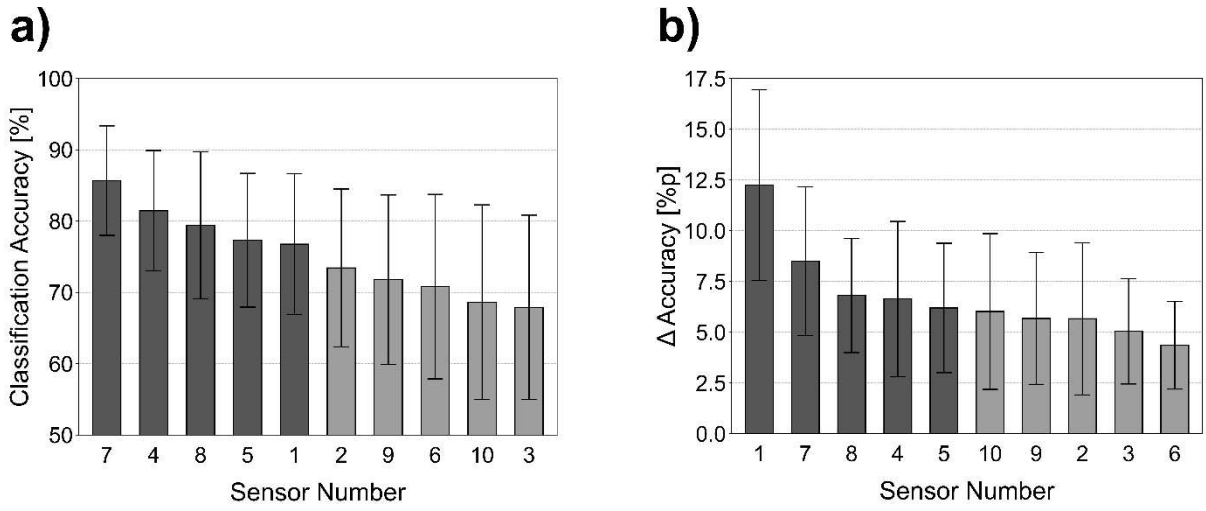


Figure 8. Individual sensor analysis based on standalone performance and perturbation-based contribution.

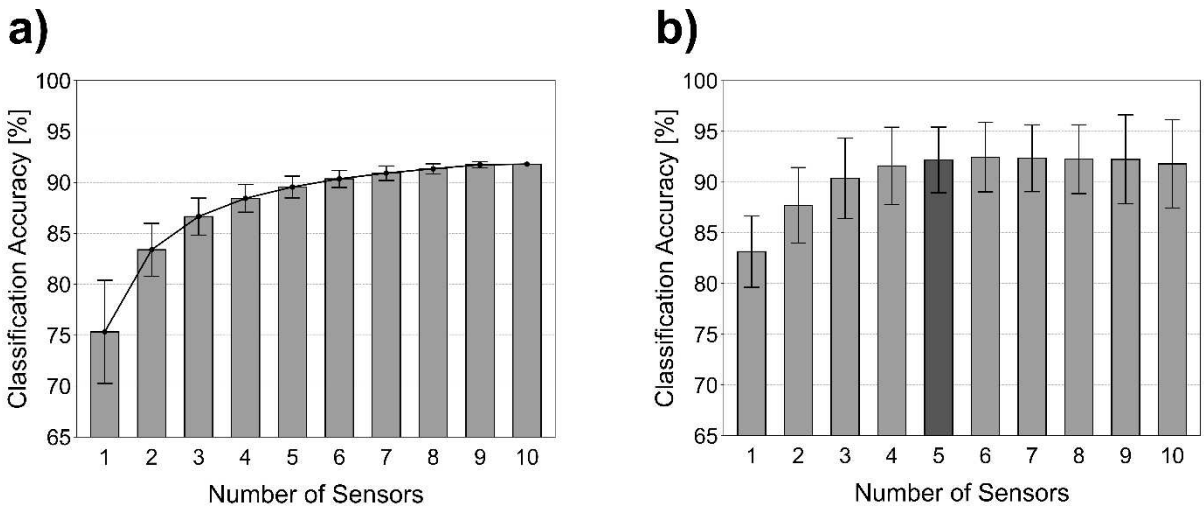


Figure 9. Classification accuracy for the new word set as a function of sensor count.

