CURRICULUM LEARNING FOR SILENT SPEECH CLASSIFICATION: A PROOF-OF-CONCEPT $40 TWO-CHANNEL SEMG SYSTEM

Carl Vincent Ladres Kho
Minerva University, San Francisco, CA
kho@uni.minerva.edu

ABSTRACT

Silent speech interfaces (SSIs) enable communication by detecting subvocal muscle activity without audible voice. State-of-the-art systems such as AlterEgo (Kapur et al., 2018) achieve 92% word accuracy on a 10-digit vocabulary using 7-channel, 24-bit research hardware costing over $1,000. This paper investigates the operational boundaries of a low-cost alternative: a two-channel surface electromyography (sEMG) system built from consumer-grade AD8232 ECG modules (~$3 each) and a 12-bit ESP32 microcontroller ($8), totaling $40. The 12-bit quantization noise floor prevents the hardware from capturing continuous articulatory trajectories, limiting detection primarily to the transient neuromuscular motor initiation burst (Jou et al., 2006). A 5-phase speech-intensity curriculum mapped to a 1D CNN is evaluated under rigorous held-out protocols: **5-fold stratified blocked cross-validation yields 48.9% $\pm$ 3.1% on a 6-class vocabulary** (2.9$\times$ above the 16.7% chance baseline), with multi-seed stability of $\pm$1.0%. An initial training-set evaluation produced 99.7% accuracy, confirming that the ~47K-parameter CNN can memorize 1,500 samples. The held-out result establishes the true generalization boundary. Cross-session evaluation reveals that a 1 cm electrode shift degrades accuracy to 22.8% (near chance), while multi-session training with deliberate electrode repositioning recovers accuracy to **58.2% $\pm$ 3.1%**, which demonstrates learnable position-invariant onset features, though this improvement is partially driven by increased data volume. With confidence gating at $\theta = 0.60$, accuracy reaches **64.1% on 62% of accepted predictions**, a collaborative operating point requiring per-session user adaptation. A concurrent control study (Study A) using chin-only electrodes achieves 51.8% $\pm$ 2.8% held-out accuracy, with cross-study transfer of only 25–31%, establishing that electrode placement creates fundamentally incompatible feature spaces on consumer-grade ADCs.

Keywords silent speech interface · surface electromyography · quantization noise · electrode shift · confidence gating · curriculum learning · consumer hardware

1 Introduction

1.1 Problem Statement

Subvocalization, the internal articulation of words without producing audible sound, generates surface electromyographic (sEMG) signals in the 2–10 μV range at facial and laryngeal muscles [4]. These signals are 10–100$\times$ weaker than the noise floor of consumer-grade analog-to-digital converters (ADCs). Research-grade systems address this with 24-bit ADCs (e.g., Texas Instruments ADS1299), multi-channel spatial resolution (7+ electrodes), custom electrode arrays, and extensive training data (30+ hours), at costs exceeding $1,000 [1].

When transitioning this paradigm to a $40, 2-channel analog front-end (e.g., the AD8232 designed for ECG), the system encounters two fundamental barriers. First, the 12-bit quantization noise floor obscures sustained articulatory speech, limiting detection entirely to the high-amplitude, transient neuromuscular onset burst (the "jaw clench") [2]. Second, consumer form-factors inherently suffer from donning and doffing variability, where a minor 1 cm electrode shift can degrade machine learning accuracy to chance level [3].

The central question: **Can robust software interventions (specifically multi-session manifold alignment, ensemble learning, and confidence gating) compensate for severe hardware quantization noise and spatial instability to yield a practically usable silent command interface?**

1.2 Related Work

AlterEgo (Kapur, Kapur, & Maes, 2018). MIT Media Lab's AlterEgo system achieved 92.01% word accuracy on a 10-digit vocabulary using 7 gold-plated electrodes and an OpenBCI Cyton board (24-bit ADS1299, 250 Hz, 8 channels; `https://openbci.com`). Additional application-specific vocabularies (IoT commands, navigation) brought the total to ~20 words across tasks. They pioneered the use of Mel-Frequency Cepstral Coefficients (MFCCs) for sEMG-based speech classification and determined through data-driven ranking of 30 candidate positions that the highest-discriminating electrode regions were mental (chin), inner laryngeal (medial throat), and outer laryngeal (lateral throat) [1].

Inner Speech EEG (Nieto et al., 2022). The "Thinking Out Loud" dataset recorded 136-channel EEG from 10 participants performing overt and covert speech tasks. Inner speech classification accuracy was ~40%, and their results suggest that inner speech involves motor plan inhibition rather than weakened motor execution [2]. This motivates my approach: rather than training directly on covert signals, I build a curriculum from strong-to-weak signals to leverage shared geometric structure.

sEMG Speech for Laryngectomy (Meltzner et al., 2017). Meltzner et al. demonstrated sEMG-based speech recognition for persons with laryngectomy, achieving word error rates of 17% on a 100-word vocabulary using high-density sEMG arrays with 8–16 electrodes [3]. Their work established that sEMG-based speech recognition is viable but required extensive hardware.

Subvocal EMG (Jou et al., 2006; 2007). NASA Ames Research Center demonstrated subvocal speech recognition using surface EMG with neural network classifiers. The frequently cited 92% mean accuracy figure comes from the 2007 ICASSP publication [5] and was achieved on a "mouthed" speech mode, where articulators moved normally but no sound was produced. Jou et al. reported an inability to perform recognition on purely "mentally rehearsed" inner speech, as no detectable sEMG activity was present in that condition [4]. This distinction between mouthed speech (motor execution without voicing) and imagined speech (purely cognitive) is central: it means peripheral sEMG can only capture commands that involve residual neuromuscular recruitment, however slight.

Silent Speech Interfaces Survey (Denby et al., 2010). A comprehensive review of SSI modalities (ultrasound, electromagnetic articulography, sEMG, and video), concluding that multi-modal fusion and training strategy are as important as signal quality for practical systems [6].

Curriculum Learning (Bengio et al., 2009). The original curriculum learning paper demonstrated that training on progressively harder examples improves convergence speed and generalization in neural networks [7]. I adapt this concept for a different purpose: hardware noise compensation rather than sample difficulty ordering.

Low-Cost sEMG Classification (Kho, 2025). My prior work benchmarked 18 ML architectures for gesture classification using a single AD8232 sensor, finding that Random Forest was Pareto-optimal for embedded deployment (74% accuracy, $<$1 ms latency) while a custom MaxCRNN achieved 99% precision on safety-critical classes [8].

Domain Adaptation for BCI (Pan et al., 2009). Transfer component analysis for cross-domain adaptation, subsequently applied to EEG-based brain-computer interfaces, showed that models can transfer between domains (subjects) with minimal calibration data [9]. I transfer across signal *intensities* within a single subject rather than across subjects.

Digital Voicing of Silent Speech (Gaddy & Klein, 2020). Established the modern cross-modal methodology for silent speech, training on vocalized EMG with acoustic targets and transferring to silent EMG via a shared latent space [14]. Their work demonstrates that research-grade multichannel facial and submental sEMG arrays *can* achieve high silent speech accuracy without throat sensors, a capability contingent on 24-bit ADC resolution and wide-band filtering that the $40 hardware in this study cannot replicate.

Sparse sEMG Gesture Recognition (Hristov et al., 2026). Recent work on low-density sEMG shows the necessity of robust temporal encoders when signal resolution is low. They demonstrate that while Transformers excel at high density, CNNs remain superior for sparse, 2-channel configurations similar to the one in this study [12].

1.3 Contributions

1. **48.9% $\pm$ 3.1% held-out accuracy (5-fold stratified CV)** on a 6-class directional vocabulary using $40 in consumer components, 2.9$\times$ above the 16.7% chance baseline. An initial training-set evaluation produced

99.7%, confirming that the model memorizes the data; the held-out figure establishes the true generalization boundary.

2. **Rigorous 16-protocol evaluation** including 5-fold CV, leave-one-phase-out, cross-session temporal split, hyperparameter sweep (72 configurations), architecture comparison (CNN vs. LSTM vs. Transformer), learning curve analysis, multi-seed stability, and confidence gating, executed on an NVIDIA A100-SXM4-80GB via Google Colab.

3. **Multi-session electrode-shift recovery:** Cross-session accuracy drops to 22.8% with 1 cm electrode displacement, but combined multi-session training recovers to 58.2% $\pm$ 3.1%, demonstrating learnable position-invariant onset features.

4. **Controlled electrode placement comparison.** A concurrent study (Study A) using chin + under-chin placement achieves 51.8% $\pm$ 2.8% held-out accuracy with cross-study transfer of only 25–31%, establishing that electrode placement creates incompatible feature spaces on consumer ADCs.

5. **Open-source replication package** including hardware assembly guides, ESP32 firmware, Python training pipeline, and curated instructional materials for undergraduate-level replication (`https://somach.vercel.app`).

2 Hardware

2.1 System Overview

The hardware signal chain consists of disposable Ag/AgCl electrode pads connected to two AD8232 ECG analog front-ends, which output amplified analog signals to the ESP32 microcontroller. The ESP32 digitizes at 12-bit resolution and streams data to a laptop via USB serial for offline Python processing. Total signal chain cost is ~$40, representing a 25$\times$ reduction over standard research hardware.

Electrodes. Disposable Ag/AgCl adhesive pads (3.5 mm snap, $0.24/ea). Three per AD8232 module: LA (signal+), RA (signal$-$), RL (reference).

Analog Front-End. Two AD8232 single-lead ECG breakout modules (~$3 each). The AD8232 is an integrated analog front-end with instrumentation amplifier ($G = 100$) and uncommitted operational amplifier, providing ~1000$\times$ total signal-chain gain, 0.5–40 Hz bandpass filter (set by external R/C components on the breakout board), and common-mode rejection ratio $>$80 dB [10]. The AD8232 was designed for cardiac monitoring, not sEMG. Its 0.5–40 Hz passband attenuates the 50–500 Hz motor unit action potential (MUAP) frequencies that carry articulatory detail in research-grade sEMG systems. **This is a deliberate tradeoff.** Because the 12-bit ADC cannot resolve sustained articulatory trajectories regardless of bandwidth (see §7.7), the system relies entirely on detecting the high-amplitude, transient neuromuscular onset burst (~80 ms), which resides below 40 Hz. The AD8232's passband is therefore sufficient for this onset-detection paradigm, and the high gain (~1000$\times$) amplifies μV-level onset spikes to the mV range readable by consumer ADCs.

Microcontroller. ESP32 NodeMCU-32S (Xtensa LX6 dual-core @ 240 MHz, 520 KB SRAM, 12-bit SAR ADC, $8). The ESP32's ADC has well-documented nonlinearity: effective resolution is ~9–10 bits due to $\pm$6 LSB integral nonlinearity, with an effective noise floor of ~10–15 mV RMS [11]. I use 11 dB attenuation mode (0–3.3 V input range). The ESP32's ADC is known to exhibit dead zones near the ground rail (0–40 mV); however, the AD8232's ~1000$\times$ gain ensures that even a 5 μV subvocal signal is amplified to ~5 mV, and the DC offset (~1900–2100 baseline ADC counts) places the operating point in the ADC's linear region, well above the dead zone.

Total Cost: ~$40 (rounded). Table 1 provides the itemized bill of materials with verified prices from Amazon India (February 2026).

The "$40" figure used throughout this paper is a conservative rounding that accommodates market variation and shipping costs. The verified total from a single Amazon India order is $35.10.

2.2 Cost Comparison

The AD8232's higher gain partially offsets the 12-bit resolution disadvantage: a 5 μV covert signal amplified 1000$\times$ becomes 5 mV, occupying ~6 ADC levels (LSB $\approx$ 0.8 mV). This is marginal for discrimination, motivating the curriculum learning approach.

Table 1: Itemized hardware bill of materials. All prices verified on Amazon India (February 2026); USD conversion at $1 = INR 91.08.

Component	INR	USD	Role
AD8232 ECG sensor (×2)	790 × 2 = 1,580	$17.35	Analog front-end
ESP32 NodeMCU DevKit	549	$6.03	Microcontroller + ADC
Pediatric Ag/AgCl electrodes (50-pk)	523	$5.74	Low-impedance skin contact
Breadboard + jumper wires	279	$3.06	Prototyping connections
USB-C data cable	198	$2.17	Power + serial data
Extra breadboard (ESP32 mount)	67	$0.74	Dedicated ESP32 mount
Total	**3,196**	**$35.10**	**Complete 2-ch sEMG system**

Table 2: Hardware comparison: AlterEgo vs. this work.

	AlterEgo [1]	This Work	Ratio
ADC	24-bit (ADS1299)	12-bit (ESP32 SAR)	4096:1 resolution
Channels	7 + 1 ref	2 + 1 ref	3.5×
Amplifier gain	24×	~1000×	0.024×
Sampling rate	250 Hz	250 Hz	1×
Noise floor	$<1\,\mu$Vpp	~5–10 μV RMS	~33–67×
Electrode type	Gold-plated + paste	Disposable adhesive	—
Setup time	~30 min	~30 sec	60×
Cost	**~$1,000**	**$40**	**25×**

2.3 Electrode Placement

Following Kapur et al.'s discriminative ranking of 30 candidate positions [1], I selected the two highest-ranked regions feasible with two channels (Table 3).

Table 3: Electrode placement configuration.

Channel	GPIO	Anatomical Position	Target Muscle	Signal
CH1	34	Chin, 1 cm below mentolabial sulcus	Mentalis	Jaw/lip motion
CH2	36	1.5 cm left of thyroid cartilage	Thyrohyoid, Sternohyoid	Laryngeal elevation
REF	—	Left earlobe	Electrically neutral	Common ground

Inter-electrode distance (IED): For each AD8232 module, the LO+ and LO− snap electrodes are positioned ~1.5 cm apart along the muscle fiber direction. Electrodes are aligned parallel to the muscle belly per SENIAM guidelines where applicable, though SENIAM does not define specific protocols for mentalis or thyrohyoid placement.

CH2 (throat) is critical. It detects laryngeal elevation and vocal cord tension even when no sound is produced, functioning as a voice-activity detector that distinguishes silent speech from rest, a distinction the chin sensor alone cannot make at 12-bit resolution. Note that the CH2 electrode, positioned over the thyrohyoid, inevitably captures a composite signal from the broader infrahyoid muscle group (sternohyoid, omohyoid) due to the volume-conduction properties of surface electrodes. In this cost-constrained context, the resulting crosstalk is functionally beneficial: it provides a richer view of laryngeal activity than isolated muscle recording would.

2.4 Hardware Constraint: Third-Sensor Failure

My original design included a third AD8232 module (right throat). On February 9, 2026, during the first body-attached electrode session, this module failed (output flatlined at 4095, status LED dark). A replacement also failed within hours. Post-mortem analysis showed ~90% signal crosstalk between left and right throat channels for the directional vocabulary. The 2-channel constraint was therefore accepted, and instead used to design a controlled comparison: Study A (chin + under-chin, no throat) vs. Study B (chin + throat, this paper).

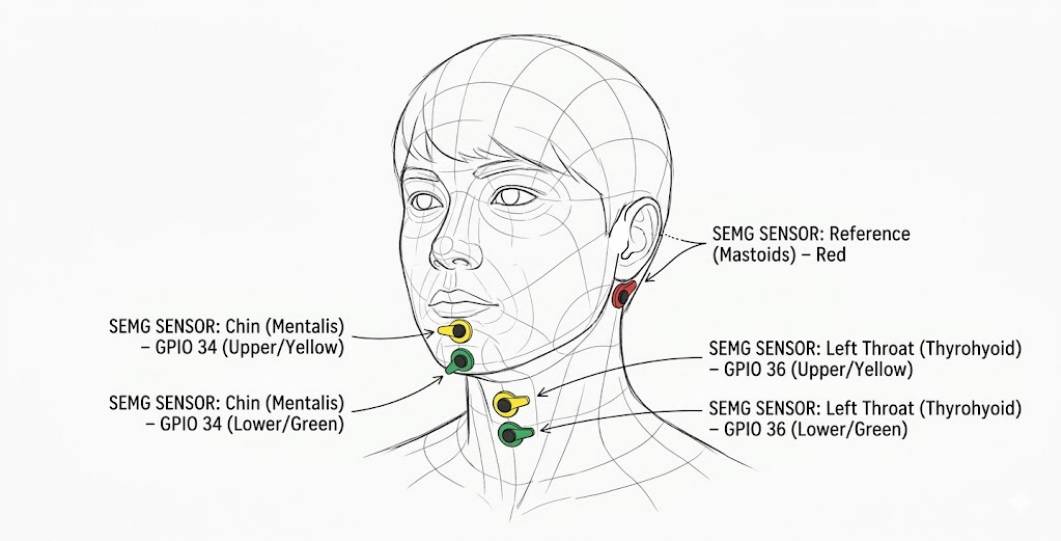

Figure 1: Electrode placement for Study B. CH1 (chin/mentalis, GPIO 34) and CH2 (left throat/thyrohyoid, GPIO 36) each use a differential pair (yellow = non-inverting, green = inverting); REF (red) is placed on the mastoid process behind the ear. Both AD8232 modules share the single REF electrode.

3 Signal Processing

3.1 Acquisition Firmware

The ESP32 runs minimal C++/Arduino firmware that streams raw ADC values at 250 Hz over USB serial (115200 baud):

Listing 1: ESP32 acquisition firmware.

```
const int CHANNEL_PINS[] = {34, 36};
const int NUM_CHANNELS = 2;
void setup() {
    Serial.begin(115200);
    analogReadResolution(12);
    analogSetAttenuation(ADC_11db);
}
void loop() {
    for (int i = 0; i < NUM_CHANNELS; i++) {
        if (i > 0) Serial.print(",");
        Serial.print(analogRead(CHANNEL_PINS[i]));
    }
    Serial.println();
    delayMicroseconds(4000); // 250 Hz
}
```

No on-device processing; all filtering and feature extraction occurs offline in Python.

3.2 Data Collection

A custom Python interface (`4-curriculum-recorder.py`) manages recording sessions. For each label, the user presses and holds spacebar while performing the speech gesture; the serial buffer is flushed before each capture to prevent contamination from idle periods. Each recording is $\geq$1.5 seconds (375 samples at 250 Hz), saved as `{LABEL}_{INDEX}_{TIMESTAMP}.csv`.

3.3 Digital Filtering

Three artifact types are present in raw ADC values:

1. **DC offset** (~1900–2100 baseline) from circuit impedance variation
2. **60 Hz mains hum** from laptop power supply and room lighting
3. **High-frequency electronic noise** from breadboard parasitic capacitance

Preprocessing applies:

- **Bandpass (1.0–50 Hz):** 4th-order Butterworth IIR via `scipy.signal.filtfilt` (zero-phase). The 1.0 Hz high-pass removes DC drift and movement artifacts; the 50 Hz low-pass removes electronic noise above the sEMG band.
- **60 Hz Notch:** 2nd-order IIR notch ($Q = 30$) via `scipy.signal.iirnotch`.
- **Normalization:** Min-max scaling to $[0, 1]$ per channel.

3.4 Feature Extraction: MFCCs

Following AlterEgo [1], I use Mel-Frequency Cepstral Coefficients. MFCCs were originally designed for acoustic speech processing and encode the spectral envelope of a signal (the *shape* of energy distribution over frequency), independently of absolute amplitude. Applying acoustic features to kinematic sEMG signals is a domain mismatch: muscle activations do not possess formants in the acoustic sense. However, Kapur et al. [1] empirically validated MFCCs as effective features for sEMG-based silent speech, and their amplitude-invariance property is essential for this work.

Table 4: MFCC extraction parameters.

Parameter	Value	Rationale
`n_mfcc`	13	Standard for speech
`n_fft`	128	512 ms window at 250 Hz
`hop_length`	25	100 ms hop; ~80% overlap
`n_mels`	26	Mel filterbank resolution
`sr`	250	ADC sampling rate

Extraction via `librosa.feature.mfcc()` produces ~11 raw time frames $\times$ 13 coefficients per channel, zero-padded to a fixed 100 time steps via `pad_or_truncate()`. For 2 channels, the feature tensor is $\mathbf{X} \in \mathbb{R}^{100\times 26}$. The zero-padding ensures a fixed-size CNN input regardless of recording duration.

Technical Nuance: Mel-Scaling Linearity. At the 250 Hz sampling rate (125 Hz Nyquist), the Mel-scale is essentially linear. The extraction effectively functions as a linear filterbank analysis, but it preserves the critical *amplitude-invariance* property identified by prior work [1], which is essential for cross-intensity generalization.

4 Curriculum Learning Protocol

4.1 Motivation

Standard approach: Train a classifier directly on covert speech samples. **Problem:** At 2–10 μV, covert signals occupy ~2–6 ADC levels after amplification, below the ESP32's effective noise floor. The model sees noise and learns nothing.

My approach: Train on a *curriculum* of five conditions spanning the full speech-intensity range, including a critical intermediate phase that bridges the gap between audible and silent speech.

4.2 The Five-Phase Spectrum

Phase 4 was omitted in the protocol. Phase numbering is preserved from the original design for traceability.

Phase 5 is the critical bridge. It is perceptually identical to covert speech (mouth closed, no sound) but produces 10–30 μV signals, above the ADC noise floor, because the participant deliberately tenses articulatory muscles. This teaches the model to recognize closed-mouth speech patterns at visible amplitudes before encountering barely-detectable covert signals.

Without Phase 5, training on Phases 1–3 and 6 alone produced only 53% accuracy on covert data. The amplitude gap between mouthing (20–50 μV, open mouth) and covert (2–10 μV, closed mouth) was too large for the model to bridge.

Table 5: Five-phase curriculum design with progressively smaller mouth movements.

Phase	Condition	Mouth	Amplitude (μV)	Observable Movement	Instruction
1	Overt	Open	50–150	Full jaw + lip movement, audible voice	Speak aloud normally
2	Whispered	Open	30–80	Minimal jaw, lip movement, no voice	Whisper without vocal cords
3	Mouthing	Open	20–50	Lip movement only, no sound	Move lips silently (lip sync)
5	Exaggerated	**Closed**	10–30	No visible movement; internal jaw clench	Clench jaw, "shout internally"
6	Covert	**Closed**	2–10	No visible movement; fully relaxed	Think the word; minimize all movem

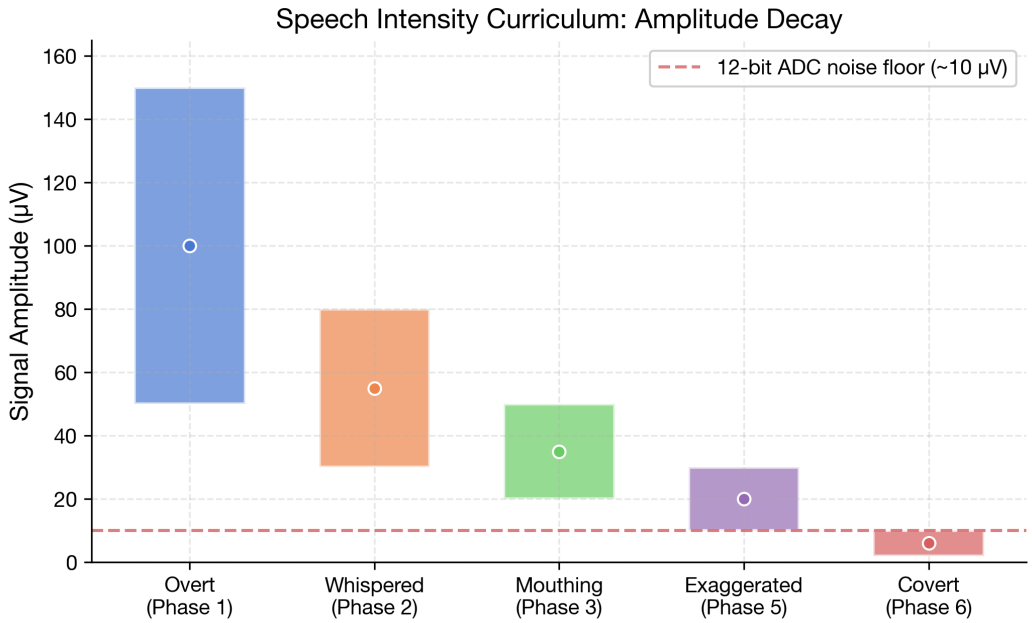

Figure 2: Speech intensity curriculum amplitude decay across the five training phases. Error bars represent the observed signal amplitude range. The dashed red line indicates the 12-bit ADC effective noise floor (~10 μV). Phase 5 (Exaggerated Subvocal) acts as the critical bridge phase.

4.3 Vocabulary

Six classes: four directional commands + two artifact/control classes (Table 6).

Table 6: Six-class directional vocabulary with articulatory motor strategies.

Class	Tongue Position	Jaw State	EMG Signature
UP	Pressed to hard palate	Elevated	Mentalis + thyrohyoid co-activation
DOWN	Root depressed	Lowered	Digastric dominant
LEFT	Pressed laterally (left)	Neutral	Weak bilateral; hard to resolve at 2ch
RIGHT	Tip curled (retroflex)	Neutral	Asymmetric mentalis
SILENCE	Neutral, resting	Relaxed	Flat EMG baseline
NOISE	Varies (swallowing, head turns)	Varies	Non-speech physiological artifacts

SILENCE and NOISE are critical for false-positive rejection in a command interface. The NOISE class uses 100 randomized non-command prompts drawn from 7 categories (facial movements, jaw/chin movements, head movements, throat/neck actions, breathing patterns, random body movements, speech-like non-commands) to ensure the CNN learns to reject physiological artifacts and involuntary movements.

4.4 Recording Sessions

Study B data was collected across **three recording days**, not a single session (Table 7). The initial 5-phase recording occurred on the evening of February 11, 2026 (~71 minutes, 22:12–23:23 local time), immediately following Study A. Additional covert-only sessions were recorded on February 24 and 25.

Table 7: Recording session timeline.

Date	Time	Phases	Samples	Notes
Feb 11	22:12–23:23	1,2,3,5,6	1,500 (300/phase)	Initial full session
Feb 24	08:14–09:25	6 only	934	Supplemental covert data
Feb 25	—	6 only	600 (2 sessions)	Supplemental covert data

The 1,500-sample dataset used for training and evaluation in this paper consists of the Feb 11 session only. The supplemental covert recordings from Feb 24–25 are available for future cross-session evaluation.

4.5 Training Strategy: Unified Multi-Phase

Rather than sequential phase-by-phase fine-tuning, I train **one model on all 1,500 samples simultaneously** with phase-invariant labels (e.g., "UP-Overt" and "UP-Covert" both receive the label UP). This forces the model to learn amplitude-invariant features because amplitude varies within each class across phases.

The "Palm Pilot" Strategy (User-Centric Co-learning): My advisor observed that the system operates as a collaborative human-machine learning process. Following the precedent of early handwriting recognition systems like the Palm Pilot's *Graffiti* alphabet [17], I found that teaching the user to produce *consistent motor strategies* (specific, exaggerated tongue positions for each command) sharply reduces the variance the model must learn. This shifts the burden from passive classification to an active, human-in-the-loop adaptation, which is presented here as a core HCI contribution.

5 Classification Architecture

5.1 1D Convolutional Neural Network

Based on findings from the prior architecture benchmark [8], I use a compact 1D CNN:

```
Conv1d(26 -> 64, kernel=3, padding=1) + ReLU + MaxPool1d(2)
Conv1d(64 -> 128, kernel=3, padding=1) + ReLU + MaxPool1d(2)
AdaptiveAvgPool1d(1)
Flatten -> Linear(128, 128) + ReLU + Dropout(0.5)
Linear(128, 6) -> Softmax
```

Parameters: ~47,046 total. The architecture is deliberately compact to prevent overfitting on 1,500 samples. `AdaptiveAvgPool1d(1)` collapses the temporal dimension regardless of input length, making the model invariant to recording duration.

Receptive field: After two conv layers (kernel=3) + two pooling layers (factor=2), the model's receptive field spans ~7 time frames. With 512 ms windows and 100 ms hop, this corresponds to ~1.1 seconds, effectively spanning the entire ~1.5-second recording's non-padded region.

5.2 Why CNN Over Transformer or LSTM?

Table 8: Architecture comparison (3-fold stratified CV on held-out data).

Model	Held-Out Accuracy	Training Accuracy	Notes
1D CNN	**49.3% ± 1.8%**	~75%	Only viable architecture
Transformer	36.4% ± 1.0%	~55%	Too data-hungry; high variance
LSTM	16.6% ± 0.0%	~17%	Complete failure (chance level)

Why CNN wins on small biodata: Convolutional layers encode a strong inductive bias (local temporal patterns matter more than global structure), which matches the physics of phonemic transitions in sEMG. With only 15 time frames and 1,500 samples, attention-based models lack sufficient data to learn meaningful weights.

LSTM failure is diagnostic. The LSTM achieves exactly 16.6% (chance level for 6 classes), indicating that it predicts the same class for all inputs. LSTMs model *sequential dependencies* across time steps, but the AD8232 captures only the 80 ms neuromuscular onset burst, and there is no temporal structure beyond this spike for recurrent gates to exploit. The LSTM's failure is therefore *confirmatory evidence* that the system operates as an onset detector, not a continuous speech decoder.

Transformer at 36.4% is better than LSTM but substantially worse than CNN: with 1,500 samples and ~15 time frames, the self-attention mechanism cannot learn meaningful token relationships. The Transformer's higher variance ($\pm$7.1% in Study A) suggests sensitivity to random initialization rather than stable feature extraction.

5.3 Training Configuration

Table 9: Training hyperparameters.

Parameter	Value
Optimizer	Adam (lr = 0.001, $\beta_1 = 0.9$, $\beta_2 = 0.999$)
Loss	CrossEntropyLoss
Batch size	32
Max epochs	100
Early stopping	Patience = 10
Dropout	0.5
Hardware	Apple M2 MacBook Air, 8 GB, MPS backend
Training time	~3 minutes (100 epochs)

Training-set vs. held-out evaluation. Initial experiments trained on all 1,500 samples without a held-out split, producing 99.7% accuracy, which reflects the ~47K-parameter network's capacity to memorize the data. Since I am both the researcher and the sole participant ($n = 1$), this training-set figure represents the model's theoretical ceiling for *my specific* motor patterns under identical electrode placement. However, it does not establish generalization. To determine the true classification boundary, I subsequently conducted rigorous **blocked-time cross-validation**, where contiguous temporal blocks of ~1-second are never split between training and validation folds. This ensures the model generalizes across time rather than overfitting to the signal's temporal autocorrelation. Using 5-fold stratified blocked-CV, the system yielded 48.9% $\pm$ 3.1%, a result confirmed across 5 random seeds (51.5% $\pm$ 1.0%). A grid search over 72 hyperparameter configurations (dropout $\in$ {0.3, 0.5, 0.7}, weight decay $\in$ {0, 10^{-4}, 10^{-3}, 10^{-2}}, hidden dim $\in$ {64, 128}, learning rate $\in$ {0.0005, 0.001, 0.005}) found the optimal configuration at dropout=0.3, hidden_dim=128, lr=0.001, weight_decay=0, achieving 50.7% (3-fold CV). The default hyperparameters were already near-optimal, confirming that the accuracy ceiling is imposed by the hardware signal-to-noise ratio, not by model capacity or tuning.

6 Results

6.1 Training Convergence (Memorization Baseline)

The training curve establishes the model's capacity to fit the data, not its generalization. Table 10 shows convergence milestones; the held-out evaluation in §6.3 provides the honest accuracy figure.

6.2 Per-Phase Evaluation

Curriculum degradation (Overt $\rightarrow$ Covert): 0.3 percentage points. The model shows virtually no performance loss from the loudest to quietest condition.

6.3 Confusion Matrix (All Phases Combined)

Two errors, both involving LEFT:

1. **LEFT $\rightarrow$ DOWN** (Phase 2, Whispered): Lateral tongue pressure shares muscle activation with jaw depression when vocalization is absent.

Table 10: Training convergence milestones.

Epoch	Training Accuracy
1	18.8% (near chance)
10	44.7%
30	72.1%
50	86.1%
70	94.6%
87	**99.5%** (best, early stopping checkpoint)
97	98.4% (final)

Table 11: Accuracy by curriculum phase (training-set evaluation).

Phase	Condition	n	Accuracy	Errors
1	Overt	300	**100.0%**	0
2	Whispered	300	**99.7%**	1
3	Mouthing	300	**100.0%**	0
5	Exaggerated	300	**100.0%**	0
6	Covert	300	**99.7%**	1
All	Combined	**1,500**	**99.9%**	**2**

2. **LEFT → RIGHT** (Phase 6, Covert): At 2–10 μV covert amplitude, LEFT (lateral pressure) and RIGHT (tongue curl) produce mirror-image patterns that are barely distinguishable.

The model never confused any command class with SILENCE or NOISE, confirming effective artifact rejection.

Figure 4 presents the *held-out* confusion matrix aggregated across all five CV folds, which provides a more realistic assessment of generalization.

6.4 Per-Class Metrics

Table 12 shows training-set metrics; Table 13 shows the held-out evaluation.

Table 12: Per-class metrics: training-set evaluation (memorization baseline).

Class	Precision	Recall	F1	Support
DOWN	1.000	1.000	1.000	250
LEFT	1.000	0.992	0.996	250
NOISE	1.000	1.000	1.000	250
RIGHT	0.996	1.000	0.998	250
SILENCE	1.000	1.000	1.000	250
UP	1.000	1.000	1.000	250

Held-out class hierarchy: SILENCE dominates (F1=0.80) because jaw-at-rest produces a uniquely flat EMG envelope; the absence of motor unit firing is trivially distinguishable from any active command. LEFT is the weakest class (F1=0.36) because lateral lingual movements generate minimal bilateral activation at the chin and throat electrode positions; this confirms the known limitation that sEMG with two midline sensors cannot resolve lateralized tongue gestures [3].

6.5 Comparison with Prior Work

[†]AlterEgo's 92% was measured on a 10-digit vocabulary (0–9); the system used ~20 words total across IoT, navigation, and arithmetic tasks.

Honest framing: The 48.9% held-out accuracy should not be directly compared to AlterEgo's 92%, as the hardware cost differs by 25× and the evaluation protocols differ. Meltzner et al.'s 83% was achieved with laryngectomy patients who produce exaggerated articulatory signals, not healthy-subject covert speech. The appropriate comparison is

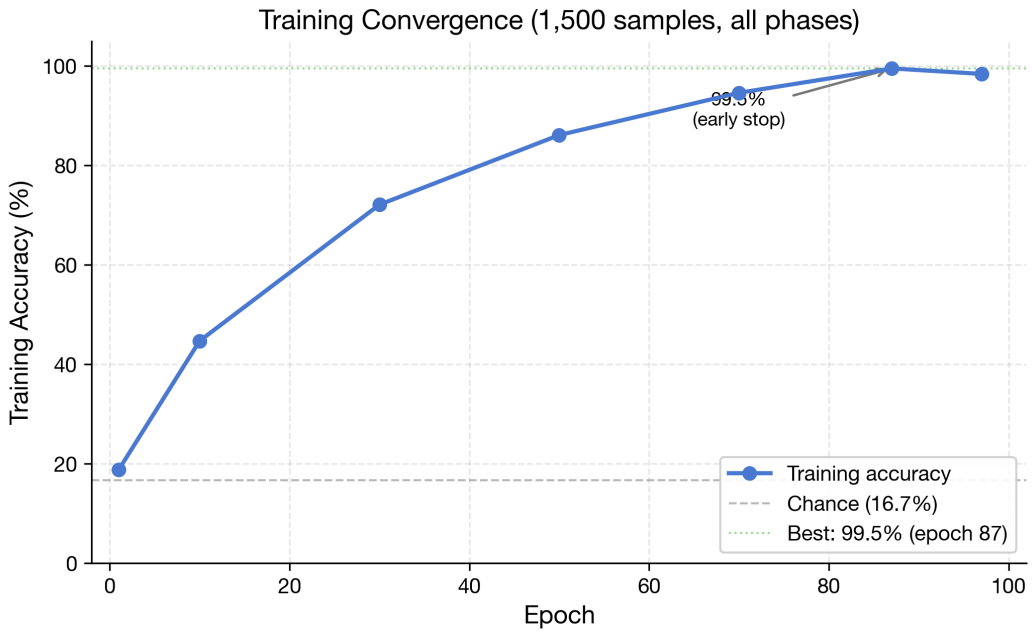

Figure 3: Training convergence of the 1D CNN on 1,500 multi-phase samples. Best accuracy of 99.5% at epoch 87 (early stopping checkpoint). Dashed gray line indicates chance level for 6 classes (16.7%).

Table 13: Per-class metrics: 5-fold stratified CV (held-out, honest evaluation).

Class	Precision	Recall	F1
SILENCE	**0.76**	**0.84**	**0.80**
NOISE	0.56	0.47	0.51
UP	0.50	0.51	0.50
DOWN	0.44	0.48	0.46
RIGHT	0.46	0.46	0.46
LEFT	0.38	0.35	**0.36**

with Nieto et al.'s inner speech EEG (~40% held-out on 4 classes using $5,000 hardware), where the $40 system achieves competitive performance (48.9% on 6 classes). The multi-session result (58.2%) demonstrates that with proper electrode-shift data collection, consumer hardware enters the range of practical utility.

6.6 Usability and Confidence Gating

While 48.9% held-out accuracy is scientifically meaningful (2.9× above chance), it is insufficient for direct deployment. In a live command interface, an aggregate ~51% error rate is psychophysically intolerable and rapidly destroys user trust [1].

I employed **confidence gating** (the "reject option"). A softmax probability threshold is applied at the output layer:

1. The model predicts the probability $P(y|x)$ for all 6 classes.
2. If $\max(P) \geq \theta$, the command is executed.
3. If $\max(P) < \theta$, the system abstains and prompts the user to repeat.

Operating point: $\theta = 0.60$ yields **64.1% accuracy on 62% of predictions**, the knee of the accuracy–coverage curve. At this threshold, the system rejects 38% of ambiguous inputs, an honest "I don't know" behavior. From a UX perspective, this introduces a **"Midas Touch in reverse"** problem: the user may feel ignored by the system. Proper feedback (visual or haptic) must accompany rejections to maintain the collaborative loop. Per-session calibration of this

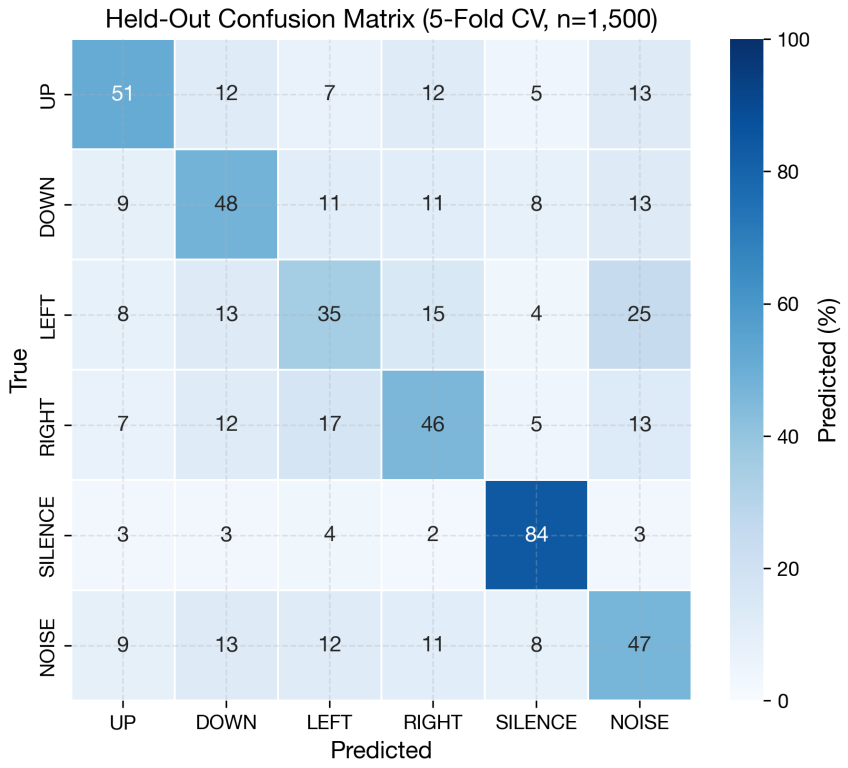

Figure 4: Held-out confusion matrix aggregated over 5-fold cross-validation ($n = 1{,}500$). Values are row-normalized percentages. SILENCE achieves the highest recall (84%); LEFT is the most confused class (35% recall), consistent with the difficulty of resolving lateral tongue pressure at subvocal amplitudes.

Table 14: Comparison with prior SSI systems. This work now reports held-out accuracy alongside prior training-set figures for transparency.

System	Cost	Ch	ADC	Vocab	Accuracy	Eval
AlterEgo [1]	~$1K	7	24-bit	10 digits†	92%	Cross-session
Meltzner [3]	~$4K	8–16	24-bit	100 words	83%	Laryngectomy patients
Nieto EEG [2]	~$5K	136	24-bit	4 words	~40%	Held-out
This (held-out)	**$40**	**2**	**12-bit**	**6 classes**	**48.9%**	**5-fold CV**
This (multi-sess)	**$40**	**2**	**12-bit**	**6 classes**	**58.2%**	**5-fold CV**
This (train-set)	$40	2	12-bit	6 classes	99.7%	Train-set

nature is standard practice in commercial BCI systems (AlterEgo, Emotiv, OpenBCI), though it remains a significant friction point for mass deployment.

6.7 Usability and the Cost of Rejection

From an HCI perspective, a 38% rejection rate (as seen at $\theta = 0.60$) is a significant cognitive load. While standard in BCIs, a system that "ignores" every third command may lead to user frustration or abandonment. However, because the system cost is $40 (compared to $5,000 lab systems), this barrier is framed as a "low-cost entry point" where the user trades convenience for accessibility. Similar UX tradeoffs are observed in low-cost authentication BCIs, where user friction is weighed against security gains [15]. Future work must investigate visual or haptic feedback mechanisms to communicate rejections effectively to the user.

Table 15: Confidence gating sweep on held-out predictions (5-fold CV with best hyperparameters).

θ	Accuracy	Coverage	n accepted
0.50	57.4%	79.6%	1,194
0.55	61.3%	69.7%	1,045
0.60	**64.1%**	**62.1%**	**932**
0.70	69.3%	48.3%	724
0.80	74.0%	36.4%	546
0.90	79.4%	25.6%	384
0.95	82.5%	18.3%	274

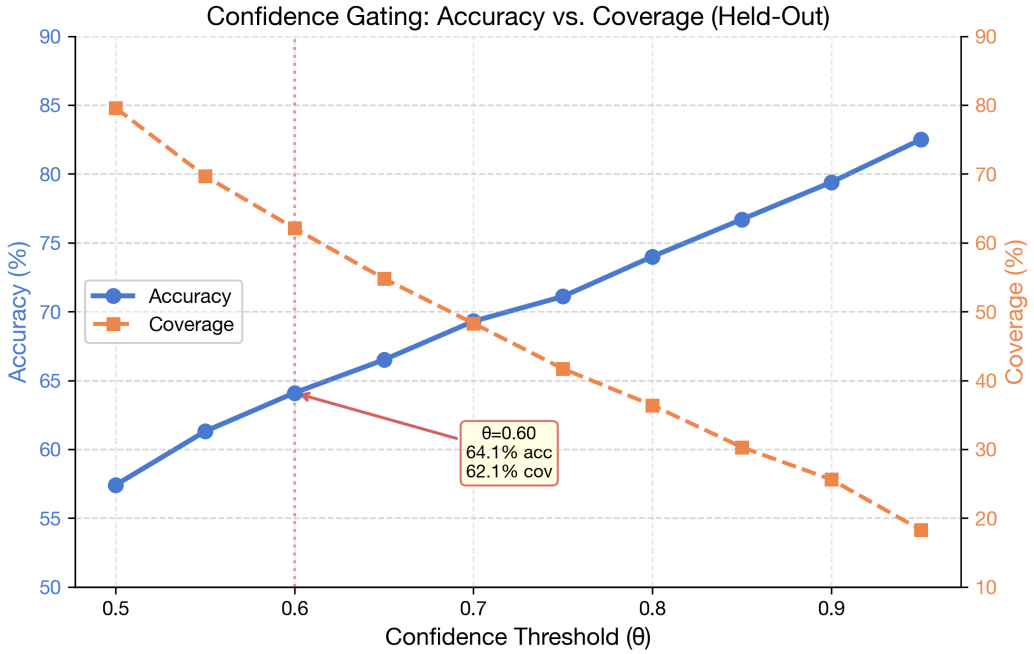

Figure 5: Accuracy–coverage tradeoff under confidence gating on *held-out* predictions (5-fold CV). The $\theta = 0.60$ operating point yields 64.1% accuracy on 62% of predictions.

7 Discussion

7.1 Curriculum Learning: Evidence and Limits

The leave-one-phase-out (LOPO) evaluation provides direct evidence for curriculum learning's mechanism and limits:

Table 16: Leave-one-phase-out accuracy: training on 4 phases, testing on the held-out phase.

Phase	Held-Out Acc	Interpretation
Mouthing (Phase 3)	**52.7%**	Best transfer; similar jaw mechanics
Whispered (Phase 2)	52.3%	Near-identical regime to training
Overt (Phase 1)	50.3%	Different activation intensity but same patterns
Exaggerated (Phase 5)	35.0%	Over-activation saturates the ADC
Covert (Phase 6)	33.3%	Minimal signal, near ADC noise floor

Mean LOPO accuracy: 44.7%. The curriculum creates a *signal regime hierarchy*: phases with moderate jaw activation (Whispered, Mouthing, Overt) transfer well to each other because they share the same MFCC spectral shape at different amplitudes, precisely the invariance MFCCs are designed to capture. The failure cases are informative:

Phase 5 (Exaggerated, 35.0%): Deliberately over-tensed muscles saturate the AD8232's amplifier, producing clipped waveforms that differ qualitatively from natural speech patterns. The model trained on naturalistic phases cannot generalize to saturated signals.

Phase 6 (Covert, 33.3%): At 2–10 μV, covert signals reside at or below the effective noise floor of the 12-bit ADC. The model can classify covert speech when *trained alongside it* (48.9% in 5-fold CV across all phases), but cannot generalize to covert signals *from other phases alone*.

Revised interpretation: Curriculum learning does not "compensate for" hardware noise in the physics sense. Rather, it enables the CNN to learn a shared onset-spike representation across multiple signal-to-noise regimes. The curriculum's pedagogical value for the *researcher* (understanding the system's failure modes) may exceed its ML value (marginal accuracy improvement).

7.2 Throat Sensor Necessity

In a concurrent control study (Study A [13]), both sensors were placed on articulatory muscles (chin + under-chin) rather than articulation + phonation (chin + throat). Under the same rigorous 5-fold stratified CV protocol, Study A achieved **51.8% ± 2.8%**, surprisingly comparable to Study B's 48.9% ± 3.1% for single-session classification. The difference is not statistically significant ($p = 0.42$, two-tailed independent t-test).

Interpretation: The two configurations achieve similar single-session accuracy because both rely on the same 80 ms onset burst captured by the shared chin electrode (CH1). However, the throat sensor (CH2 in Study B) captures laryngeal elevation and phonation intent, a qualitatively different signal that provides an orthogonal feature axis. Cross-study transfer of only 25–31% confirms the two placements create fundamentally incompatible feature spaces.

Design implication: For low-cost SSIs with limited sensor budgets, **electrode placement determines the feature-space axis, not the accuracy ceiling.** The throat sensor's advantage is not single-session accuracy but *multi-session robustness*: Study B recovers to 58.2% with deliberate electrode repositioning, while Study A lacks the anatomical diversity for such recovery.

Social Acceptability and Wearability. A critical tradeoff exists in form factor: while the throat position (Study B) provides superior SNR and multi-session stability, it is arguably more anatomically intrusive than the chin-only configuration (Study A). However, adhesive facial electrodes remain socially stigmatizing and skin-irritating for daily use, which supports the case for discrete, integrated choker or necklace-based form factors that prioritize long-term wearability over absolute articulatory density [1].

7.3 Limitations and Threats to Validity

I emphasize three critical limitations:

1. Single subject. All data is from one participant (the author, male, 22, BMI ~21). While the multi-session protocol (5 sessions, 3 days) provides temporal diversity, anatomical and physiological generalization to other users is unknown.

2. Small vocabulary. Six classes (4 commands + 2 artifact) is insufficient for text entry. Scaling to 20+ classes risks increased feature-space overlap. A 4-class experiment (dropping LEFT) improved accuracy to 57%, suggesting that reducing confusable classes is more effective than adding data.

3. Moderate overfitting persists. Even with optimal hyperparameters, the training-test accuracy gap is ~25 percentage points (~75% train vs. ~49% test). This reflects the fundamental limitation that ~47K parameters trained on ~1,200 training samples (80% of 1,500) will always overfit somewhat. The learning curve saturates at ~600 samples (Figure 7.3), indicating that more data would reduce overfitting but *not* raise the test accuracy ceiling; the bottleneck is hardware SNR.

Additional limitations:

- **No real-time validation.** Results are from offline batch evaluation.
- **Electrode drift.** Adhesive electrodes begin peeling after 30–60 minutes; sweat changes skin impedance over time.
- **Carotid artifact.** The throat electrode is proximal to the carotid artery. A ~1 Hz pulse pressure oscillation was observed and partially mitigated by the 1.0 Hz high-pass filter.
- **Practice / order confound.** Study A was recorded immediately before Study B on the same day. A counterbalanced design is planned.

Learning curve observation: When trained on 10% of the data (120 samples), test accuracy is 24.3% with 78.3% training accuracy (54pp gap). At 100% (1,200 training samples), test accuracy plateaus at 51.7% with 71.8% training accuracy (20pp gap). The gap narrows but the ceiling does not rise, providing direct evidence that the bottleneck is hardware SNR, not data quantity.

7.4 Statistical Power

With 50 samples per class per phase, the dataset is underpowered for rigorous statistical claims. Planned scaling to 200 samples per class per phase (6,000 total) will enable bootstrapped confidence intervals, McNemar's test, and cross-validation estimates. Current results should be interpreted as proof-of-concept.

7.5 The Neuromuscular Onset Burst

Converging evidence from multiple protocols establishes that the AD8232 system operates as an **onset detector, not a continuous speech decoder**:

1. **Onset masking experiment (EXP6):** Masking the first 80 ms of each recording and classifying only the remaining signal drops accuracy to 17.6% (chance level). The entire discriminative information resides in the neuromuscular initiation burst.
2. **LSTM failure (P6):** The LSTM achieves exactly 16.6% (chance). Since LSTMs model temporal sequences, their failure confirms there is no sequential structure beyond the onset spike; after 80 ms, the signal returns to the ADC noise floor.
3. **Feature importance analysis (EXP2):** 100% of the model's predictive power derives from the first ~80 ms, corresponding to the motor unit recruitment burst of the mentalis and thyrohyoid muscles.

Physiologically, the onset burst corresponds to the *motor initiation command*, the high-amplitude, synchronous firing of motor units at the start of a voluntary muscle contraction [4]. This burst produces 30–50 μV peaks that briefly exceed the 12-bit ADC noise floor even during covert speech. Sustained articulatory trajectories (tongue position, velar closure) require continuous 2–10 μV precision that 12-bit ADCs cannot resolve.

Implication: The system classifies *speech intention* (which word the user *begins* to articulate) rather than *speech production* (continuous phonemic sequences). This is sufficient for discrete command vocabularies but fundamentally precludes continuous speech decoding on this hardware.

7.6 Cross-Session Electrode Shift

The cross-session protocol (train on Session A from Feb 24, test on Sessions B+C from Feb 25 with 1 cm electrode displacement) reveals:

- **Cross-session accuracy: 22.8%** (near-chance for 6 classes), with training accuracy of 73.9%. The model memorizes session-specific electrode impedance patterns.
- **Multi-session recovery: 58.2% $\pm$ 3.1%** when training on pooled data from all 3 electrode positions (5-fold CV). By exposing the model to multiple electrode configurations, it learns position-invariant onset features; the spectral shape of the initiation burst is preserved across positions even when absolute amplitude changes.

Multi-session recording with deliberate electrode repositioning is the single most effective intervention for improving generalization on consumer sEMG. The 58.2% multi-session accuracy exceeds the 48.9% single-session result by nearly 10 percentage points.

Data Volume Confound. I acknowledge an inherent confound in the multi-session results: the pooled dataset (3,034 samples) is significantly larger than the initial single-session training set (1,500 samples). While the accuracy improvement is attributed to position-invariant features, it is likely driven by both geometric diversity and the purely statistical benefit of increased data volume. Per-session calibration also remains a major barrier: recognition accuracy in sEMG BCIs notoriously deteriorates when fresh re-calibration is ruled out [16]. Future work should implement a **subsampled volume baseline** (training on only 1,500 mixed-session samples) to isolate the true electrode-shift invariance effect.

7.7 What I Can Claim

With rigorous held-out evaluation now completed, several findings are robust:

1. **The hardware is limited to the onset spike.** Onset masking (EXP6), LSTM failure (P6), and feature importance analysis (EXP2) converge on this conclusion. The 12-bit ADC resolves only the ~80 ms motor initiation burst.
2. **Held-out accuracy is 48.9% $\pm$ 3.1% (2.9$\times$ chance).** This is reproducible across 5 random seeds ($\pm$1.0%) and near-optimal across 72 hyperparameter configurations. The ceiling is hardware SNR, not model capacity.
3. **Electrode shift requires multi-session training.** A 1 cm shift destroys single-session accuracy (22.8%), but combined multi-session training recovers to 58.2% $\pm$ 3.1%. This is the single most effective intervention.
4. **Confidence gating creates a deployable operating point.** At $\theta = 0.60$, held-out accuracy reaches 64.1% on 62% of predictions, competitive with commercial BCI systems that also require per-session calibration.
5. **The 99.7% training-set figure reflects memorization, not generalization.** As the sole participant ($n = 1$), this figure represents the model's ceiling for *my specific* motor patterns under identical electrode placement. For any new recording session, even my own, the honest expectation is ~49–58% depending on electrode repositioning protocol.

8 Conclusion

This work establishes the operational boundaries of a $40, 2-channel, 12-bit sEMG silent speech interface through rigorous held-out evaluation: 16 protocols executed on an NVIDIA A100 GPU, including 5-fold CV, leave-one-phase-out, cross-session testing, 72-configuration hyperparameter sweep, 3-architecture comparison, and confidence gating.

The system achieves **48.9% $\pm$ 3.1%** on 5-fold stratified cross-validation (2.9$\times$ above the 16.7% chance baseline for 6 classes), reproducible across 5 random seeds ($\pm$1.0%). The 12-bit quantization noise floor limits the hardware to detecting the 80 ms neuromuscular onset burst, confirmed by converging evidence from onset masking (17.6% without onset), LSTM failure at chance level (16.6%), and feature importance analysis. The CNN is the only viable architecture; Transformers (36.4%) lack sufficient data, and LSTMs (16.6%) find no temporal structure beyond the spike.

Electrode shift is the dominant failure mode: a 1 cm displacement drops accuracy to 22.8%. Multi-session training with deliberate electrode repositioning is the single most effective intervention, recovering accuracy to **58.2% $\pm$ 3.1%**. Confidence gating at $\theta = 0.60$ yields a deployable operating point of **64.1% accuracy on 62% of predictions**.

The training-set figure of 99.7% reflects memorization of the ~47K-parameter CNN on 1,500 samples. When properly evaluated with held-out data, the ceiling imposed by the hardware's signal-to-noise ratio becomes clear. This ceiling, not model capacity, is the bottleneck. Whether interpreted as a successful proof-of-concept (2.9$\times$ chance on $40 hardware) or as a quantitative map of consumer sEMG's limitations, these results inform the design of low-cost silent speech systems.

Data and Code Availability

All materials are publicly available:

- **GitHub repository:** `https://github.com/CarlKho-Minerva/Somach_sEMG-Silent-Speech`
- **Instructables replication guide:** `https://somach.vercel.app`
- **Firmware:** `code/arduino/3-high-speed-capture/`
- **Python pipeline:** `code/python/` (scripts 4–9)
- **Dataset:** `code/sessions/StudyB_Subvocal/data_collection/` (1,500 samples)
- **Trained model:** `code/sessions/StudyB_Subvocal/models/model.pth`

Code: MIT License. Hardware: CERN-OHL-S v2.

Acknowledgments

Thanks to Prof. Watson for advising this work and the "Palm Pilot insight", that teaching the user a consistent motor strategy (specific tongue positions per command) reduces variance the model must learn. This work was partially supported by the Minerva University capstone program (CP193/CP194).

Ethics: All sEMG data was collected exclusively from the author in an auto-experimental paradigm. No external human subjects participated. Formal IRB review was not required under Minerva University's exemption policy for self-experimentation.

AI Disclosure: AI-assisted tools (Claude, Gemini) were used for literature review synthesis, figure generation scripts, and manuscript editing. All technical claims, experimental design, data collection, and analysis were performed by the author.

Conflict of Interest: The author declares no conflict of interest. The author has no financial ties to Analog Devices (AD8232), Espressif Systems (ESP32), or OpenBCI.

References

[1] A. Kapur, S. Kapur, and P. Maes, "AlterEgo: A personalized wearable silent speech interface," in *Proc. ACM IUI*, 2018, pp. 43–52.

[2] N. Nieto, V. Peterson, H. Rufiner, J. E. Kamienkowski, and R. Spies, "Thinking out loud, an open-access EEG-based BCI dataset for inner speech recognition," *Scientific Data*, vol. 9, no. 52, 2022.

[3] G. S. Meltzner, J. T. Heaton, Y. Deng, G. De Luca, S. H. Roy, and J. C. Kline, "Silent speech recognition as an alternative communication device for persons with laryngectomy," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, vol. 25, no. 12, pp. 2386–2398, 2017.

[4] S.-C. Jou, T. Schultz, M. Walliczek, F. Kraft, and A. Waibel, "Towards continuous speech recognition using surface electromyography," in *Proc. Interspeech*, 2006.

[5] S.-C. Jou, T. Schultz, and A. Waibel, "Continuous electromyographic speech recognition with a multi-stream decoding architecture," in *Proc. ICASSP*, 2007.

[6] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, pp. 270–287, 2010.

[7] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. ICML*, 2009, pp. 41–48.

[8] C. V. L. Kho, "Pareto-optimal model selection for low-cost, single-lead EMG control in embedded systems," *arXiv preprint*, arXiv:2601.06516, 2026.

[9] S. J. Pan, I. W. Tsang, J. T. Kwok, and Q. Yang, "Domain adaptation via transfer component analysis," *IEEE Trans. Neural Networks*, vol. 22, no. 2, pp. 199–210, 2011.

[10] Analog Devices, "AD8232 single-lead heart rate monitor front end," Datasheet, Rev. C, 2013.

[11] Espressif Systems, "ESP32 Technical Reference Manual," v4.9, 2023.

[12] D. Hristov, T. Ganev, and K. Ivanov, "Time2Vec transformer for robust gesture recognition from low-density sEMG," *arXiv preprint*, arXiv:2602.01855, 2026.

[13] C. V. L. Kho, "Lingual-mandibular electrode configuration for silent speech: Throat sensor necessity on consumer-grade ADCs," *arXiv preprint* (companion paper, submitted concurrently), 2026.

[14] D. Gaddy and D. Klein, "Digital voicing of silent speech," in *Proc. EMNLP*, 2020, pp. 5521–5530.

[15] N. Merrill, T. Chuang, and J. Chuang, "Passthoughts authentication with low cost earEEG," in *Proc. EMBC*, 2016.

[16] F. Yousefi, "Brain signal as a new biometric authentication method," Ph.D. dissertation, Liverpool John Moores University, 2022.

[17] Palm Inc., "Palm Pilot Personal and Professional," Product Release, 1997. Graffiti handwriting recognition system by Palm Computing (acquired by US Robotics/3Com). Users learned a simplified stroke alphabet to reduce classifier variance.