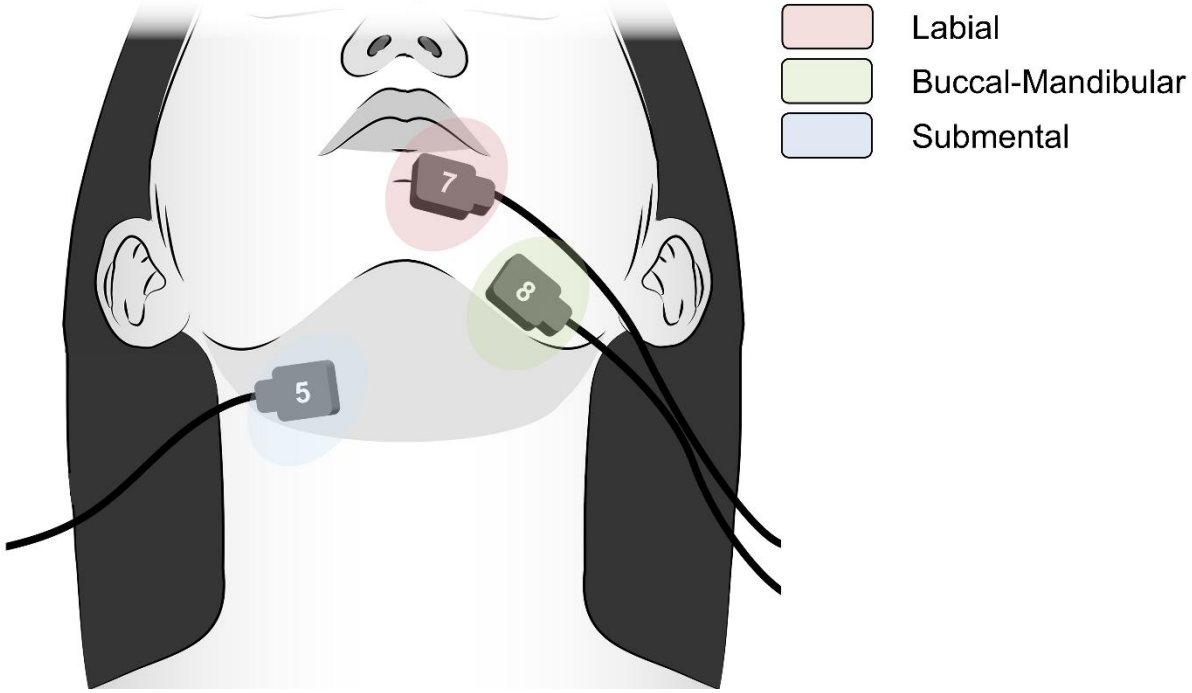


Figure 10. Sensor locations of the practical reduced-sensor configuration {5, 7, 8}.

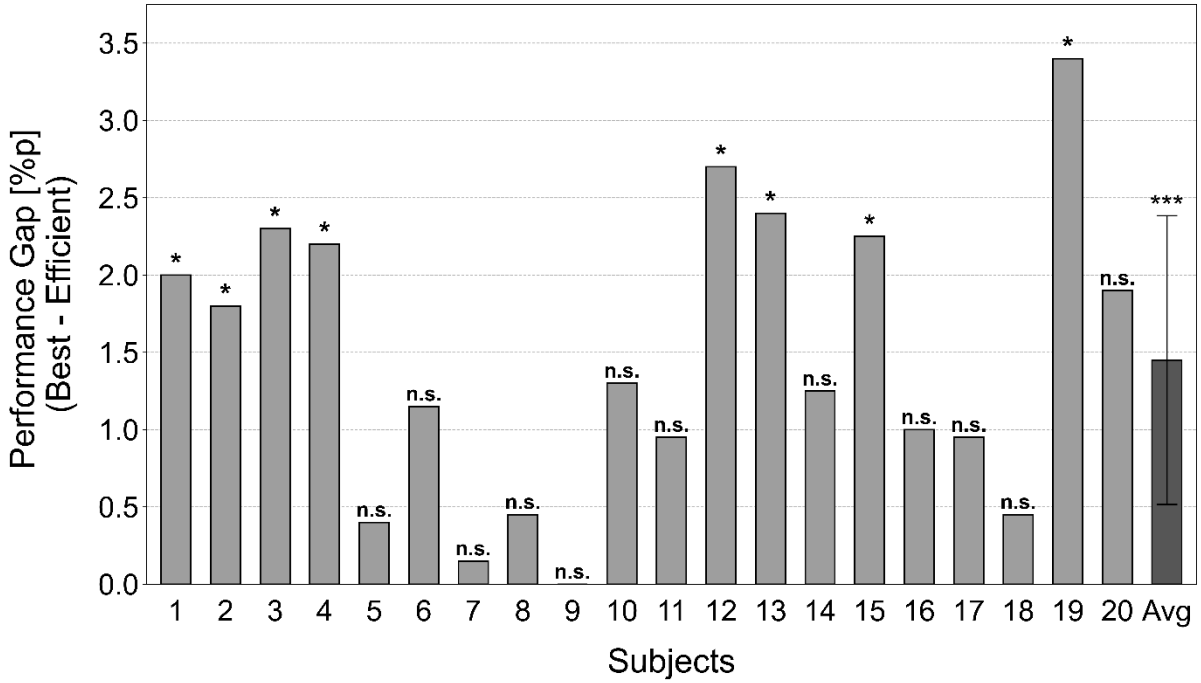


Figure 11. Performance gap between the subject-specific best-performing configuration and the efficient configuration {1, 4, 5, 7, 8}.

Figure captions

Figure 1. System implementation and data communication architecture. (a) Experimental setup showing two shoulder-mounted devices connected to ten IMU sensors attached to the participant's face and neck. (b) Mainboard configuration inside each shoulder-mounted device, consisting of an ESP32 microcontroller, a TCA9548A I2C multiplexer, sensor connectors, a power switch, and a 3.3 V voltage regulator. (c) Schematic of the data acquisition and communication flow. Acceleration signals from five IMU sensors on each side were transmitted to the ESP32 through an I2C multiplexer and then sent to the PC via Wi-Fi for data collection.

Figure 2. Sensor locations and functional articulatory regions. (a) Photograph showing ten IMU sensors attached to the participant's lower face and neck. (b) Schematic illustration of the same sensor locations, with sensors grouped into functional articulatory regions: labial (Sensors 1, 2, and 7), buccal-mandibular (Sensors 3, 4, 6, and 8), and submental (Sensors 5, 9, and 10).

Figure 3. Experimental paradigm for silent speech data collection. The figure illustrates the visual interface and timing sequence of each trial. After viewing the target word, participants pressed a key to start the trial. Signal recording began with the brief presentation of a fixation cross. A 0.5 s pre-recording buffer was followed by a start beep and a 4 s recording window, during which participants silently articulated the target word. An end beep marked the end of the recording window, and recording continued for an additional 0.5 s post-recording buffer before the trial ended.

Figure 4. CNN-BiLSTM model architecture for 100-word classification. Four convolutional blocks extract local temporal features from normalized accelerometer signals, and a BiLSTM layer captures temporal dependencies before global average pooling and softmax-based word classification.

Figure 5. Average classification accuracy across all possible sensor configurations for each sensor count. Gray bars represent the mean classification accuracy across all configurations with the same number of sensors, evaluated using 10-fold cross-validation. Error bars indicate the corresponding standard deviation across configurations, and black markers with connecting lines show the mean accuracy trend across sensor counts.

Figure 6. Statistical identification of the saturation point by sensor count. (a) Highest classification accuracy achieved at each sensor count. Bars represent the mean classification accuracy across subjects, and error bars indicate the standard deviation across subjects. The darker bar indicates the five-sensor saturation point. (b) Pairwise statistical comparisons between sensor counts using Wilcoxon signed-rank tests with FDR correction. Gray levels indicate significance thresholds ($p < 0.001$, $p < 0.01$, and $p < 0.05$), and white cells indicate non-significant differences.

Figure 7. Sensor locations of the efficient configuration {1, 4, 5, 7, 8}. Shaded areas indicate the functional articulatory regions represented in the configuration: the labial region (Sensors 1 and 7), the buccal-mandibular region (Sensors 4 and 8), and the submental region (Sensor 5).

Figure 8. Individual sensor analysis based on standalone performance and perturbation-based contribution. (a) Standalone classification accuracy obtained using each sensor independently. (b) Accuracy drop after perturbing each sensor input in the full 10-sensor model. Bars represent mean values across subjects, and error bars indicate the standard deviation across subjects. In each panel, sensors are sorted in descending order according to the corresponding metric, and darker bars indicate sensors included in the efficient configuration {1, 4, 5, 7, 8}.

Figure 9. Classification accuracy for the new word set as a function of sensor count. (a) Average classification accuracy across all possible sensor configurations for each sensor count. Gray bars represent the mean accuracy across configurations, error bars indicate the standard deviation across configurations, and black markers with connecting lines show the mean accuracy trend. (b) Highest classification accuracy achieved at each sensor count. Bars represent the mean accuracy across subjects, and error bars indicate the standard deviation across subjects. The darker bar indicates the five-sensor configuration with the highest accuracy.

Figure 10. Sensor locations of the practical reduced-sensor configuration {5, 7, 8}. The configuration includes one sensor from each functional articulatory region: Sensor 7 from the labial region, Sensor 8 from the buccal-mandibular region, and Sensor 5 from the submental region.

Figure 11. Performance gap between the subject-specific best-performing configuration and the efficient configuration {1, 4, 5, 7, 8}. For each subject, the bar indicates the difference in classification accuracy between the subject-specific best-performing configuration and the efficient configuration, calculated from the 10-fold cross-validation results. Positive values indicate that the subject-specific best configuration outperformed the efficient configuration. Statistical significance for each subject was assessed using a Wilcoxon signed-rank test with FDR correction. The rightmost dark bar represents the mean performance gap across subjects, and the error bar indicates the standard deviation across subjects. This group-level difference was tested using a Wilcoxon signed-rank test without FDR correction. Asterisks indicate significance levels (***$p < 0.001$, ** $p < 0.01$, * $p < 0.05$; n.s., not significant).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx