POLITECNICO DI TORINO

Master's Degree in Computer Engineering



Politecnico
di Torino

Master's Degree Thesis

From Biosignals to Words: Exploiting Novel Deep Learning Architectures for Speech Understanding

Supervisors

Prof. Alessio BURRELLO

Prof. Daniele Jahier PAGLIARI

PhD. Giovanni POLLO

Candidate

Carola BONAMICO

Academic Year 2025/2026

Abstract

Silent speech (SS) interfaces decode biosignals of speech production into text without sound. One possibility is to use the neck surface electromyography (sEMG) that suits wearable devices and it is not invasive, capturing articulatory muscle activity even during silent mouthing. However, existing wearable systems based on sEMG recognize a few isolated commands, with limited evaluation after repositioning. This thesis extends them to closed-set classification and continuous recognition of sentences. The first contribution is a multimodal corpus including seven participants, six sessions collected at least one day apart and devices repositioned between them. Each participant recorded fifteen words and twenty sentences, ten times each, aloud and silently, for 29 400 utterances over about 39 h 40 min. We used two BioGAP-Ultra units recorded in parallel, one with a fourteen-channel sEMG neckband, one with a sixteen-channel EEG headband, both dry-electrode, each with a microphone to record synchronized EEG, sEMG, and audio signals. The second contribution is the alignment of the two devices to a mean offset of 0.12 ms without drift, and a 7.7 ms delay inside each unit. The EEG takes no part in the decoding, which reads sEMG alone. The third contribution is a decoding pipeline in which a convolutional backbone feeds a Bidirectional LSTM (BiLSTM) or a Transformer encoder. The use of temporal and frequency domain signals has been explored, together with the utilization of different losses, the connectionist temporal classification (CTC) and the cross-entropy (CE). The best configuration uses Short-time Fourier Transform (STFT) as preprocessing and the CTC loss. For closed-set classification of twenty sentences and rest, the Transformer performs best, at an average accuracy across participants of 85.8 ± 5.2% in vocalized and 82.1 ± 10.5% silent across sessions, versus 94.2 ± 3.9% and 90.7 ± 6.1% in intra-session evaluation. The silent-to-vocalized gap is 3.7 points across sessions, against 11.8 in the published word-level baseline. For continuous recognition, the recurrent model achieves 16.3 ± 5.0% and 22.4 ± 14.1% word error rate (WER) across sessions, and 6.9 ± 4.5% and 12.2 ± 8.5% with the intra-session protocol, with approximately 911000 parameters and no language model. The Transformer achieves comparable performance for silent speech across sessions at 21.2 ± 12.6%. The fourth contribution is a causal onset detector, which avoids needlessly executing the decoding pipeline when the user is not speaking.

Table of Contents

List of Tables V

List of Figures VII

Acronyms IX

1 Introduction 1
- 1.1 Speech Production and Neuromuscular Basis of sEMG 2
- 1.2 Modalities for Silent Speech Interfaces 3
- 1.3 EMG-based Silent Speech Recognition 3
  - 1.3.1 From Neuromuscular Foundation to Sensing Techniques . . . 3
  - 1.3.2 Limitations of sEMG 5
  - 1.3.3 sEMG for SSR: Hardware and Form Factors 6
  - 1.3.4 sEMG for SSR: Data Collections and Algorithms 7
  - 1.3.5 Algorithms. 7
  - 1.3.6 Towards End-to-End Wearable Systems 8
- 1.4 EEG-based Silent Speech Recognition 8
  - 1.4.1 General Sensing Principles 8
  - 1.4.2 EEG for SSR: Related Works 9
- 1.5 Multimodal EMG–EEG Systems 10
- 1.6 Motivation of the Thesis 11

2 EMG and EEG Acquisition: Materials 12
- 2.1 Hardware Interfaces 12
  - 2.1.1 EMG Neckband 12
  - 2.1.2 EEG Headband 13
- 2.2 Software and Multimodal Synchronization 14
  - 2.2.1 Alignment and Delays 14
- 2.3 Data-Collection Protocol 18
  - 2.3.1 Vocabulary Selection 18
  - 2.3.2 Session Structure and Participants 21

3 Closed-Set Classification: Methods 23
3.1 Problem Formulation . . . 23
3.2 Signal Processing Pipeline . . . 24
3.2.1 Segmentation and Windowing . . . 24
3.3 Feature Extraction and Representations . . . 25
3.4 Training Objectives . . . 28
3.4.1 Cross-entropy and CTC . . . 28
3.5 Deep Learning Models . . . 29
3.5.1 Reference Model . . . 30
3.5.2 Convolutional Backbone . . . 30
3.5.3 Bidirectional LSTM . . . 32
3.5.4 Transformer Encoder . . . 33
3.6 Decoding . . . 36
3.7 Evaluation Settings and Metrics . . . 37
3.7.1 Cross-Validation Protocols . . . 37
3.7.2 Training Schedule and Class Balancing . . . 37
3.7.3 Evaluation Metrics and Representational Probing . . . 38
3.8 Automatic Speech Detection from EMG . . . 39
3.8.1 Detection Algorithm . . . 40
3.8.2 Evaluation against the Trigger . . . 41

4 Closed-Set Classification: Results 43
4.1 Order of the Experiments . . . 43
4.2 Baseline Reproduction and Validation of the Corpus . . . 44
4.3 Training Objective and Sequence Model . . . 45
4.4 Input Representations . . . 47
4.4.1 STFT Representation . . . 47
4.4.2 MFCC Representation . . . 49
4.5 BiLSTM versus Transformer at Equal Number of Parameters . . . 50
4.5.1 Learned Representation . . . 51
4.6 Decoding the CTC Classifier . . . 52
4.7 Ablation Studies . . . 55
4.7.1 Session-Count Scaling . . . 55
4.7.2 Sliding-Window Augmentation . . . 56
4.8 Trigger-Free Speech Detection . . . 58
4.8.1 Corpus Yield and Classification Performance . . . 58

5 Sentence Continuous Recognition: Methods 62
5.1 Problem Formulation . . . 62
5.2 From Classification to Continuous Recognition . . . 63
5.3 Decoding . . . 63

5.4 Evaluation Settings and Metrics . . . 63
5.4.1 Cross-Validation Protocols . . . 63
5.4.2 Training Schedule and Class Balancing . . . 64
5.4.3 Evaluation Metrics . . . 65

6 Sentence Continuous Recognition: Results 67
6.1 Continuous Recognition Performance . . . 67
6.2 Decoder Sweep: Greedy Decoding versus Prefix Beam Search . . . 69
6.3 BiLSTM versus Transformer at Equal Parameters . . . 70

7 Conclusion 72

A Data-Collection Protocol 74
A.1 Corpus . . . 74
A.1.1 Phase A: Words . . . 74
A.1.2 Phase B: Sentences . . . 74
A.2 Lexical Analysis . . . 75
A.3 Trigger Module Configuration . . . 77
A.4 Trial and Session Timing . . . 78
A.5 Shuffling and Balance Rules . . . 78
A.6 Nominal Dataset Size . . . 80
A.7 Scheduling Rules and Realized Schedule . . . 80

B Word and Sentence Lexicon 82

C Per-Participant Results 83
C.1 Sentence Closed-Set Classification, per Participant . . . 83
C.2 Sentence Continuous Recognition, per Participant . . . 84
C.3 Per-Participant Confusion Matrices . . . 84

Bibliography 87

List of Tables

1.1 Hardware of representative wearable EMG-based SSI systems. . . . 7
1.2 Dataset and algorithms of representative wearable EMG-based SSR systems. . . . 8
1.3 Sentence-level and open-vocabulary silent speech systems. . . . 9

2.1 Acquisition parameters of the two units. . . . 15
2.2 Muscular-to-acoustic lead at three instants of one block. . . . 18
2.3 Representative data-collection protocols surveyed for the design of this study. . . . 20
2.4 Design parameters of the acquisition protocol. . . . 21

3.1 Width of the output layer, by objective and task. . . . 30
3.2 The three convolutional stacks. . . . 33
3.3 The three sequence stages and the resulting model sizes. . . . 35
3.4 Constants of the trigger-free detector. . . . 41

4.1 Baseline reproduction and validation of the new corpus. . . . 45
4.2 Input domain, sequence stage and objective. . . . 48
4.3 Length of the STFT analysis window. . . . 49
4.4 Comparison of the two sequence models at an equal number of parameters. . . . 51
4.5 Nearest-centroid classification of the learned representations. . . . 52
4.6 Impact of decoding strategies on classification performance. . . . 55
4.7 Sliding-window augmentation at constant training-set size. . . . 58
4.8 Sliding-window augmentation at augmented training-set size. . . . 59
4.9 The detector as a binary speech-against-rest classifier. . . . 61

6.1 Continuous recognition performance. . . . 69
6.2 Impact of decoding strategies on recognition performance. . . . 70

A.1 Phase A word set. . . . 75
A.2 Phase B sentence set. . . . 76

A.3 Content-word frequency across the Phase B sentences. 76
A.4 Trigger configurations used for the two acquisition phases. 78
A.5 Timing parameters of the acquisition protocol. 79
A.6 Within-session block structure. 79
A.7 Nominal dataset size. 80

B.1 Sentence lexicon, in file order. 82

C.1 Per-participant closed-set classification. 83
C.2 Per-participant continuous recognition. 84

List of Figures

1.1 Overview of the thesis pipeline 2
1.2 Building blocks of a wearable sEMG interface. 3
1.3 Frequency composition of the sEMG signal. 4
1.4 Frequency composition of the EEG signal. 10

2.1 Multimodal acquisition setup. 12
2.2 The SilentWear wearable sEMG platform used in this thesis. 13
2.3 The multimodal acquisition setup. 15
2.4 Controlled setup for the instrumental delay. 17
2.5 Muscular-to-acoustic lead at three instants of one block. 19
2.6 Structure of the data collection. 22

3.1 Trigger-driven cropping, segmentation, and windowing. 25
3.2 From the raw recording to the tensor entering the network. 26
3.3 The three architectures compared in this thesis. 31
3.4 Tensor shapes inside the network. 34
3.5 The classification pipeline, from the head to the class. 36

4.1 Vocalized words and sentences with related triggers. 44
4.2 The three input representations of a window. 47
4.3 Learned representations across sessions. 53
4.4 From per-frame CTC posteriors to a class label. 54
4.5 Session-count scaling on the single word subset. 56
4.6 Sliding-window augmentation at constant training-set size. 57
4.7 Sliding-window augmentation at augmented training-set size. 57
4.8 Cue-anchored and trigger-free windows for a single sentence. 60
4.9 The detector against the cue boxes of one recording. 61

5.1 The continuous recognition pipeline, from the head to the error rates. 64

Acronyms

SSI
Silent Speech Interface

SSR
Silent Speech Recognition

EMG
Electromyography

sEMG
surface Electromyography

EEG
Electroencephalography

ECoG
Electrocorticography

HMI
Human–Machine Interface

SCM
Sternocleidomastoid

ITI
Inter-Trigger Interval

STFT
Short-Time Fourier Transform

MFCC

Mel-Frequency Cepstral Coefficients

HPF

High-Pass Filter

DFT

Discrete Fourier Transform

DCT

Discrete Cosine Transform

FFT

Fast Fourier Transform

DTW

Dynamic Time Warping

SNR

Signal-to-Noise Ratio

ML

Machine Learning

DL

Deep Learning

CNN

Convolutional Neural Network

RNN

Recurrent Neural Network

LSTM

Long Short-Term Memory

BiLSTM

Bidirectional Long Short-Term Memory

CTC

Connectionist Temporal Classification

CE

Cross-Entropy

LDA

Linear Discriminant Analysis

LM

Language Model

LLM

Large Language Model

ReLU

Rectified Linear Unit

GELU

Gaussian Error Linear Unit

WER

Word Error Rate

CER

Character Error Rate

SER

Sentence Error Rate

ASR

Automatic Speech Recognition

CV

Cross-Validation

LOSO

Leave-One-Session-Out

t-SNE

t-distributed Stochastic Neighbor Embedding

FLOP

Floating-Point Operation

HDF5

Hierarchical Data Format version 5

CSV

Comma-Separated Values

YAML

YAML Ain't Markup Language

JSON

JavaScript Object Notation

SoC

System-on-Chip

AFE

Analog Front-End

ADC

Analog-to-Digital Converter

PGA

Programmable-Gain Amplifier

BLE

Bluetooth Low Energy

IIS

Integrated Systems Laboratory

Chapter 1

Introduction

Speech represents the primary human interface, but it fails in various situations such as noisy environments, scenarios requiring privacy, or for individuals having phonation disorders. A SSI addresses these limitations by capturing physiological signals and decoding them into text or audio without the need for sound [1, 2]. Among the available sensing modalities of SSI, sEMG, which targets facial and neck muscles, is the most promising because it offers a non-invasive, wearable approach to record the motor commands of articulation even during unvocalized speech [3, 4, 5]. Despite this, however, many of the SSR systems are limited to a closed set of commands or words collected via a single sensing modality, and the models are evaluated on data from a single session [4, 6, 7]. Moving to the comprehension of longer texts, such as at the sentence level, is non-trivial, requiring the modeling of temporal structures and the decoding of open vocabularies. Furthermore, the EEG signal allows for the collection of complementary data regarding the cortical planning of language, which is why it is useful to record both signals in parallel [8, 9, 10]. Specifically, this work builds upon the SilentWear wearable setup [4] to implement a multimodal acquisition framework and analyze silent speech at the sentence level. The main contributions are:

- **A multimodal dataset.** Synchronized sEMG neckband, EEG headband, and microphones capture parallel utterances, both vocalized and silent, at the word and sentence levels from multiple subjects.
- **Sentence-level sEMG processing.** The baseline neural architecture [4] is adapted from isolated command to sentences classification.
- **Analysis of sEMG-audio temporal alignment.** A controlled experiment demonstrates that acquisition latency is negligible compared to the lead time of muscle activity relative to the sound.

- **Architectural exploration and continuous recognition:** Recurrent sequence models (BiLSTM) and Transformers are compared using a shared convolutional backbone and an equivalent number of parameters. The task shifts from closed-set classification to continuous recognition, evaluating different decoding strategies.
- **Trigger-free onset detection:** A causal algorithm segments utterances using only the sEMG signal and is evaluated as a binary classifier to distinguish between speech and rest.

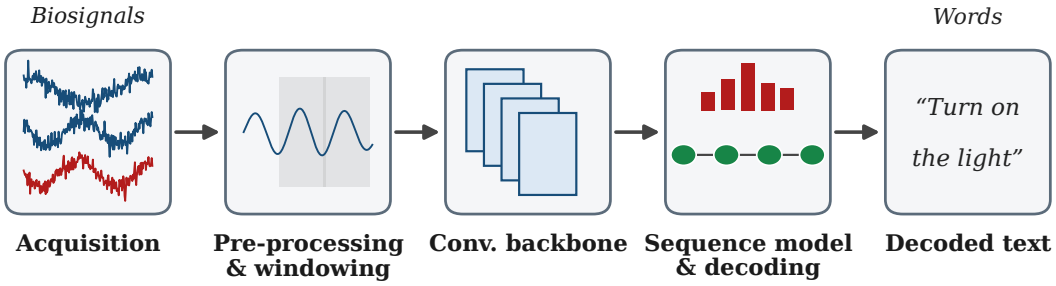


Figure 1.1: Processing pipeline developed in this thesis

The remainder of this chapter provides the theoretical background. Section 1.1 reviews speech production and sEMG. Section 1.2 surveys biosignal modalities. Sections 1.3 and 1.4 detail EMG and EEG systems. Section 1.6 motivates multimodal acquisition. The remaining chapters follow the pipeline in Figure 1.1: hardware (Chapter 2), classification (Chapters 3 and 4), continuous recognition (Chapters 5 and 6), and conclusions (Chapter 7).

1.1 Speech Production and Neuromuscular Basis of sEMG

Speech production requires a coordinated sequence of muscular activations. These contractions are triggered by action potentials propagating along muscle fibers. sEMG measures the spatial combination of these electrical fields at the skin surface. Because electrical potentials precede mechanical movement and sound, EMG encodes utterances before they become audible [11]. Section 2.2.1 empirically validates this temporal lead. For wearables, the neck provides optimal access to articulatory muscles, specifically the infrahyoid group and the SCM, via a textile neckband [3, 4]. These myoelectric signals persist during *silent* speech. When a speaker mouths words without airflow, the muscles still activate, allowing the interface to work where a microphone captures nothing. However, silent

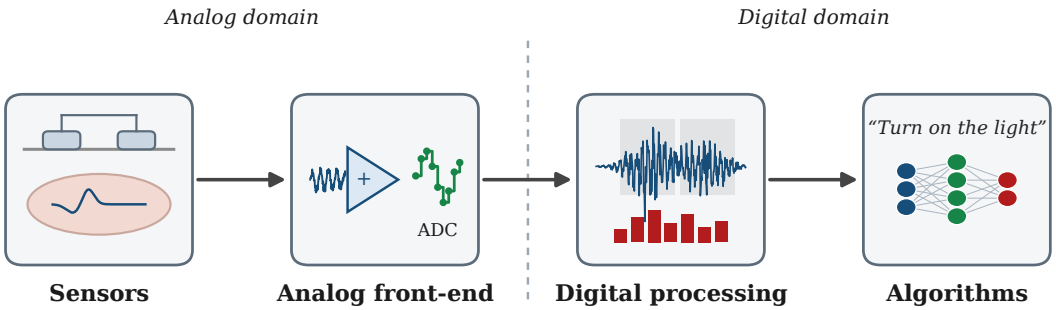


Figure 1.2: The four building blocks of a wearable sEMG interface, presented in the order an utterance traverses them.

articulation degrades signal quality, creating a core challenge addressed in this thesis (Section 1.3.2).

1.2 Modalities for Silent Speech Interfaces

Silent speech can be captured via several biosignals modalities, each monitoring a different physiological stage. EEG targets the central pathway, reflecting cortical speech planning [8, 12]. Surface EMG operates at the peripheral stage, recording articulatory muscle execution [13, 1]. Invasive techniques like ECoG offer higher fidelity but require surgery [14]. These modalities are complementary: EEG encodes early intent, while sEMG provides localized data synchronized with articulation. Fusing both domains improves decoding robustness [12, 9, 8]. However, for an ambulatory wearable, sEMG is the more practical primary modality due to its better SNR, reliable placement, and direct link to vocabulary [2, 6]. In these systems, EEG serves best as a supplementary channel [9, 10].

1.3 EMG-based Silent Speech Recognition

1.3.1 From Neuromuscular Foundation to Sensing Techniques

Translating neuromuscular activity into decoded text requires a specialized acquisition chain. Figure 1.2 illustrates the building blocks of a wearable sEMG interface [15]. *Sensors* transduce muscle potentials, which an AFE amplifies and digitizes [16]. *Digital processing* then structures the data for the *algorithmic* layer mapped onto discrete outputs. The ADC sets the sampling rate and resolution, establishing physical limits on signal quality and power consumption [13, 4]. Electrodes are placed over muscles recruited during speech, namely facial and neck

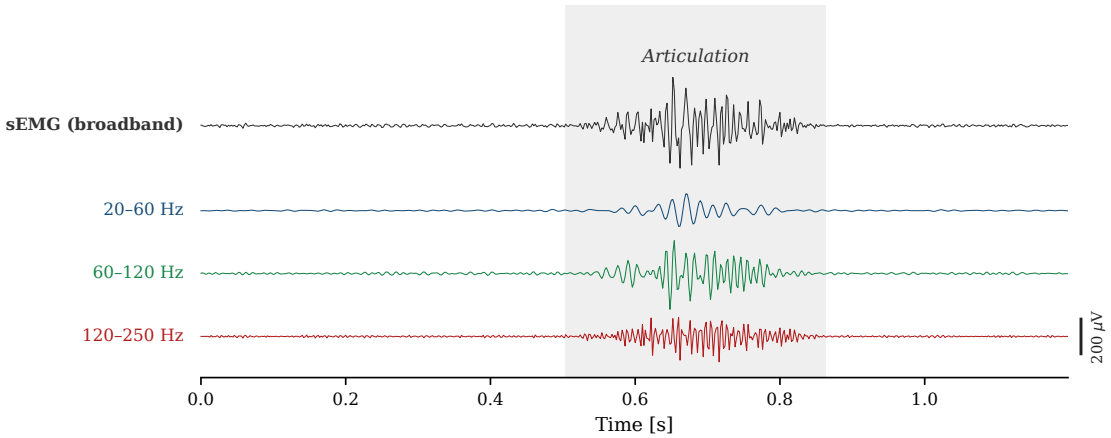


Figure 1.3: Frequency composition of a surface EMG signal. The trace is synthesized from spectral parameters in the literature [13, 18]; it is high-pass filtered at 20 Hz, notched at 50 Hz, and sampled at 500 Hz.

muscles [1, 3]. Specifically, these locations are chosen because they are the anatomical sites where electrical fields reach the skin with minimal attennuation. The electrode-skin interface (wet or dry) dictates contact impedance and SNR, while the topology (common-reference or differential) determines common-mode noise rejection. Section 1.3.3 analyzes how current systems navigate these trade-offs. Muscle potentials range from tens to hundreds of microvolts. Therefore an AFE is required to isolate these signals from environmental noise [13]. More specifically, an AFE is constituted by an instrumentation amplifier of high input impedance and high common-mode rejection ratio. After amplification, the signal is band-limited and digitized by adn ADC, typically at hundreds of hertz and 10 to 24 bits of resolution [1, 5]. These parameters follow from the morphology of the EMG signal itself, which is detailed in the subsequent paragraph.

Signal morphology. A surface electrode aggregates randomly timed motor-unit action potentials into a zero-mean *interference pattern* reflecting the underlying neural drive [13, 17]. Amplitude tracks contraction strength, alternating between a baseline of a few microvolts and bursts reaching hundreds of microvolts. The signal lacks discrete narrow-band rhythms: its energy spreads continuously, yielding similar non-oscillatory contributions across low, mid, and high-frequency sub-bands (Figure 1.3). The resulting power spectrum is a broad hump spanning 20–450 Hz, effectively empty below ∼20 Hz and concentrating most energy between 50 and 150 Hz [13, 18]. This morphology directly dictates the required front-end sampling rates and digital filter cutoffs [19].

Digital processing. Despite optimal AFE design, signals require high-pass and notch filtering to remove motion artifacts, baseline drift, and power-line interference [3, 7]. The filtered stream is then segmented into discrete windows. Historically, systems relied on handcrafted time- and frequency-domain features [5, 3], whereas modern architectures increasingly feed lightly processed representations directly into deep networks [6, 4]. A widely-used middle ground borrows acoustic speech transforms, specifically MFCC [20, 21, 1, 22]. Yet, transferring acoustic tools to EMG is debated. Indeed, research has shown plain spectral and time-domain features outperforming cepstral ones [23, 24], while more recently, mel filter bank energies have outscored the cepstra derived from them [25]. To address EMG's narrower bandwidth compared to speech, the filter bank is reduced [20]. The optimal representation for an EMG network is therefore an open question, which Section 4.4 addresses empirically for this corpus.

Algorithms. The final stage maps extracted representations onto linguistic constructs, acting as either a discrete classifier for a closed command vocabulary or a continuous recognizer for arbitrary-length sequences (surveyed in Section 1.3.4). Architecture choice is dictated by both raw accuracy and the deployment target, and unlike offline decoding, on-device inference via ultra-low-power SoCs imposes strict bounds on parameter counts and computational operations [4, 26].

1.3.2 Limitations of sEMG

Four unresolved challenges shape the methods in this thesis:

- **Electrode repositioning.** Removing and reapplying a device introduces placement shifts, causing electrodes to misalign from muscle fibers, contact impedance to fluctuate, and signal distributions to drift. This variability compromises model performance across separate recording sessions [3, 4]. Consequently, inter-session evaluation is more realistic to assess system robustness.
- **Electrode repositioning.** Physiological factors, including facial anatomy, adipose tissue thickness, skin condition, and articulation habits, vary between people. Consequently, a model trained on one speaker rarely generalizes to unseen individuals. The literature mitigates this variability through domain adaptation, adversarial training, transfer learning, and extensive data augmentation [27, 7, 28].
- **Silent versus vocalized articulation.** Silent articulation recruits muscles weakly and lacks laryngeal engagement. The drop in accuracy is documented across speaking modes and persists in current wearable systems [29, 4].

- **Recording conditions and crosstalk.** Electrodes capture spatial crosstalk from adjacent muscles. Artifacts from swallowing or head movements can mimic articulatory gestures [13]. These confounders must be managed through careful protocols and data annotation.

1.3.3 sEMG for SSR: Hardware and Form Factors

Research leverages the sEMG signal for two distinct objectives. Silent speech recognition maps biosignals directly to text through closed-set classification or continuous transcription. On the other hand, silent speech production synthesizes audio [30, 31, 15]. While production systems are routinely evaluated by passing synthesized audio through an ASR model to compute a WER, this metric confounds the biosignal decoder's performance with the external recognition model [30, 32]. Furthermore, system architectures are not monolithic. Data collected by low-power wearables can train computationally intensive offline models, including LLMs [14, 33]. Eventually, systems are differentiated by three core design choices: the sensing hardware, the dataset and algorithmic pipelines, and the proportion of processing executed locally on the body.

Sensors and form factors. Hardware design balances signal quality against user wearability. To maximize quality, AlterEgo uses seven wet, gold-plated electrodes with conductive paste located on the face, jaw, and neck [1]. However, recent developments prioritize comfort such as epidermal tattoos utilizing dry Au/Cr traces on conformable PET films, shifting rigid components to the ear [5]. In addition, headphone earmuffs featuring graphene and PEDOT:PSS textile electrodes [6] are also available. Recent research also targets the neck, an unobtrusive location exposing both infrahyoid and SCM muscles. Implementations range from textile neckbands pairing dry gold-plated electrodes with a wet reference [34] to the fully dry, conductive polymer elastomer neckband of SilentWear [3, 4], which serves as the primary interface for this thesis. The field is transitioning from rigid, wet setups to integrated, dry garments. However, dry contacts increase electrode-skin impedance and lower the baseline SNR, explaining why modern architectures increasingly rely on high channel counts and differential positioning.

Acquisition electronics. AlterEgo streams data via a standard OpenBCI board [1], while the tattoo and headphone systems rely on custom readout circuits, the latter interfacing a commercial EMG front-end to a WiFi-enabled ESP32-S3 microcontroller [5, 6]. The textile neckband in [34] runs on WANDmini [35], a wireless neuromodulation platform consuming 46 mW. In contrast, SilentWear uses the BioGAP-Ultra platform [16] executing simultaneous acquisition, on-device inference, and wireless streaming at 20.5 mW. As illustrated in Table 1.1, only

Table 1.1: Hardware of representative wearable EMG-based silent speech interfaces.

System	Form factor	Electrodes	Ch.	Acq. electronics	Power	Edge AI
AlterEgo [1]	Face–neck apparatus	Wet, gold-plated + paste	7 CR	OpenBCI	n.r.	No
Tattoo [5]	Facial tattoo + ear unit	Dry Au/Cr on PET	4 CR	Custom	n.r.	No
Necklace [34]	Textile neckband	Dry gold-plated + wet ref.	10 CR	WANDmini [35]	46 mW	No
Headphones [6]	Headphone earmuffs	Dry textile (PEDOT:PSS)	4 CR	Custom (ESP32-S3)	n.r.	No
SilentWear [4]	Textile neckband	Dry polymer, Ag/AgCl coating	14 diff	BioGAP-Ultra [16]	20.5 mW	GAP9
This work	EMG Neckband + EEG headband	As SilentWear	14 diff	BioGAP-Ultra ×2	—	No

two of the five evaluated systems disclose quantitative power profiles. Then, only SilentWear incorporates a processor capable of executing the decoding model on the edge.

1.3.4 sEMG for SSR: Data Collections and Algorithms

Data collections. The datasets in Table 1.2 vary in size and rigor and most systems evaluate on highly constrained: AlterEgo uses a ten-digit vocabulary (750 utterances per subject) [1], the epidermal tattoo tests 110 words on a single speaker [5], WANDmini extracts a 92.7% accuracy from just 110 total samples [34], and the headphone interface collects 100 utterances per subject [6]. However, none enforce sensor repositioning while SilentWear addresses this gap by recording 4800 utterances per subject (half vocalized, half silent) over multiple days, refitting the device between sessions [4]. In contrast, state-of-the-art continuous recognition requires hours of aligned data [30, 32], and massive neuromotor models achieve generalization only by training thousands of users [36].

1.3.5 Algorithms.

Closed-set algorithms have shifted from handcrafted features using LDA or random forests [5, 34] to deep learning, culminating in edge-optimized architectures like SilentWear's 15 k-parameter SpeechNet [1, 6, 4]. On the other hand, continuous recognition requires sequence modelling and CTC for unaligned inputs [37]. Building on foundational EMG-to-speech systems [30, 32], modern networks synthesize audio [31, 38] or achieve competitive error rates by pairing acoustic models with LLMs [14, 33]. Meanwhile, session variability is mitigated through high-density grids, transfer learning, adversarial training [39, 7, 27, 28], and edge-deployable foundation models [26]. These paradigms are shown in Table 1.2).

Table 1.2: Datasets and algorithms of the systems in Table 1.1.

System	Vocab.	Mode	Subj.	Rep.	Model	Acc. (%)
AlterEgo [1]	10 digits	Silent	10	No	CNN	92.0
Tattoo [5]	110 words	Silent	1	No	LDA	92.6
Necklace [34]	11 words	Voc.	2	No	Random forest	92.7
Headphones [6]	10 commands	Silent	10	No	1D SE-ResNet	96.0
SilentWear [4]	8 commands + rest	Voc. + silent	4	Yes	SpeechNet (CNN)	84.8 / 77.5
This work	15 words, 20 sentences	Voc. + silent	7	Yes	SpeechNet + BiLSTM / Transformer	Ch. 4, 6

1.3.6 Towards End-to-End Wearable Systems

Processing data locally on the edge offers three advantages: sensitive EMG transmits only discrete labels to safeguard privacy and eliminates bidirectional streaming latency. Despite that, only SilentWear assesses this (Table 1.1). Its 8-bit quantized SpeechNet model executes on the BioGAP-Ultra GAP9 accelerator, costing 63.9 μJ over 2.47 ms per inference for a total system budget of 20.5 mW [4]. In addition, Table 1.3 tracks sentence-level, open-vocabulary, and hybrid EEG and EMG interfaces (Section 1.4.2). For what concerns open-vocabulary WERs, they have improved from 68.0% for the original model [30], to 42.2% via convolutional-Transformer transduction [32], and 12.2% using LLM rescoring [14]. The *Decoder* column positions this thesis because while many systems attach large language models and low-power wearables restrict tasks to simple commands (Table 1.3), this work bridges both paradigms. It explores sentence-level sequence modelling while operating strictly within SilentWear's constrained, LM-free wearable regime.

1.4 EEG-based Silent Speech Recognition

1.4.1 General Sensing Principles

EEG records cortical electrical activity through electrodes placed on the scalp; compared to sEMG, the signal is weaker, spatially coarser and more susceptible to artefacts. Standard practice applies a band-pass filter to preserve the informative low-frequency components and uses multiple channels to compensate for the limited information provided by each channel [8, 12].

Signal morphology. Morphologically, EEG recorded at the scalp consists of continuous, low-amplitude fluctuations, typically a few tens of microvolts with clinical gel electrodes. Its activity is distributed across the following frequency

Table 1.3: Sentence-level and open-vocabulary silent speech decoding systems.

System	Signals	Corpus	Subj.	Decoder	Error
Gaddy & Klein [30]	EMG	Open vocab. (O)	1	ASR	68.0 WER
Gaddy & Klein [32]	EMG	Open vocab. (O)	1	ASR	42.2 WER
Benster et al. [14]	EMG	Open vocab. (O)	1	LLM	12.2 WER
Mohapatra et al. [33]	EMG	67 words (C)	1	LLM	49 WER
Li et al. [8]	EEG+EMG	Silent-speech corpus	8	Seq2seq	7.22 CER
Satterlee et al. [10]	EEG+EMG	4 sentences (C)	1	LM	4.75 SER
This work	EMG	20 sentences (C), open vocab. (O)	7	None	Ch. 6

bands: delta (0.5–4 Hz), theta (4–8 Hz), alpha (8–13 Hz), beta (13–30 Hz), and gamma (>30 Hz) (Figure 1.4) [40]. Because the power spectrum approximately follows a $1/f$ distribution [41], these components have very different amplitudes. Delta and theta account for most of the signal power, alpha appears as a ∼10 Hz oscillation, while beta and gamma contribute only small fluctuations. These properties have two implications for silent speech. First, most of the signal energy lies below ∼40 Hz, so the acquisition chain can use a corresponding low-pass range and relatively low sampling rates. Second, the low amplitude, the contribution of multiple cortical sources, and the lack of articulation-locked events make EEG more difficult to decode alone than sEMG [8, 12].

1.4.2 EEG for SSR: Related Works

Research on EEG-based silent speech decoding considers different articulation paradigms. In *overt* speech, the subject speaks aloud, providing a clear acoustic reference but also introducing EMG artifacts into the EEG signal [42]. In *silent* speech, the subject articulates without phonation, while *imagined* or covert speech involves only the internal representation of speech and therefore provides neither an acoustic signal nor a direct physical event for temporal alignment [2, 43]. Performance therefore depends strongly on the task and acquisition setup, and results should be interpreted relative to the corresponding chance level rather than directly compared across modalities. For example, a 16-channel cEEGrid system achieved 56.6% accuracy on a 64-word Japanese vocabulary across 24 subjects, compared with a 1.56% chance level, while a locked-in participant reached 47.3% accuracy and 76.0% top-5 accuracy in an online setting [44]. These results required 282 hours of training data, illustrating the data requirements of non-invasive EEG decoding.

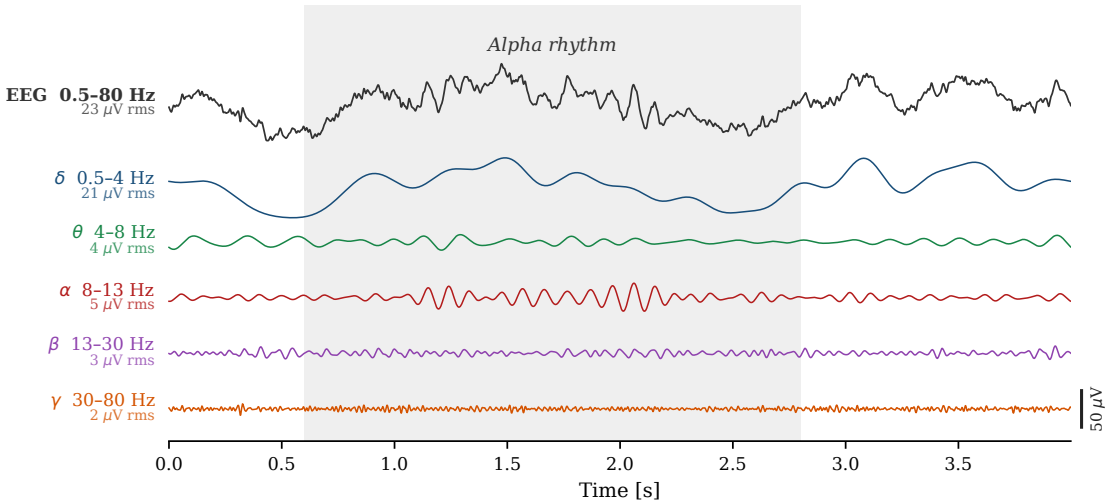


Figure 1.4: Frequency composition of a scalp EEG signal, synthesized from the composition and the relative band powers reported in the literature [40]. The shaded span marks the interval over which the alpha rhythm is present.

Other approaches reconstruct speech-related representations directly from EEG, including mel-spectrograms, without requiring explicit temporal alignment [43], while shared latent representations have also been used to transfer information between EEG and EMG [12]. At the other, intracortical systems provide much higher decoding performance by recording neural activity directly from the cortex. A recent multi-participant neuroprosthesis reduced relative WER using a shared Transformer decoder and achieved below 7% WER for new participants after fewer than 200 sentences of fine-tuning [45]. The gap between invasive and non-invasive approaches reflects the attenuation and spatial mixing introduced by the skull. For wearable silent speech interfaces, these limitations make sEMG an attractive alternative because it measures the peripheral muscular activity associated with articulation directly and does not require surgical intervention.

1.5 Multimodal EMG–EEG Systems

Integrating both modalities is advantageous because they capture different stages of the speech production process and can provide complementary information. This has been explored through convolutional fusion architectures [9], cross-domain transfer learning [12], and hybrid systems combining scalp EEG with facial and neck EMG for speech synthesis [8]. More recent wearable systems also combine multimodal sensing with language models for sentence-level recognition [10]. However, these systems differ substantially in their experimental settings. For example, the former reports an offline CER averaged over eight speakers, whereas the latter

evaluates sentence accuracy on only four sentences from a single subject and uses a LM for post-correction, as summarized in Table 1.3. Despite these advances, data availability remains a major limitation. The JapanEEG dataset provides 1020 hours of simultaneous scalp EEG, facial EMG, and audio recordings from overt and imagined Japanese speech [42]. However, it was collected from only three subjects over extended sessions using laboratory setups with 62–128 channels. Its facial EMG also consists of only three bipolar channels placed around the lips and eyes and is used to characterize articulatory artifacts rather than as a primary decoding modality. Therefore, a corpus combining a wearable dry-electrode neckband, multi-channel EMG, repeated sensor repositioning, and a task-oriented vocabulary is still lacking. Addressing this gap is the first objective of this thesis.

1.6 Motivation of the Thesis

The motivations of this thesis address the main gaps identified in current wearable sEMG speech interfaces: the lack of paired multimodal datasets, and the gap between isolated-command and continuous sentence decoding. A key motivation is the lack of paired datasets for studying multiple modalities under the same recording conditions. Simultaneous recording of sEMG, EEG, and audio enables accurate temporal alignment and provides reference targets for future speech reconstruction [31]. Collecting and synchronizing such a dataset is therefore a key objective of this thesis. While the present work focuses on sEMG-based decoding, the aligned EEG recordings are released to support future multimodal sensor-fusion studies. In addition, current wearable sEMG systems mainly classify isolated commands from a small number of speakers within a single session (Table 1.2), whereas continuous speech decoding is typically demonstrated on large single-speaker datasets using computationally expensive offline models (Table 1.3). This thesis addresses this gap by targeting continuous sentence decoding across multiple speakers and sessions, including sessions with device repositioning. Such repositioning can alter electrode contact impedance, and signal distributions (Section 1.3.2), making single-session evaluation overly optimistic. Models are therefore also evaluated on held-out sessions (Section 3.7).

Chapter 2

EMG and EEG Acquisition: Materials

This chapter details the hardware, software, and protocol used to collect the multimodal corpus. Data is acquired simultaneously via an EMG neckband and an EEG headband (Figure 2.1), each driven by an independent hardware interface equipped with a reference microphone.



Figure 2.1: The multimodal acquisition setup worn by a participant, featuring the EEG headband and the sEMG neckband.

2.1 Hardware Interfaces

2.1.1 EMG Neckband

Surface EMG is acquired with the SilentWear neckband, a 14-channel fully dry textile interface powered by the BioGAP-Ultra platform [4, 3, 16] (Figure 2.2).

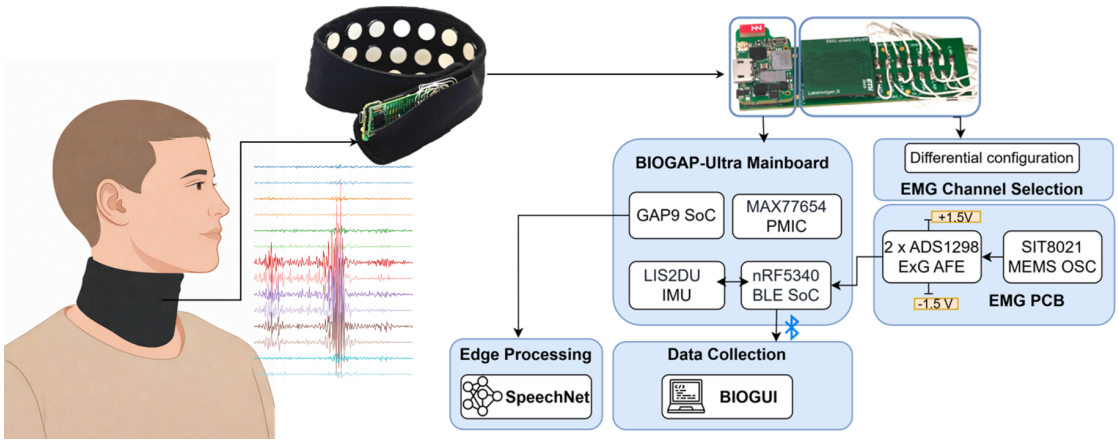


Figure 2.2: Hardware architecture of the EMG acquisition platform. Reproduced from [4]. The edge processing in the original architecture is not applied to the setup of this thesis.

The neckband consists of a soft fabric band with adjustable straps carrying fifteen fully dry Datwyler SoftPulse electrodes [46], which require no skin preparation for placement. The electrodes are arranged in three horizontal rows and wired as vertically overlapping differential pairs. This configuration provides 10 differential channels over the infrahyoid muscles and four additional channels over the SCM muscles. Four shorted electrodes at the back of the neck provide the ground reference. The band therefore provides 14 differential EMG channels in total [4, 3]. Acquisition is handled by the BioGAP-Ultra platform [16]. The biopotential board uses two Texas Instruments ADS1298 AFEs to amplify the differential channels with a PGA gain of 6 and digitize them at 500 Hz with 24-bit resolution. The mainboard integrates GAP9, an ultra-low-power SoC with a nine-core compute cluster and a neural-network accelerator, together with a Nordic nRF5340 dedicated to BLE telemetry. A reference microphone is mounted on the board to simultaneously sample audio at 16 kHz. The assembled platform runs on a 150 mAh Li-Po battery and streams data via BLE [4, 16].

2.1.2 EEG Headband

EEG signal is acquired with an adjustable textile EEG headband connected to a second BioGAP-Ultra unit, using the same acquisition hardware described in Section 2.1.1 [16]. The headband uses fully dry Datwyler SoftPulse electrodes [46]. It features 16 channels arranged equidistantly around the band and equipped with *Brush Medium* electrodes, while the bias and reference electrodes are placed on the hair-free forehead and use the flat *Dome* variant [16]. The BioGAP-Ultra digitizes the 16 EEG channels at 500 Hz with a gain of 6 and simultaneously samples a reference microphone at 16 kHz mounted on the board. The data are streamed

through a separate BLE link.

2.2 Software and Multimodal Synchronization

The open-source BioGUI application drives both hardware units, managing data acquisition, cue presentation, and labelling [47].[1] Each BioGAP-Ultra streams data through a BLE link and dongle, which the host exposes as a serial port. BioGUI opens both ports, sends the appropriate streaming commands, and decodes, displays, and saves the incoming data. The two devices operate independently and share neither a clock nor a hardware trigger. Their only common timing reference is the start of the recording issued by BioGUI, but no synchronization is maintained afterwards. A host timestamp is associated with each received packet, providing the timing information needed to align the two recordings offline. This motivates the two alignment stages described in Section 2.2.1. Labelling is handled through the *trigger* channel, which contains the numeric identifier of the item currently displayed on screen, or zero when no item is being uttered. The displayed item is referred to as a *cue* throughout this thesis. BioGUI shows one cue at a time and sends its identifier to all active recording streams. Thus, the EMG and EEG recordings share the same sequence of labels despite being acquired independently. The cue sequence and durations implement the protocol described in Section 2.3 and are reported in full in Appendix A.4. Importantly, BioGUI records one trigger value for each received packet rather than for each sample. The resulting label stream therefore contains approximately 375 values per second, compared with the 500 Hz sampling rate of the biopotential signals. Since the two time grids do not coincide, each sample must be assigned the trigger value corresponding to its timestamp before the labels can be used for further processing (Section 3.2).

2.2.1 Alignment and Delays

Reconstructing a common timeline across the multimodal recordings requires three aspects to be considered: the alignment of the biopotential and audio streams within each device, the synchronization of the two independent devices, and the intrinsic delay introduced by the acquisition chain (Table 2.1 and Figure 2.3). The first two are implemented as a standalone procedure and do not enter the decoding pipeline of this thesis.

[1] The upstream repository is available at `https://github.com/pulp-bio/biogui`, while the fork used in this work is available at `https://github.com/carolabonamico/biogui`.

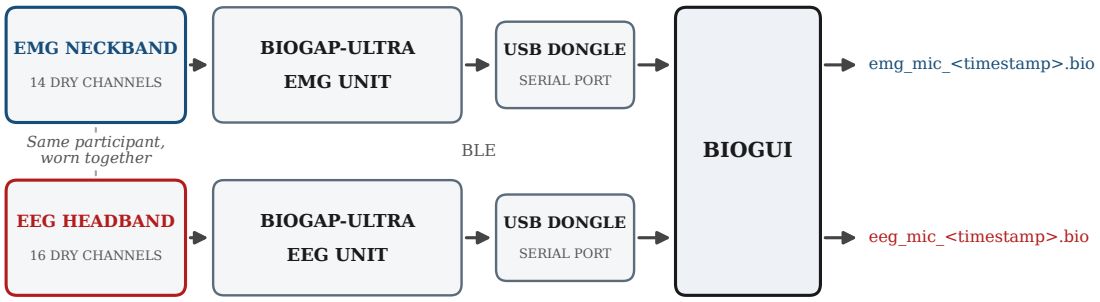


Figure 2.3: The multimodal acquisition setup. Each sensing interface is powered by a dedicated BioGAP-Ultra unit, streaming data over independent BLE links to the host.

Table 2.1: Acquisition parameters of the two identical BioGAP-Ultra units.

Parameter	EMG unit	EEG unit
Sensing interface	Neckband, 14 dry channels	Headband, 16 dry channels
Biopotential sampling rate	500 Hz	500 Hz
Microphone sampling rate	16 kHz	16 kHz
Host link	BLE to serial	BLE to serial

Within-device alignment. The first step aligns the biopotential and microphone streams belonging to the same device. Although both streams are generated by the same device, they are transmitted independently and therefore do not start at exactly the same instant. The microsecond timestamps associated with the packets are first converted to a common time base, after which all streams are restricted to their common temporal interval, from the latest initial timestamp to the earliest final timestamp. Packet counters are then used to identify transmission losses: a gap in the counter indicates a missing packet, and the corresponding samples are inserted as `NaN` values rather than being removed, preserving the original temporal structure of the recording. The resulting correction after applying the alignment is small because the two streams buffer different numbers of samples before generating a packet. At 500 Hz, a biopotential packet contains four samples, while an audio packet contains 64 samples at 16 kHz. In the vocalized sentence block recorded in the first session with participant S02, the common-window operation removes 8.5 ms at the beginning and 4.2 ms at the end. These margins occur before and after the actual recording protocol, which starts with a five-second START screen. Packet loss is similarly limited. Indeed, over the same 13.6-minute recording, one biopotential packet and two microphone packets are lost, corresponding to only four EMG samples out of 408 956 and 128 audio samples out of 13 086 656.

Between-device alignment. The second step aligns the recordings produced by the two independent BioGAP-Ultra units. Since the devices share neither a clock nor a hardware synchronization signal, the cue trigger provides the common reference. The rising edge of each cue is identified using the host timestamp associated with the corresponding packet. The k-th cue onset in one recording is then matched with the k-th onset in the other recording, and the measured offsets are linearly interpolated between consecutive cue events. Interpolation is required because the two devices use independent oscillators, so their sampling clocks can slowly diverge over time. The cue sequences were verified to be identical across the two recordings of each acquisition, providing the one-to-one correspondence required for this procedure. The two files are already closely aligned before this correction. For the sentence acquisition considered here, 200 cue onsets are available for the comparison. The measured EEG-to-EMG offset has a mean of $+0.12$ ms and a standard deviation of 2.4 ms, with no measurable drift. A linear fit to the same offsets remains below 1 ms per 1000 s. The remaining variability is due to the limitations of BLE transmission: signal recording is not istantaneous, but rather occurs only when data reaches the host. It is therefore subject to the inherent latency of Bluethoot which communicates every 7.5 ms. Consequently, the 2.4 ms value represents only the time difference between the two incoming signals, not the absolute accuracy of the devices.

Instrumental versus physiological delay. The preceding alignment phases do not identify the nature of the latency between an EMG signal stream and the corresponding microphone input. This latency must therefore be distinguished into the physiological delay between muscle activation and the resulting acoustic signal, and the delay attributable to the acquisition chain. To characterize the instrumental contribution, a microcontroller generated a 100 Hz sinusoidal burst lasting 502 ms every 1159 ms. On the one hand, the was attenuated to the millivolt range and injected into the electrodes of both interfaces, while the other drove a loudspeaker positioned close to the reference microphones. Since the same event was generated simultaneously on both paths, any measured difference between the biopotential and microphone signals was attributable to the acquisition chain. Across four test acquisitions, 4570 events were recorded. The experiment showed that the microphone signal arrives, on average, 7.7 ms after the corresponding EMG signal, with an interquartile range of 5.4–9.0 ms. This delay is of the same order of magnitude as the biopotential packet period (8 ms) and the BLE connection interval (7.5 ms). It is therefore reasonable to attribute it to packet transmission and handling, rather than to the sensors. After aligning the respective streams from the same device, the residual delay was much smaller, with a median of -2.9 ms. This value is less than half the biopotential packet period, indicating that the alignment effectively corrects the delay introduced during acquisition. The same

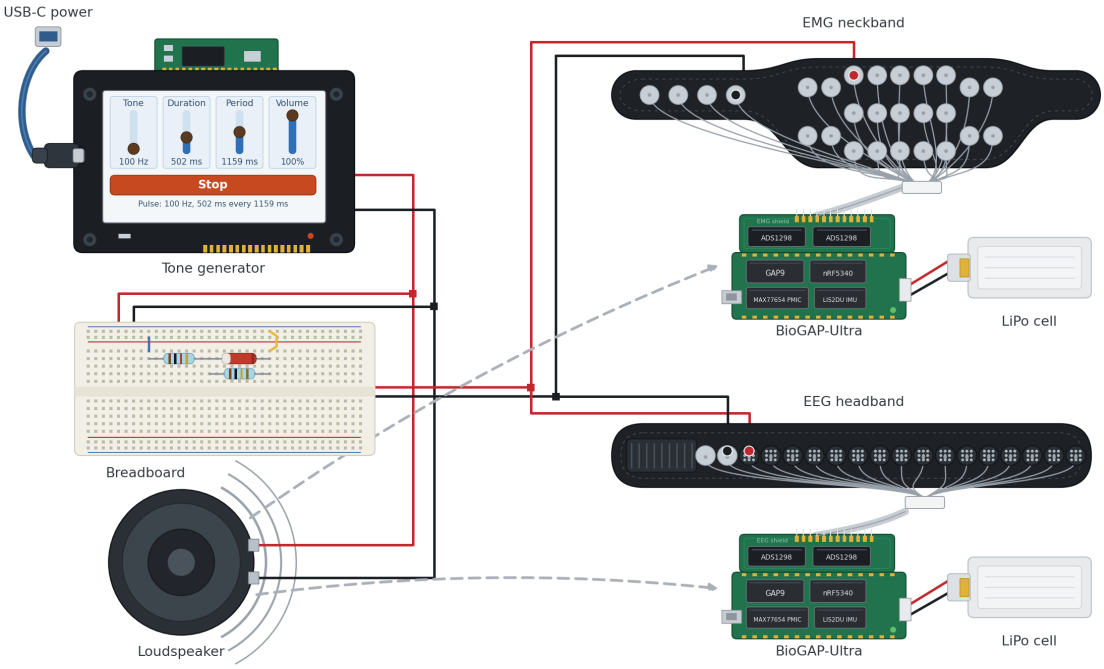


Figure 2.4: Controlled hardware setup injecting simultaneous synthetic signals to measure instrumental delay.

experiment also demonstrated that the two BioGAP-Ultra units [16] do not always have the same time offset. Offsets of +14, +6, +14, and -14 ms were measured across the four test blocks, respectively. Thus, both the magnitude and the sign of the offset can vary between acquisitions.

The controlled measurement can then be compared with the delay observed during real speech production. To quantify the latter, the same onset-detection procedure was applied to the EMG envelope and the microphone signal recorded during the data collection. The muscular-to-acoustic delay was measured on the same BioGAP-Ultra unit, so no cross-device offset was involved, and the intra-device latency had already been removed by the stream-alignment procedure described in Paragraph 2.2.1. In the vocalized sentence recording of the second session of participant S03, the delay was measured at three points of the trace, corresponding to the words *down*, *up*, and *reverse*. The measured values were 336, 394, and 343 ms, respectively, with a mean delay of 358 ms (Figure 2.5 and Table 2.2). The three measurements were taken approximately 450 s apart and remained within about 50 ms of each other, indicating that the observed difference was not caused by progressive timing drift. This delay is much larger than the instrumental delay measured in the controlled experiment and is therefore attributed to the physiological delay between muscle activation and phonation. This is consistent

Table 2.2: Quantitative readings from Figure 2.5, showing a stable physiological lead averaging 358 ms.

Instant	Word	Muscular onset [s]	Acoustic onset [s]	Δ [ms]
Beginning, $t \approx 6\,\mathrm{s}$	*Down*	6.1815	6.5175	336
Middle, $t \approx 230\,\mathrm{s}$	*Up*	230.0712	230.4653	394
End, $t \approx 456\,\mathrm{s}$	*Reverse*	456.2098	456.5527	343
Mean				358

with the anticipatory recruitment of articulatory muscles before the acoustic onset of speech [11]. Consequently, the acoustic signal cannot be used as the temporal ground truth for sEMG classification. Instead, all supervised labels in this thesis are derived from the visual cue trigger.

2.3 Data-Collection Protocol

Before designing the protocol, studies covering wearable EMG and EEG silent-speech systems, clinical and dysarthric-speech corpora, and spoken human-robot-interaction datasets were surveyed (Table 2.3). They were also compared in terms of sensing modality, corpus and application domain, number of participants and sessions, repetitions per item, recording duration, block structure, and reported performance. Three requirements were particularly relevant. First, the corpus had to contain sentence-level commands representative of a potential wearable HMI. Second, individual sessions had to remain short enough to limit fatigue. Prolonged vocal effort can alter the spectral characteristics of muscle activity [48]; since no established threshold defines an appropriate session length for speech EMG acquisition, the protocol adopts a duration of approximately one hour, consistent with the range reported by comparable studies such as Zhao et al. [49] and Zhou et al. [50]. Third, the acquisition was organized over multiple sessions, with the sensing hardware removed and repositioned between sessions, so that the resulting corpus captures the placement variability encountered in a wearable interface.

2.3.1 Vocabulary Selection

The vocabulary has two phases (Table 2.4). Phase A contains fifteen isolated command words. Eight are keywords from the baseline SilentWear system [4]; each also appears in at least one Phase B sentence, so that the same word appears both in isolation and in continuous speech. Seven additional commands (*advance*, *reverse*, *rotate*, *halt*, *begin*, *grab* and *place*) do not occur in Phase B. Phase B contains twenty sentences selected from a pool of 367 candidates. The candidates were collected from ten public sources and grouped into five domains: general robotic interaction

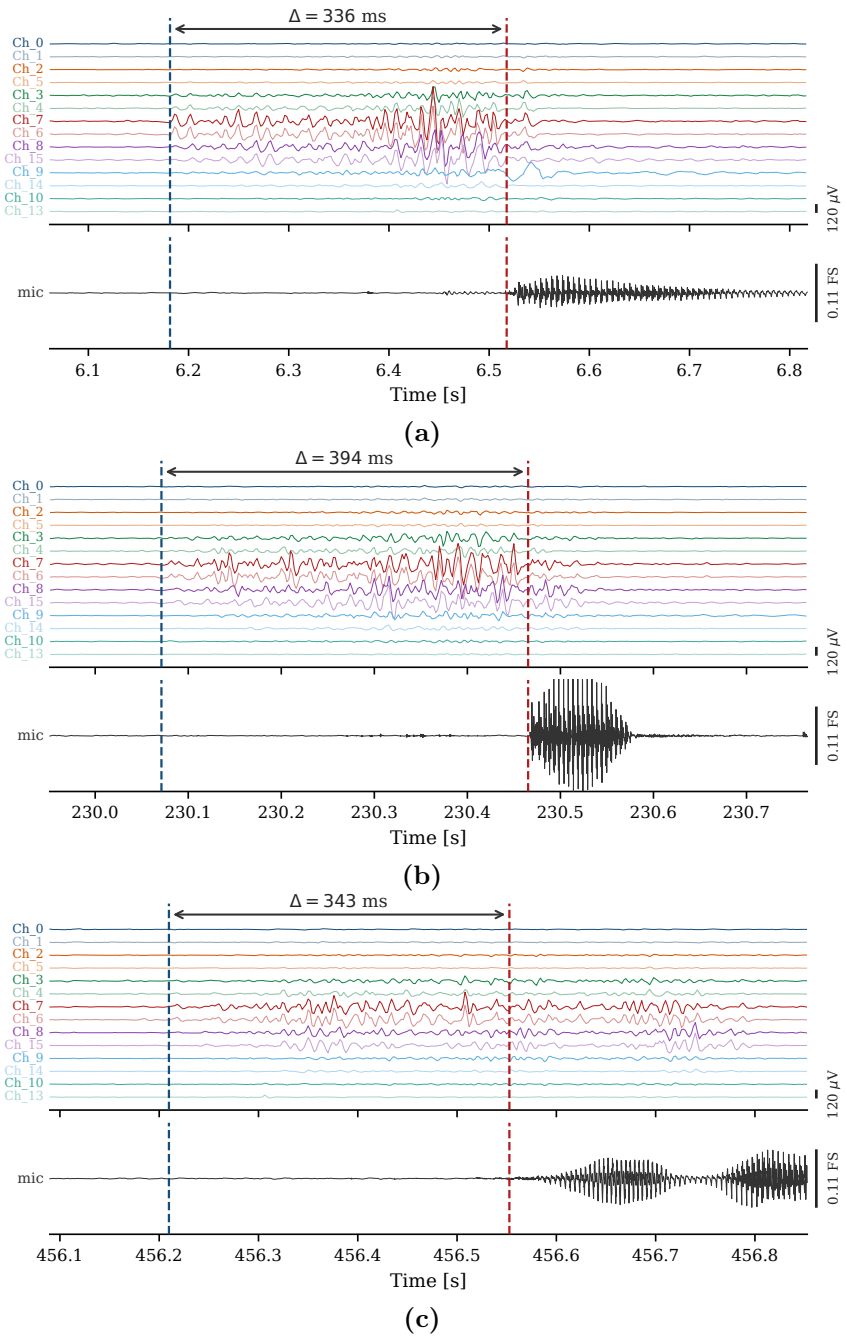


Figure 2.5: The interval between muscular and acoustic onset at three instants of one vocalized word block of participant S03, session 2: near the beginning on *down* ((a)), in the middle on *up* ((b)), and near the end on *reverse* ((c)). Dashed lines mark the two onsets, demonstrating physiological leads.

Table 2.3: Representative data-collection protocols surveyed before designing the acquisition. "Items" is the size of the closed set and "Reps" the number of repetitions of each item per subject. "Rep." states whether the sensing hardware is repositioned between sessions.

Study	Signals	Items	Domain	Subj.	Sess.	Rep.	Reps
Spacone et al. [4]	EMG	8 words	HMI	4	3	n.r.	20
Satterlee et al. [10]	EMG+EEG	10 w / 4 s	General	1	1	n.r.	150 / 100
Tang et al. [51]	Strain	47 w / 20 s	General	10	1	n.r.	100 / 50
Zhao et al. [49]	EMG+EEG	10 characters	Mandarin	30	1	n.r.	50
Zhou et al. [50]	A+V+EMG	100 sentences	Mandarin	100	10	n.r.	—
Turrisi et al. [52]	Audio	67 commands	Telephony	24	1	—	—
Andriella et al. [53]	Audio	74–79 sentences	Social robot	21	1	—	—
Shrestha et al. [54]	A+gesture	1143 commands	HMI	18	2	—	—
Lukin et al. [55]	A+visual	89,056 utts.	Robot dialogue	—	—	—	—
This work	EMG+EEG+A	15 w / 20 s	Assistive robot	7	6	Yes	60 / 60

(142 candidates), task-specific robotic interaction (110), everyday sentences (57), medical requests (40), and telephony commands (17). The principal sources are AVE Speech [50], SCOUT [55], MUTEX [56], SayTap [57], BridgeData V2 [58], and EasyCall [52]. Four criteria reduce the pool to the final closed set:

1. *Domain coherence:* only hands-free mobile and assistive-robot interaction is retained; medical and telephony candidates are discarded because they do not belong to the target application domain.

2. *Length:* sentences contain 2 to 4 words, with a mean of 3.2 words, short enough to fit comfortably within the 2.5 s utterance window at a natural speaking rate and consistent with the distribution observed in AVE Speech [50].

3. *Keyword coverage:* sentences containing one of the eight legacy command words are preferred, so identical lexical units can be compared between isolated and continuous speech.

4. *Lexical overlap:* a restricted set of content words is reused in different positions

Table 2.4: Parameters of the acquisition protocol.

Parameter	Value
Vocabulary	15 words, 20 sentences
Repetitions per item and block	10
Repetitions per item, participant	60 vocalized, 60 silent
Speaking conditions	Vocalized and silent, in separate blocks
Sessions per participant	6, on separate days, devices repositioned
Participants	7
Utterances per session	700

and syntactic roles, preventing the sentence classes from being reduced to the detection of a single discriminative word.

The twenty sentences comprise eleven navigation commands, five object and assistive-interaction commands, and four control commands. Twelve are taken or adapted from the analyzed sources and eight were written specifically for this study. The complete sentence set is reported in Appendix A.1.2.

2.3.2 Session Structure and Participants

Seven healthy volunteers (two female and five male) participated in the study and written informed consent was obtained from them before the data collection. The study was conducted under ethics approval EK-2025-E-13 and in accordance with the Declaration of Helsinki. Each participant completed 6 sessions recorded at least 24 hours apart, with the EMG neckband and EEG headband repositioned between sessions. Each session composes of 4 recorded blocks: one vocalized and one silent block of the fifteen words, followed by one vocalized and one silent block of the 20 sentences. Within each trial, the participant receives a 1.5 s utterance window for a word or a 2.5 s window for a sentence, followed by a fixed 1.5 s rest interval shown as a visual countdown in BioGUI [47]. The rest interval defines the *rest* class and is not collected in dedicated trials; rest samples are drawn from the intervals between consecutive items. A 5 minute seated rest separates the two blocks of each phase, while a 5 minute interval separates Phase A from Phase B. During the breaks the acquisition setup remains unchanged. The operator introduces the participant to the protocol (Appendix A) before the session and provides no further directives during the recorded runs. Each item is repeated ten times per block, giving 150 word utterances and 200 sentence utterances per condition and 700 recorded utterances per session. Then, a new pseudo-random order is generated independently for every block, session, and participant. In total, the dataset yields 42 sessions, 168 blocks, and 29 400 utterances, evenly split between vocalized and silent speech. While this thesis only processes the EMG data, the parallel EEG and audio streams are preserved for future multimodal fusion.

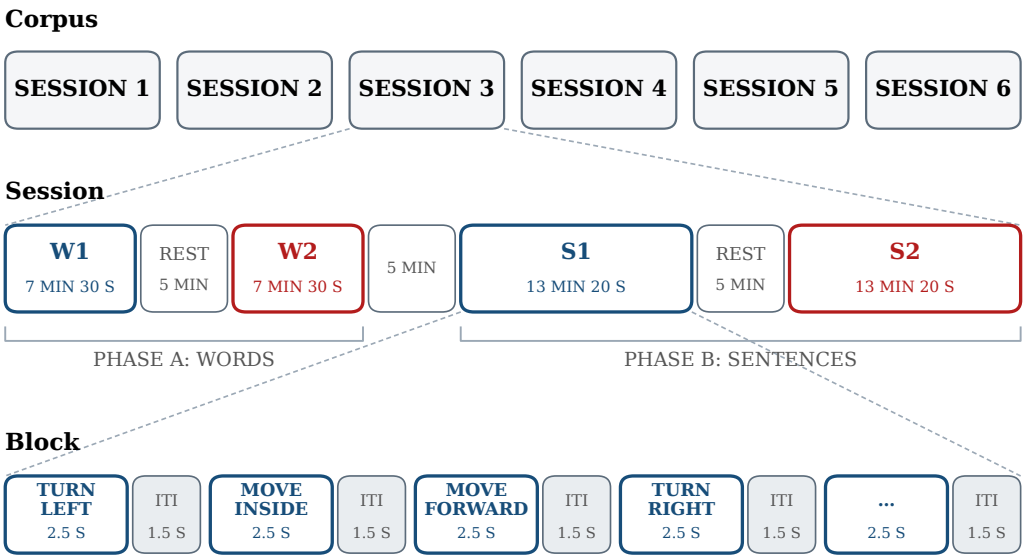


Figure 2.6: Structure of the data collection. Each session alternates vocalized (blue) and silent (red) blocks separated by rests.

Chapter 3

Closed-Set Classification: Methods

This thesis addresses two related problems: closed-set classification and continuous recognition. From this point onward, all analyses use only the EMG data. The two tasks share the same signal processing and modelling framework. This chapter therefore introduces the closed-set classification problem and then describes the common preprocessing pipeline, input representations, model architectures, and training objectives.

3.1 Problem Formulation

The classifier takes as input a window of 14 filtered EMG channels and assigns it to one label from a fixed set. The two classification tasks differs only in the label set: at the word level, the classifier distinguishes 16 classes, corresponding to the 15 command words and rest, whereas at the sentence level it distinguishes 21 classes, corresponding to the 20 sentences and rest. In both cases, rest is a regular class rather than a rejection option, so every window is assigned a label. These label sets are defined by the acquisition protocol described in Section 2.3.2. Of the two training objectives in Section 3.4, one directly outputs a predefined label, while the other outputs a character string that is then mapped back to the label set. Character-level errors therefore affect the classification result only when they change the assigned class. Chapter 5 removes this final mapping and evaluates the decoded string directly.

3.2 Signal Processing Pipeline

The processing pipeline converts the raw `.bio` recordings into labelled, fixed-length windows of filtered EMG. The trigger generated by BioGUI [47] provides the labels, and all data streams are clamped to ensure that signals and labels are aligned and end simultaneously [4].

Filtering. The EMG signals are filtered with a fourth-order Butterworth HPF at 20 Hz and a 50 Hz notch filter with quality factor 30 to reduce power-line noise [4]. The applied forward-backward filtering is zero-phase, so it does not introduce a temporal shift and preserves the alignment between the signal and the labels. Its magnitude response is, however, the square of the single-pass response [59, 60].

Cropping and channel selection. After filtering, the recordings are cropped to the labelled interval with an additional one-second margin, corresponding to 500 samples at each end [4]. This removes the idle portions before the first cue and after the last one. The acquisition provides 16 digitized inputs, of which 14 correspond to the EMG channels used in this thesis. The two unused inputs, `Ch_11` and `Ch_12`, are removed, and the remaining channels are reordered according to the fixed electrode ordering of the neckband [4].

3.2.1 Segmentation and Windowing

Windowing. As illustrated in Figure 3.1, the filtered EMG signal is segmented into fixed-length windows [4] but only segments which do not exceed the acquisition boundaries are retained. Following the acquisition protocol of Section 2.3.2, the window duration is 1.4 s for isolated words and 2.0 s for full sentences.

Sliding-window augmentation. To increase training variability, a sliding-window augmentation procedure [28] can be applied to extract shifted copies of each segment at

$$n_0 + j\Delta, \qquad j \in \{-N, \ldots, -1, 0, 1, \ldots, N\}, \tag{3.1}$$

where:

- n_0 is the segment onset;
- Δ is the temporal stride;
- N is the number of shifts on each side;
- $j = 0$ corresponds to the original window.

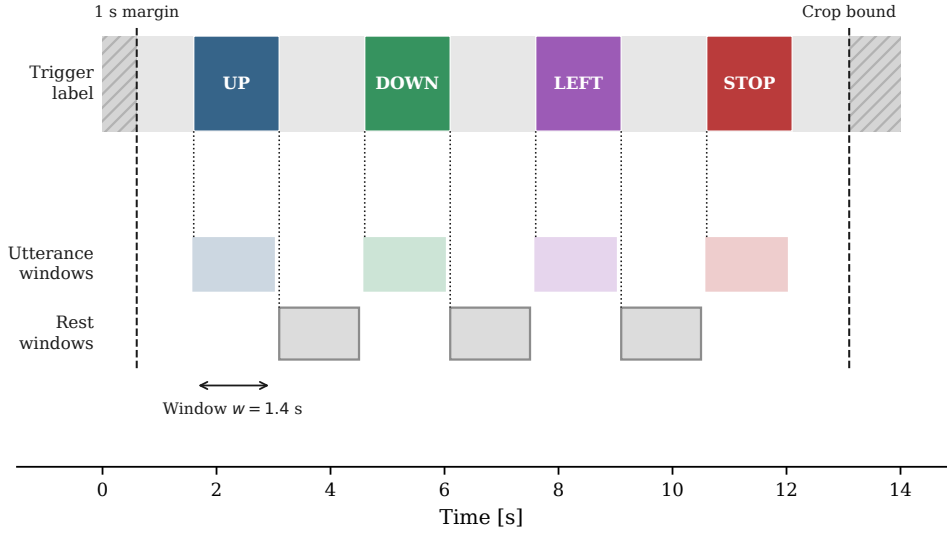


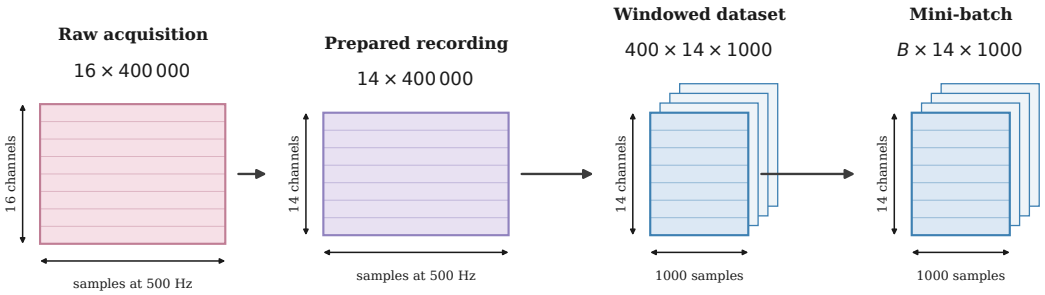
Figure 3.1: Cropping, segmentation and windowing. The trigger track alternates rest and utterances; shaded areas are the idle margins discarded during cropping. One fixed-length window is extracted from each segment and anchored at its onset.

Two augmentation modalities are possible. Either all shifted windows are added to the training set, increasing its size, or the augmented pool is sampled so that the training set keeps the size of the original one. The two modes are treated as separate ablations and are compared with the non-augmented baseline. Augmentation is disabled in the main pipeline of this thesis and evaluated separately.

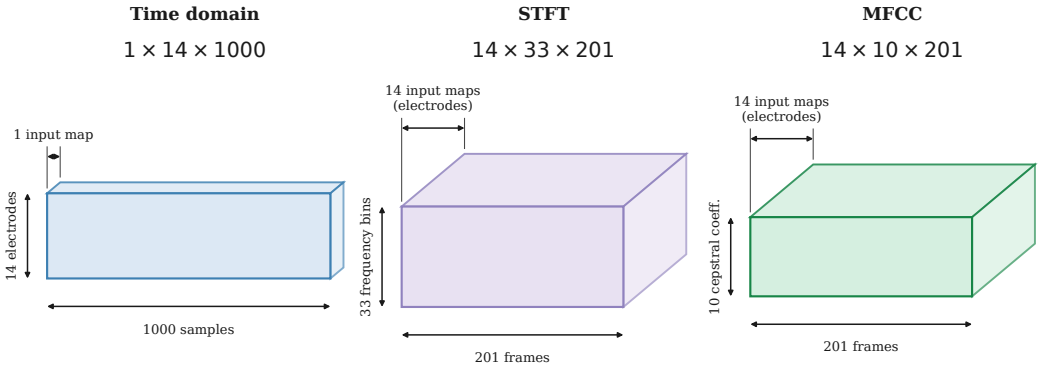
3.3 Feature Extraction and Representations

After preprocessing (Section 3.2), each window is represented by a matrix with 14 channels and T samples, where $T = 1000$ for 2.0 s sentence windows and $T = 700$ for 1.4 s word windows. Three input representations are compared: the raw time signal, the short-time Fourier transform (STFT), and the mel-frequency cepstral representation. Their performance is compared in Section 4.4, and the selected representation is then passed to the models of Section 3.5.

Time domain. In the time domain representation, each window consists of a 14×1000 tensor, with one row per electrode and one column per sample. This preserves the original 500 Hz signal temporal resolution and has the smallest data volume of the three options, with 14,000 values per window.



(a) Pipeline from raw recording to mini-batch.



(b) Tensor dimensions at the model input.

Figure 3.2: From the raw recording to the tensor entering the network. Panel (a) shows the preprocessing from the raw `emg` stream to the mini-batch. Panel (b) shows the dimensions of the time-domain, STFT, and mel representations at the model input. The axis order is (batch, map, rows, columns), with the batch axis omitted.

Short-time Fourier transform. The STFT produces a power spectrogram for each channel following Allen [61, 59]. For channel c,

$$X_c[m,k] = \sum_{n=0}^{N-1} w[n]\, x_c[mH+n]\, e^{-j2\pi kn/N}, \qquad S_c[m,k] = 10\log_{10}\left|X_c[m,k]\right|^2, \tag{3.2}$$

where:

- $x_c[\cdot]$ is the filtered signal;
- $w[\cdot]$ is the window;

- m is the time index;
- k is the frequency index.

The final configuration uses $N = 64$ samples, corresponding to 128 ms, and provides a frequency resolution of 7.8 Hz. This gives 33 frequency bins up to the 250 Hz Nyquist limit. The hop size is $H = 5$ samples, or 10 ms, producing 201 time frames for a 2.0 s sentence window and 141 for a 1.4 s word window. The resulting sentence tensor is therefore $14 \times 33 \times 201$, or 92,862 values per window. A longer analysis window improves frequency resolution but reduces temporal precision, while the hop size controls the density of the output frames. As discussed in Section 4.4.1, the window length was also swept from 32 to 512 ms.

Mel-frequency cepstral coefficients. The mel cepstrum follows the definition of Davis and Mermelstein [62] and is widely used in EMG speech studies [30, 50]. It first maps the spectrogram onto a set of triangular filters spaced on the mel scale and then applies a DCT [63]. The q-th coefficient is

$$c_c[m, q] = \sum_{b=1}^{B} \log \left(\sum_k H_b[k] \; |X_c[m, k]|^2 \right) \cos \left(\frac{\pi q (b - 1/2)}{B} \right), \qquad q = 0, \ldots, Q - 1, \tag{3.3}$$

where:

- $H_b[\cdot]$ is the b-th triangular filter;
- B is the number of filters;
- Q is the number of retained cepstral coefficients.

The transform uses a 128-sample analysis window and the same 10 ms hop as the STFT, giving 65 frequency bins before mel filtering. The filters are equally spaced on the mel scale [64], in the closed form given by O'Shaughnessy [65],

$$m(f) = 2595 \log_{10} \left(1 + \frac{f}{700} \right)$$

Because the EMG bandwidth ends at 250 Hz and the mel scale is nearly linear below 1 kHz, this operation does not introduce perceptual warping in the present setting. The filter bank is fitted to the EMG bandwidth, with $B = 15$ and $Q = 10$ [20]. This yields a $14 \times 10 \times 201$ tensor, corresponding to 28 140 values per window.

Spectral augmentation. SpecAugment [66] is applied during training to the STFT and mel representations. Specifically, one frequency mask (up to 8 bins wide) and one temporal mask (up to 20 frames wide) are generated per window at random offsets, setting the corresponding entries to zero. The temporal mask is applied both to MFCC and to the STFT, while the frequency mask is applied only to the latter.

3.4 Training Objectives

The two tasks use the same input windows and network up to the output head. They differ in the prediction target, the width of the head, and how the output is decoded. This section describes the two objectives and the width of the head. Decoding operates on the outputs of a trained network and is described in Section 3.6.

3.4.1 Cross-entropy and CTC

The two objectives differ in the level at which the prediction is made. With CE, the temporal features are first pooled into a single representation, and the network outputs one distribution over the predefined classes (Table 3.1). With CTC, the temporal sequence is preserved and each output frame produces a character-level distribution. These frame-level predictions are then decoded into a text string, allowing the model to learn character-level patterns that can be shared across sentences with overlapping vocabulary, such as *turn left*, *turn right*, and *turn on the light*. For closed-set classification, the decoded string is subsequently mapped to one of the predefined classes.

Formulation of the CTC loss. CTC [37] trains a frame-level classifier when the target sequence is shorter than the frame sequence and no frame-to-character alignment is provided. A blank symbol is added to the token set. The collapse operation merges consecutive repetitions and removes blanks. Every frame-level path that collapses to the target is therefore a valid alignment. The loss is

$$\mathcal{L}_{\mathrm{CTC}}(\mathbf{y}) = -\log \sum_{\boldsymbol{\pi} \in \mathcal{B}^{-1}(\mathbf{y})} \prod_{t=1}^{T'} p_t\big(\pi_t\big), \tag{3.4}$$

where:

- $\mathbf{y}$ is the target character sequence;
- $\mathcal{B}$ is the collapse operator and $\mathcal{B}^{-1}(\mathbf{y})$ is the set of frame-length paths that collapse to $\mathbf{y}$;

- $p_t(\cdot)$ is the frame-level distribution over the token set plus the blank;
- T' is the number of output frames.

The sum is computed with a forward dynamic program because explicitly enumerating all valid paths would be exponential in T'. The model can therefore place each character at the frames where the evidence occurs, provided that the character order is preserved. The alignments are monotonic given that frames cannot be reordered or reused. Since every input window has the same duration, all sequences contain the same number of output frames and no padding mask is required. CTC also requires sufficient output frames to represent the target sequence. A target with L characters needs at least L frames, plus one additional frame for every pair of adjacent identical characters that must be separated by a blank. The longest sentence target contains 24 characters, so this condition sets a lower bound on the temporal resolution the network must preserve.

Text normalization and tokenization. Each class has an integer label and a corresponding text string stored in the lexicon (Appendix B). Ground truths and predictions undergo the same normalization before comparison, including transliteration of accented characters, conversion of hyphens and colons into spaces, removal of punctuation, lowercasing and trimming. A target is therefore obtained by taking the text of the class and mapping its characters, spaces included to preserve word boundaries.

Width of the output layer. The output dimension is determined by the training objective. With CE, the output units correspond to the target classes, resulting in 21 outputs for sentences and 16 for words, including the rest class. With CTC, instead, the output units correspond to the distinct characters in the task lexicon, plus the blank symbol. The sentence lexicon contains 23 symbols, yielding 24 outputs, while the word lexicon contains 20 symbols, yielding 21 outputs. The head is a linear layer, so its size is V times the number of features it reads, plus V biases. The difference between the two CTC heads is therefore about one thousand parameters. Continuous recognition instead uses a fixed alphabet, independent of the lexicon, defined in Chapter 5.

3.5 Deep Learning Models

Each model in this thesis consists of the input representation described in Section 3.3, a convolutional backbone, and a sequential stage that processes the backbone's frame-level outputs. The final predictions are produced by a linear head with dimensions outlined in Table 3.1. An architecture is therefore defined by three

Table 3.1: Width V of the output layer. Under CE, the head enumerates classes; under CTC, it enumerates character tokens. The last row is the fixed alphabet of the continuous recognition task, defined in Chapter 5.

Objective and task	**What the head enumerates**	V
CE, sentences	20 sentence classes and rest	21
CE, words	15 word classes and rest	16
CTC classification, sentences	23 lexicon characters + blank	24
CTC classification, words	20 lexicon characters + blank	21
CTC recognition	26 letters, space + blank	28

choices: the input domain, time, STFT or MFCC; the sequence model, none, BiLSTM or Transformer encoder; and the training objective, CE or CTC. The starting point is the published baseline, which combines the time domain, no sequence model and CE. Components are then changed one at a time, so a difference in accuracy attaches to a single modification. Figure 3.3 summarizes the three architectures compared.

3.5.1 Reference Model

The baseline model is the SpeechNet classifier [4] composed of five convolutional blocks (Table 3.2a). The first three blocks learn temporal patterns independently on each electrode and reduce the time dimension by a factor of 128. The last two mix electrodes, after which adaptive average pooling produces one value per feature map and a linear layer predicts the nine classes of the original corpus. The model contains 15 489 parameters and uses 1.4 s windows. It is reimplemented in the framework of this thesis as a fixed reference. Section 4.2 first compares the reimplementation with the published system on the four participants of the original SilentWear corpus[1] and then evaluates it on the corresponding word subset of the present corpus. Two properties of the baseline are incompatible with CTC. It leaves five output frames for a 1.4 s word window and eight for a 2.0 s sentence window, which violates the condition on the frame count of Section 3.4. Its head also reads one pooled vector per window, while a CTC head reads each frame.

3.5.2 Convolutional Backbone

The backbone follows the block structure of the baseline, while modifying the pooling, layout, and channel width to match the input representation and the requirements of frame-level decoding. The resulting stacks are summarized in

[1] `https://huggingface.co/datasets/PulpBio/SilentWear`

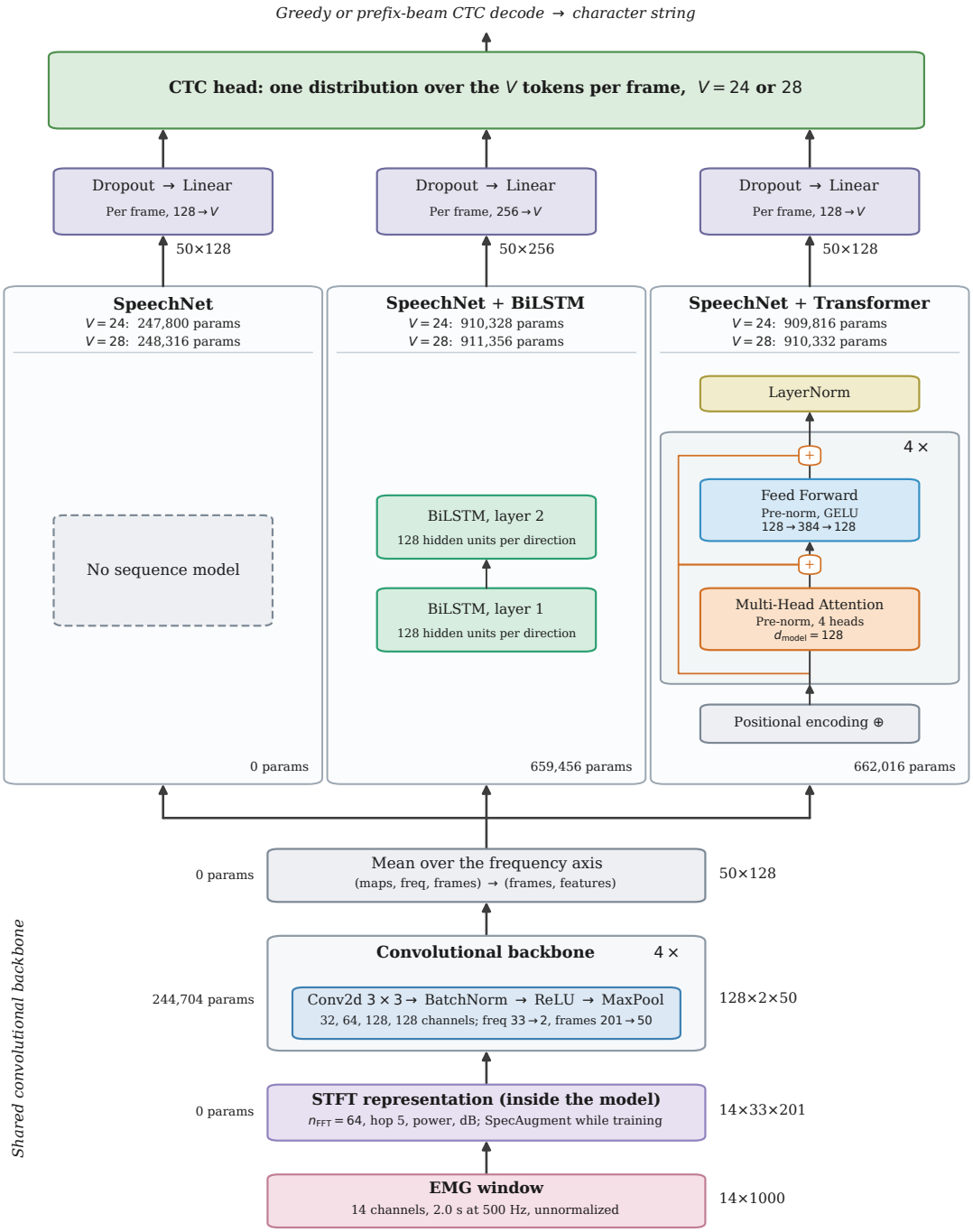


Figure 3.3: The three architectures compared in this thesis. The convolutional backbone is shared, while the sequence stage changes. The head width is the only component that varies with the task, with $V = 24$ for closed-set classification and $V = 28$ for continuous recognition (Table 3.1).

Table 3.2. In the time domain, the first three blocks apply temporal filtering independently to the 14 electrodes and downsample the window by a factor of two at each stage, leaving $T/8 = 125$ time steps. The fourth block then spans the full electrode axis, collapsing the spatial dimension, and the final 1×1 block produces 128 features per frame. In the spectral domains, the 14 electrodes are treated as input maps and are fused from the first convolution; the 3×3 kernels therefore jointly process the frequency and time dimensions. Pooling reduces the frequency axis through all four blocks but stops along time after the second block, leaving 50 time steps. The wider filters are motivated by the frame-level output required by CTC. Each frame must preserve enough information to support character prediction, which motivates the larger channel widths used in both backbones. These widths were fixed before the experiments and were not ablated. The resulting backbone contains 79 328 parameters in the time domain and 244 704 in the spectral domains (Table 3.2); these counts refer to the convolutional backbone alone, while the complete model additionally includes the input representation, sequence stage, and output head. After the convolutional blocks, each backbone produces a three-dimensional feature map with 128 channels, a residual row axis, and a temporal axis. The row axis contains frequency bins for the spectral representations, whereas the time-domain stack has a single row because the fourth block has already collapsed the 14-electrode axis (Table 3.2b). A mean over this row dimension, equivalent to global average pooling restricted to that axis [67], converts the map into a sequence of 128-dimensional feature vectors, one per frame. In particular, the 33 rows of the STFT representation and the 10 rows of the mel-based representation are all reduced to the same 128 features per frame, as is the single row of the time-domain stack. Consequently, the same sequence stage can be used across all three input domains without changing its parameters, keeping the downstream model independent of the input representation.

3.5.3 Bidirectional LSTM

The first sequence model is a two-layer bidirectional LSTM [68, 69, 70] with $H = 128$ hidden units per direction and dropout [71] between the two layers. The choice behind its introduction is motivated by the frame-level nature of CTC, which was originally introduced with bidirectional recurrent networks [37, 70]. The gated structure of the LSTM also helps maintain stable gradients over the sequence [68]. Each layer processes the complete frame sequence in both temporal directions and concatenates the two hidden states [69], producing 256 features per frame. This recurrent stage complements the local temporal context captured by the convolutional backbone: while an output frame of the backbone only depends on a limited portion of the input window, the recurrent pass can propagate information from the entire stored window to every frame. This is useful when distinguishing

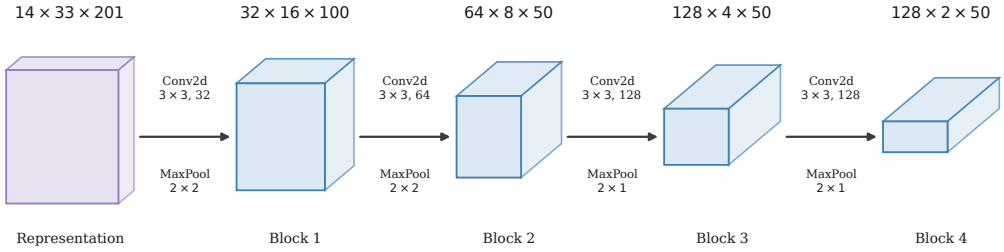
Table 3.2: The baseline stack and the two convolutional backbones of this thesis. Panel (a) is the baseline at its 1.4 s word window [4], panels (b) and (c) are the modified stacks at the 2.0 s sentence window. Each block follows the `Conv2d-BatchNorm2d-ReLU-MaxPool2d` pattern. Shapes are given as (channels × rows × columns), where rows represent the electrodes (for the time domain) or frequency bins (for spectral domains), and columns represent the time steps; the batch axis are omitted.

Block	Kernel	Pool	Output shape	Params
(a) Baseline SpeechNet [4]; input $1 \times 14 \times 700$				
1	1×4, 8	1×8	$8 \times 14 \times 87$	56
2	1×16, 16	1×4	$16 \times 14 \times 22$	2 096
3	1×8, 16	1×4	$16 \times 14 \times 5$	2 096
4	7×1, 32	1×1	$32 \times 14 \times 5$	3 680
5	7×1, 32	1×1	$32 \times 14 \times 5$	7 264
			Total	15 192
(b) Time domain; input $1 \times 14 \times 1000$				
1	1×5, 16	1×2	$16 \times 14 \times 500$	128
2	1×5, 32	1×2	$32 \times 14 \times 250$	2 656
3	1×5, 64	1×2	$64 \times 14 \times 125$	10 432
4	14×1, 64	1×1	$64 \times 1 \times 125$	57 536
5	1×1, 128	1×1	$128 \times 1 \times 125$	8 576
			Total	79 328
(c) Spectral domains; input $14 \times 33 \times 201$				
1	3×3, 32	2×2	$32 \times 16 \times 100$	4 128
2	3×3, 64	2×2	$64 \times 8 \times 50$	18 624
3	3×3, 128	2×1	$128 \times 4 \times 50$	74 112
4	3×3, 128	2×1	$128 \times 2 \times 50$	147 840
			Total	244 704

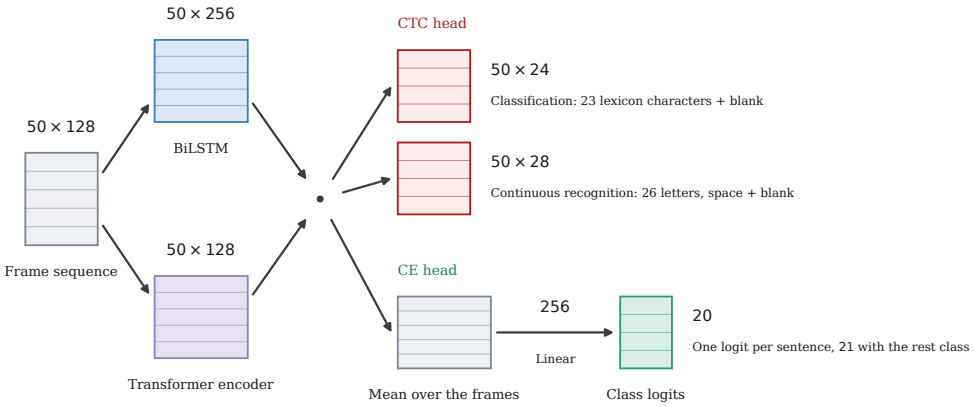
similar utterances depends on evidence appearing later in the sequence, such as the difference between *turn on the light* and *turn off the volume*.

3.5.4 Transformer Encoder

The second sequence model replaces the BiLSTM with a Transformer encoder [72] and keeps the rest of the architecture unchanged. The encoder takes the same 128-dimensional feature vectors produced by the convolutional backbone, so its model dimension is set to $d_{\text{model}} = 128$ and no additional input projection is required. The encoder uses four layers, four attention heads, and a feed-forward width of 384, chosen to match the BiLSTM parameter budget. It contains 661 760 parameters versus 659 456 for the BiLSTM, restricting the parameter discrepancy to 0.3% for this stage and approximately 0.1% for the complete model. This capacity match



(a) Tensor shape through the convolutional backbone.



(b) Tensor shape through the reduction, sequence stage, and output head.

Figure 3.4: Tensor dimensions inside the network for the STFT representation and the 2.0 s window. All shapes were measured with forward hooks on the instantiated models.

ensures that any performance variation isolates the core model architecture.

Sequence representation and tokenization. The Transformer processes the frame sequences emitted by the convolutional stage. For a 2.0 s window, this yields exactly fifty tokens spaced by 40 ms, with each token summarizing an overlapping 378 ms receptive field. Since the preceding convolutions fuse the 14 electrodes and pool the frequency axis, no spatial or channel identity reaches the encoder. As a consequence, the multi-head self-attention computes temporal dependencies. Deriving these overlapping tokens from a convolutional stage aligns the architecture with acoustic models like the Speech Transformer [73] and Conformer [74].

Optimization corrections at equal number of parameters. Each of the four layers adopts a pre-norm residual block structure [75], placing the layer normalization before the multi-head attention and the feed-forward networks. The attention mechanism uses the scaled dot-product form over $A = 4$ heads [72],

Table 3.3: The three sequence stages and the size of the complete model they produce. Every count is for the spectral backbone of Table 3.2c, whose 244 704 convolutional parameters are included in the totals. The four right-hand columns give the total for a linear head of each width of Table 3.1. The Transformer count includes the `LayerNorm` on the features entering the encoder; the encoder itself holds 661 760 parameters. All counts are read from the instantiated models.

			Complete model			
Sequence stage	**Feat./frame**	**Stage**	$\boldsymbol{V} = 16$	$\boldsymbol{V} = 21$	$\boldsymbol{V} = 24$	$\boldsymbol{V} = 28$
None	128	0	246 768	247 413	247 800	248 316
BiLSTM	256	659 456	908 272	909 557	910 328	911 356
Transformer encoder	128	662 016	908 784	909 429	909 816	910 332

allowing the query-key product to relate any two frames in a single operation, unlike the sequential state propagation of recurrence. The feed-forward block integrates a GELU non-linearity [76] and replaces the ReLU of the convolutional stage, with an inner width of $d_{\text{ff}} = 384$. Optimization is further tuned via Xavier initialization [77] for all projections and an intra-encoder dropout of 0.1, which is decoupled from the 0.5 dropout of the convolutional path. The behavior driving these choices is analyzed in Section 4.5. Finally, the architecture adds two external `LayerNorm` layers [78]. The first normalizes the features entering the encoder, as the global mean over frequency leaves the per-frame scale unnormalized after the final `BatchNorm`, to which the Transformer is sensitive. The second normalizes the final encoder output.

Positional encoding and computational cost. Temporal order is injected via the sinusoidal positional encoding [72]. Because this encoding is computed dynamically rather than learned, it consumes no parameter budget and adapts automatically to any sequence length: 50 frames in the spectral domains, 125 in the time domain, and 35 at the word window. Consequently, a single encoder supports all feature configurations without enforcing a fixed maximum context length. On the other hand, equal parameter counts do not automatically guarantee equal computational cost. Measuring a single 2.0 s window reveals that a forward pass requires 231 million FLOPs for the convolutional stage, rising to 296 million with the Transformer encoder. Although the recurrent kernel cannot be directly measured, estimating two operations per parameter per frame places the BiLSTM sequence stage at the same 65-million FLOP order of magnitude as the Transformer. Therefore, the two sequence models remain strictly comparable in both memory footprint and computational cost.

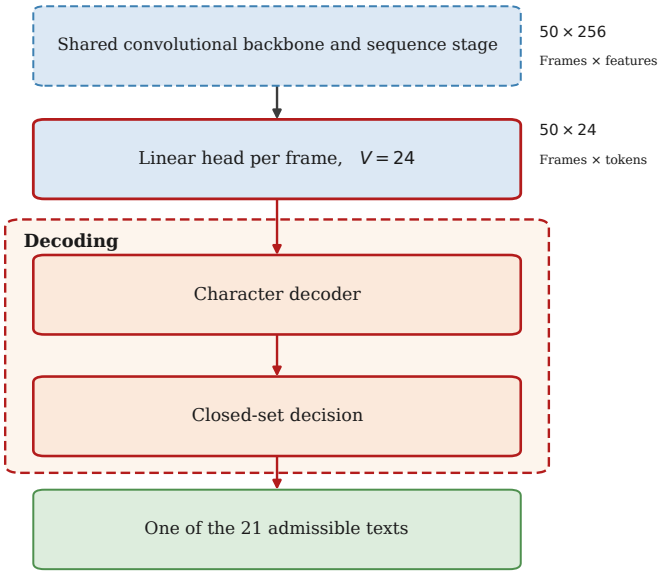


Figure 3.5: The classification pipeline of this chapter, drawn for the STFT representation, the BiLSTM sequence stage and the CTC objective. The stages before the head are those of Figure 3.3 and are named.

3.6 Decoding

Under CTC the network emits one distribution per frame, and a decoder turns the frame sequence into a string. For closed-set classification a second step maps that string to one of the predefined classes. Figure 3.5 summarizes the process from the frame-level predictions to the final class decision.

Greedy and prefix beam search. Two decoders operate on the frame-level log-probabilities before the CTC collapse. Greedy decoding selects the most probable token at each frame and applies the collapse [37]. Prefix beam search [79, 80] instead keeps multiple candidate prefixes and sums the probabilities of all alignments producing the same prefix. This allows a sequence to be selected when no single alignment path is individually optimal. Neither decoder developed for this thesis uses a language model. Three parameters control the beam search. A temperature T_θ divides the logits before the softmax [81, 82], flattening the distributions to increase the weight of alternative alignments. A blank penalty β subtracts a constant from the blank log-probability to reduce the CTC tendency to overproduce blanks. A length bonus λ compensates for the bias towards short strings. At default values ($T_\theta = 1$, $\beta = 0$, $\lambda = 0$), the mechanism reduces to plain prefix search. For closed-set classification, the decoded string resolves to the lexicon entry with the minimum character edit distance [83]. If the decoder produces an empty string, the system scores every candidate text using the CTC forward

algorithm [37] and selects the most probable one.

Evaluation of the decoder. Decoding parameters do not influence training. Each model is therefore trained once, and its per-frame log-probabilities for all test folds are saved to files. The decoder parameter grid is then evaluated offline on these fixed outputs, which separates the effect of the search strategy from the effect of the model.

3.7 Evaluation Settings and Metrics

All models are subject-specific, meaning a dedicated model is trained for each participant under every speaking condition. The overall evaluation framework follows the SilentWear architecture [4] while adapting the dataset splits to the six-session acquisition described in Section 2.3.2.

3.7.1 Cross-Validation Protocols

To measure distinct generalization capabilities, the evaluation implements two cross-validation protocols that differ in how the training and testing folds are partitioned. The *global* protocol measures model performance when training and test data share the same sessions. It implements a five-fold stratified cross-validation [84] over the corpus of each participant, applying a constraint to preserve the distribution of session and class labels across all folds. The *inter-session* protocol evaluates model robustness against electrode repositioning implementing a leave-one-session-out cross-validation across the six recorded sessions. Each fold trains on five sessions and holds out the remaining session for testing. Because the test session was acquired on a different day with a new physical device placement, this protocol enforces generalization across varying skin-electrode impedances and placement shifts [4]. Eventually, all cross-validation splits are built from the unshifted base windows, and to prevent data leakage, augmented copies (Section 3.2.1) are only injected into the training splits. Within each training fold, a stratified random split reserves 20% of the data for validation to guide early stopping and learning rate scheduling [4]. In addition, random seeds are reset at the beginning of each fold to guarantee reproducibility.

3.7.2 Training Schedule and Class Balancing

Each evaluation fold trains a model from scratch in mini-batches of 32 windows and the optimization pipeline is specifically tuned to stabilize the sequence-modelling stages. The models are optimized using AdamW [85] with a base learning rate of 10^{-3} and a weight decay of 10^{-4}. To prevent gradient explosion in the recurrent

and attention layers, gradient norms are clipped at 1.0 [86]. The training sequence begins with a 10-epoch linear warm-up phase, scaling the learning rate from zero to its base value [87]. Following warm-up, a plateau scheduler monitors the validation loss, dividing the learning rate by ten after 12 stable epochs. Training runs for a maximum of 150 epochs and stops early after 30 consecutive epochs without validation loss improvement. The checkpoint that achieves the lowest validation loss is restored for testing. This training schedule extends the baseline configuration according to sentence-level sequences requirements but the baseline reproduction (Section 3.5.1) keeps the standard Adam optimizer [88] at a fixed learning rate of 10^{-3}, a plateau scheduler with a patience of 2 epochs, a maximum of 100 epochs, and early stopping after 10 epochs, omitting both the warm-up phase and gradient clipping. In addition, to prevent the rest class from dominating the loss during training, its windows are downsampled to match the size of the smallest active utterance class [4]. This downsampling applies only to the training and validation splits while the test folds keep their distribution where the rest state accounts for half of the incoming stream.

3.7.3 Evaluation Metrics and Representational Probing

The evaluation monitors both the final classification accuracy and the underlying quality of the learned latent spaces.

Classification metrics. Standard accuracy calculates the proportion of correct predictions over the entire evaluation set

$$\text{Accuracy} \;=\; \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\hat{y}_i = y_i), \tag{3.5}$$

where the components are:

- N: the total number of predictions;
- $\hat{y}_i$: the predicted class for window i;
- y_i: the true class for window i;
- $\mathbf{1}(\cdot)$: the indicator function.

Because the test sets maintain the original and imbalanced class distribution due to the un-downsampled rest windows, standard accuracy is an invalid metric. The primary evaluation metric is therefore the balanced accuracy [89], calculated as the

unweighted mean of the per-class recalls to ensure that every command and the rest state carry equal statistical weight

$$\text{Balanced Accuracy} = \frac{1}{|C|} \sum_{c \in C} \frac{TP_c}{TP_c + FN_c}, \tag{3.6}$$

where:

- C: the set of all classes;
- TP_c: the number of true positives for class c;
- FN_c: the number of false negatives for class c.

Macro- and weighted-averaged precision, recall, F_1-score [90], and row-normalized confusion matrices are computed alongside the primary metric [4]. All evaluation metrics are first averaged across the validation folds for each subject and then averaged across the seven participants.

Nearest-centroid representation. To evaluate the feature structures formed by the sequence stage independently of the final linear classification head, a non-parametric analysis is conducted on the fully trained models. Activation vectors are extracted at two specific boundaries: the output of the convolutional stage and the input to the final linear layer. A nearest-centroid classifier evaluates these latent spaces. First, an explicit geometric centroid is computed for each class by averaging the activation vectors across the training and validation windows. Second, every test window is assigned to its closest class centroid based on Euclidean distance. The resulting assignments are scored using the balanced accuracy metric. This approach also acts as a measurement of the inherent discriminative power of the learned representations. The numerical results of these activation spaces are detailed in Section 4.5.1.

3.8 Automatic Speech Detection from EMG

During acquisition, utterance boundaries are defined by the visual cue whose boundaries are however unavailable during real-time operation. To extend to practical and online systems, they must be able to detect them directly from the continuous EMG signal. This section introduces a trigger-free onset detector and evaluates it as a binary speech-versus-rest classifier against the original cue segmentation. The detector is evaluated separately from the recognition models (Section 4.8) and is not used to generate the windows for the experiments reported elsewhere in this thesis.

3.8.1 Detection Algorithm

The detector processes the fourteen filtered EMG channels and identifies intervals in which the subject articulates an utterance. The procedure consist of four stages [91, 92]: envelope extraction, normalization, channel aggregation, and threshold-based onset detection. All operations depend only on current and past samples, so the complete procedure can be implemented causally and it is important to mention that the trigger signal is employed only as ground truth for the evaluation described in Section 3.8.2.

Stage 1: envelope. Each channel is rectified and averaged over the previous $w_e = 30\,\text{ms}$ [18]. Denoting by $x_c[n]$ the filtered signal of channel c, the envelope is

$$e_c[n] \;=\; \frac{1}{W} \sum_{k=n-W+1}^{n} \big|x_c[k]\big|, \qquad W = \lfloor w_e f_s \rfloor, \tag{3.7}$$

where:

- $e_c[n]$ is the envelope of channel c at sample n;
- $x_c[k]$ is the filtered signal of channel c at sample k;
- W is the length of the averaging window in samples, and $w_e = 30\,\text{ms}$ its duration;
- $f_s = 500\,\text{Hz}$ is the sampling rate.

Stage 2: normalization. The envelope is measured in microvolts, and its level when the participant is not articulating an utterance differs from its level during an utterance. These levels also vary between channels, participants, and sessions. For this reason, a threshold applied directly to $e_c[n]$ would need to be adjusted for each recording. Each envelope is instead normalized with respect to its recent activity

$$z_c[n] \;=\; \frac{e_c[n] - \mathrm{LO}_c[n]}{\mathrm{HI}_c[n] - \mathrm{LO}_c[n]}, \tag{3.8}$$

where:

- $z_c[n]$ is the activation of channel c at sample n;
- $e_c[n]$ is the envelope of Equation (3.7);
- $\mathrm{LO}_c[n]$ is the 20th percentile of the previous $B = 20\,\text{s}$ of the envelope;
- $\mathrm{HI}_c[n]$ is the 80th percentile of the same interval.

Table 3.4: Constants of the trigger-free detector used for all participants, sessions, speaking conditions, and corpora.

Symbol	Parameter	Value	Role
W	Smoothing window	30 ms	Length of the causal moving average
B	Baseline window	20 s	History used to estimate the percentiles
U	Baseline update period	1 s	Frequency of percentile updates
	Baseline percentiles	20 / 80	Resting level and dynamic range
τ_w	Warm-up	5 s	History required before detection starts
K	Channels averaged	4	Number of active channels
t_{high}	Confirmation threshold	2.5	Threshold that confirms an event
t_{low}	Onset threshold	1.0	Threshold used to place the onset
M	Backward search	0.30 s	Maximum backward search for the onset
G	Merge gap	0.30 s	Events closer than this are merged
D	Minimum duration	0.15 s	Shortest accepted event
	Match tolerance	0.50 s	Maximum lead allowed for a cue match

The value z_c is dimensionless by construction, and a value close to zero denotes a channel at its non-articulating level, regardless of its absolute amplitude in microvolts. This level is estimated from recent samples, since movement artifacts can produce amplitudes far exceeding those of speech. Using the global signal would require the entire recording to be available, making it unsuitable for a causal implementation.

Stage 3: channel aggregation. At each sample, the activity trace $a[n]$ is computed as the mean of the $K = 4$ largest normalized envelopes $z_c[n]$ channel values. Using the most active channels allows the detector to follow the electrodes that carry the strongest activity without making assumptions about which muscles are dominant. The sensitivity to the number of selected channels is evaluated in Section 4.8.1.

Stage 4: dual threshold and backward search. An event is accepted only when the activity trace crosses a high confirmation threshold t_{high}. Once confirmed, the onset is placed at the last preceding crossing of the lower threshold t_{low}, searched backwards for up to $M = 0.30$ s [93, 92]. Then, events separated by less than $G = 0.30$ s are merged, so that short pauses inside a sentence do not split one utterance into multiple events, and events shorter than $D = 0.15$ s are discarded.

3.8.2 Evaluation against the Trigger

The detector is evaluated as a binary speech-versus-rest classifier using the cue segmentation employed by the windowing stage (Section 3.2.1). Each cue interval is considered detected when it contains at least one onset, allowing detections up to

`match_tolerance_s`=0.50 s before the cue. A cue without a corresponding onset is a false negative; an onset falling inside a rest interval is a false positive, while a rest interval without an onset is a true negative. The number of detected events inside one cue does not affect the decision at the sentence level: one or more events still count as a single detected cue.

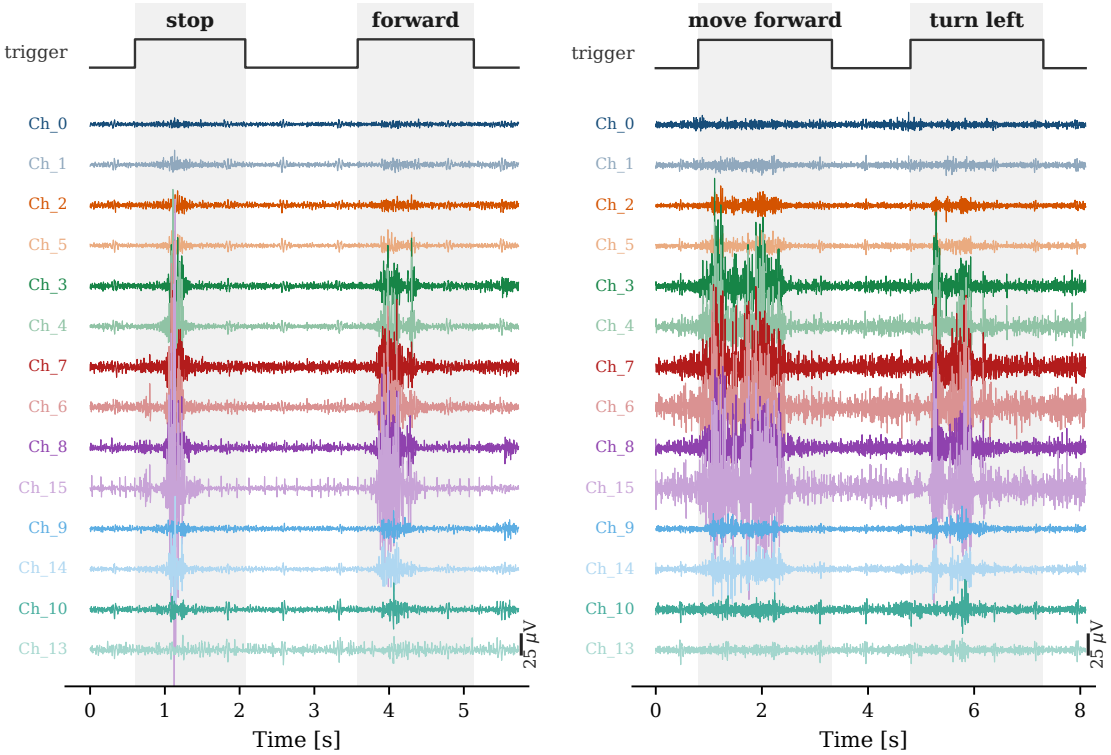
Chapter 4

Closed-Set Classification: Results

This chapter presents the results of the closed-set classification study conducted using the methods described in Chapter 3. First, the word subset is used to reproduce the published baseline [4] and evaluate the ablations, while the following analyses focus on the sentence task. Unless stated otherwise, the results are expressed as balanced accuracy, as defined in Section 3.7, and values reported after a $\pm$ sign denote the standard deviation across participants and therefore reflect variability across the population. Figure 4.1 shows the signal from which these results are obtained.

4.1 Order of the Experiments

The experiments in Section 3.5 compare three elements: the training objective, the input representation, and the sequence model, all evaluated under two protocols and two speaking conditions. Training every combination is computationally too expensive. Therefore, the evaluation proceeds with each experiment building on the best settings from the previous one. This progression mirrors the order in which the system was developed. First, the system baseline [4] is reproduced on both its original corpus and the new dataset. This confirms that the reimplemented pipeline and the new recordings match the published results before analyzing any sentence-level data. The evaluation then moves to the training objective. The training objective and the recurrent stage are tested together and once a working CTC model is in place, the experiments compare three input formats, eventually keeping the STFT spectrogram. Finally, the sequence models are compared at an equal number of parameters and both of them move forward to the decoder sweep and into Chapter 6. Three ablations and the trigger-free detector are then



(a) Word task. **(b)** Sentence task.

Figure 4.1: The input signals from which the results are obtained. The panels display two consecutive words, *stop* and *forward*, alongside two consecutive sentences, *move forward* and *turn left* (subject S02, vocalized, session 1), following processing of Chapter 3.

evaluated separately: the detector is scored as a binary classifier separating speech from rest, and its output is not used to train any models in this chapter.

4.2 Baseline Reproduction and Validation of the Corpus

SpeechNet is rebuilt from its configuration (Section 3.5.1 and Table 3.2a) and trained on the four participants of the SilentWear corpus at the 1.4 s window. Under global cross-validation, it reaches 78.8 ± 8.3% balanced accuracy vocalized and 73.3 ± 5.7% silent, compared to the 84.8 ± 4.6% and 77.5 ± 6.6% reported by the authors [4]. Under the inter-session protocol, it scores 72.8 ± 9.5% and 63.5 ± 6.9%, slightly outperforming the declared 71.1 ± 8.3% and 59.3 ± 2.2%. When the SpeechNet model is trained on the word subset of this thesis corpus,

balanced accuracy falls to 63.1 ± 9.9% vocalized and 50.4 ± 1.6% silent. This drop is due to task difficulty since the new label set expands from nine to sixteen classes and includes seven near-synonyms, making the vocabulary much more confusable (Section 2.3.1). To confirm that this issue is not attributable to the quality of the collected data, the architecture is upgraded. By replacing the original backbone with the STFT stack of Table 3.2c and a recurrent stage, the global balanced accuracy on the new corpus rises to 75.0 ± 10.5% vocalized and 71.4 ± 6.5% silent. In the same configuration, the inter-session evaluation reaches 57.4 ± 11.6% vocalized and 60.3 ± 7.4% silent. Furthermore, applying this upgrade to the published corpus yields 77.2 ± 11.7% vocalized and 66.0 ± 6.2% silent in the inter-session experiment, and 92.8 ± 3.5% vocalized and 89.9 ± 2.3% silent in the global one, outperforming the original baseline. Finally, applying the CTC objective to the SpeechNet causes balanced accuracy to collapse, dropping to 48.9 ± 7.4% vocalized and 42.3 ± 6.5% silent evaluated with the global setup and to 44.0 ± 7.8% vocalized and 36.8 ± 5.0% silent with the inter-session protocol. In this case, pooling the time axis by a factor of 128 leaves only five frames per 1.4 s window, which is not enough with words of up to eight characters and violates the constraint of Section 3.4. It proves that CTC cannot just replace the CE head on an existing network, but rather requires a backbone that preserves more time steps, which also motivates the pooling introduced in Table 3.2.

Table 4.1: Balanced accuracy (%) for the reproduced SpeechNet baseline. Each value is the mean over the folds and then over the participants, ± one standard deviation across them, four participants in the upper block and three in the lower. Reported results are evaluated with the global and the inter-session protocol.

		Global evaluation		**Inter-session evaluation**	
Corpus and model	**Obj.**	**Voc**	**Sil**	**Voc**	**Sil**
SilentWear corpus					
SpeechNet, as published [4]	CE	84.8 ± 4.6	77.5 ± 6.6	71.1 ± 8.3	59.3 ± 2.2
SpeechNet	CE	78.8 ± 8.3	73.3 ± 5.7	72.8 ± 9.5	63.5 ± 6.9
SpeechNet	CTC	48.9 ± 7.4	42.3 ± 6.5	44.0 ± 7.8	36.8 ± 5.0
STFT backbone + BiLSTM	CE	92.8 ± 3.5	89.9 ± 2.3	77.2 ± 11.7	66.0 ± 6.2
New corpus					
SpeechNet	CE	63.1 ± 9.9	50.4 ± 1.6	61.5 ± 13.0	50.1 ± 4.1
STFT backbone + BiLSTM	CE	75.0 ± 10.5	71.4 ± 6.5	57.4 ± 11.6	60.3 ± 7.4

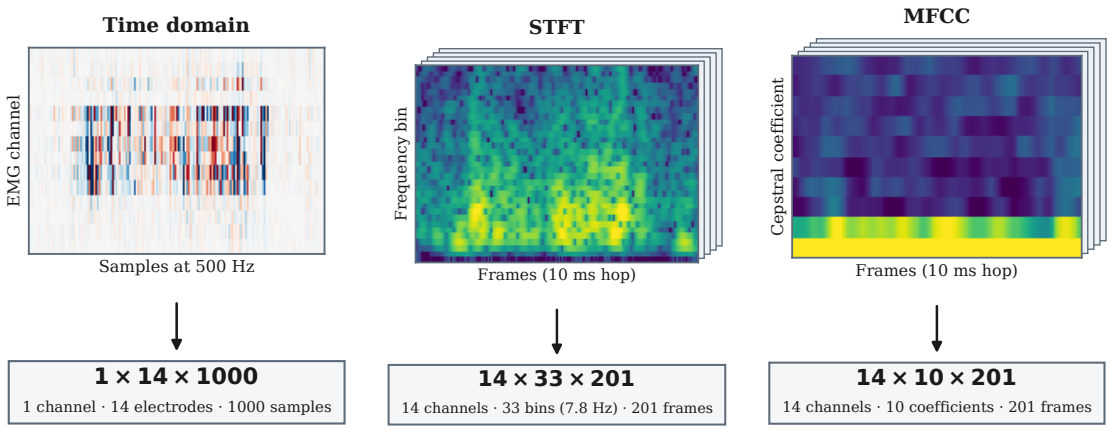
4.3 Training Objective and Sequence Model

The baseline SpeechNet model cannot be trained with CTC as it is. It is replaced by the backbone of Table 3.2, which leaves 50 frames per window and 128 features

per frame, and a two-layer BiLSTM is placed above it. This section evaluates the objective and the sequence stage in all seven participants, using greedy decoding with a 2.0 s window for sentences, and a 1.4 s window for words. These results correspond to the STFT rows of Table 4.2, while the next section extends the comparison to the other input representations. The sequence model yields the most significant performance gain. Using CE and the STFT representation, adding the recurrent stage to the backbone increases the balanced accuracy on sentences by 14.6 points (from 65.8 ± 17.7% to 80.4 ± 10.3%) in the global vocalized setting and by 17.9 points (from 62.5 ± 17.6% to 80.4 ± 8.0%) in the global silent one. Under the inter-session protocol, the same addition is worth 23.0 points (from 45.3 ± 14.1% to 68.3 ± 7.3%) vocalized and 18.3 points (from 47.7 ± 14.8% to 66.0 ± 10.6%) silent. In the time domain, where the CNN reaches 52.4 ± 16.3% globally and 39.6 ± 12.7% across sessions in the vocalized condition, this addition provides even larger gains of 23.6 and 21.7 points. This improvement comes from the expanded receptive field. Indeed, a convolutional network only connects frames locally through kernels and pooling layers, failing to capture the long-range dependencies between the beginning and the end of a sentence. Capturing these distant relations is precisely the capability that the corpus was designed to evaluate. When combined with the sequence model, the objective CTC also outperforms CE in the sentence task. It yields a 13.6 point increase in the global vocalized setting (94.0 ± 3.5% versus 80.4 ± 10.3%) and a 7.4 point increase in the global silent one (87.8 ± 8.9% versus 80.4 ± 8.0%). Under the inter-session protocol the gains are 15.6 points vocalized (83.9 ± 5.9% versus 68.3 ± 7.3%) and 11.9 points silent (77.9 ± 13.3% versus 66.0 ± 10.6%). In contrast, applying CTC to the word task causes a 2.6-point drop in the global vocalized setting (73.5 ± 19.1% versus 76.1 ± 11.6%) and a 2.9-point increase in the inter-session vocalized one (63.4 ± 15.4% versus 60.5 ± 11.8%). This asymmetry arises from the linguistic structure available to the objective function. Sentences consist of up to 24 characters and share a restricted vocabulary by design (Section 2.3.1). This shared sub-word structure allows a per-frame character model to transfer learned patterns across different commands, such as between *turn left*, *turn right*, and *turn on the light*. Isolated words lack these properties, and targets are short and classes share very little overlap. Consequently, CTC is effective for the sentence task, which demonstrates that the word subset serves as a control experiment. Finally, CE is excluded from further evaluation, as it does not produce per-frame token probabilities and cannot generate a character sequence. Therefore, it is incompatible with the continuous recognition task detailed in Chapter 5.

4.4 Input Representations

The three input representations from Section 3.3 are compared between all seven participants and both sets of labels, the twenty sentences and rest at the 2.0 s window and the fifteen words and rest at the 1.4 s one. All evaluations include the rest condition and rely on greedy decoding, as detailed in Table 4.2. To ensure that the performance of these representations is independent of the training objective and the sequence model, the comparison is repeated across all three model configurations: the convolutional backbone alone under CE, the backbone with the recurrent stage under CE, and the same recurrent model under CTC. Finally, Figure 4.2 illustrates the output of each representation for the exact same signal window.



(a) Time domain **(b)** STFT **(c)** MFCC

Figure 4.2: The three input representations of an EMG window. Panel (a): raw window without transformation. Panel (b): $N = 64$ samples (128 ms), hop 5 samples (10 ms), power, dB. Panel (c): 15 mel filters, 10 coefficients, hop 5 samples (10 ms).

4.4.1 STFT Representation

The STFT representation outperforms the time-domain representation across all comparisons. Under global cross-validation on sentences, STFT achieves a 4.4 point advantage in the vocalized setting (80.4 ± 10.3% versus 76.0 ± 11.5%) and an 8.9 point advantage in the silent setting (80.4 ± 8.0% versus 71.5 ± 10.1%). The same pattern is observed in the inter-session protocol, with margins of 7.0 points (68.3 ± 7.3% versus 61.3 ± 9.2%) and 7.4 points (66.0 ± 10.6% versus 58.6 ± 10.4%), respectively. With the CTC objective, the differences become

even larger, reaching 7.1 and 15.8 points under global cross-validation and 7.2 and 17.1 points in the inter-session setting. Model capacity does not explain this difference in performance and the two architectures cannot be matched in size, with 79 328 parameters in the time-domain backbone and 244 704 in the spectral one. However, the results obtained with the MFCC representation rule out the number of parameters as the main cause. Since the MFCC representation uses the same spectral backbone, it has the same number of parameters as the STFT model. Despite this, the MFCC representation performs 5.2 points below the smaller time-domain baseline in the global vocalized setting (70.8 ± 14.2% versus 76.0 ± 11.5%) and is level with it in the inter-session setting (61.2 ± 11.6% versus 61.3 ± 9.2%). The larger model therefore does not outperform the smaller one, indicating that the input representation, rather than model capacity, is the main factor determining the results.

Table 4.2: Balanced accuracy (%) for all combinations of the three input domains, sequence stages, and training objectives. Each value, obtained with greedy decoding, is the mean over the folds and then over the participants, ± one standard deviation across them, seven participants over six sessions. Reported results are evaluated with the global and the inter-session protocol, and the best domain of each configuration is in bold.

		Global evaluation		**Inter-session evaluation**	
Domain	**Configuration**	**Voc**	**Sil**	**Voc**	**Sil**
(a) Sentences, 20 sentences + rest, 21 classes, 2.0 s					
Time	CE	52.4 ± 16.3	41.8 ± 18.9	39.6 ± 12.7	32.8 ± 14.9
MFCC 15/10	CE	51.1 ± 19.6	49.9 ± 16.1	35.9 ± 13.3	40.1 ± 13.6
STFT	CE	**65.8** ± 17.7	**62.5** ± 17.6	**45.3** ± 14.1	**47.7** ± 14.8
Time	CE + BiLSTM	76.0 ± 11.5	71.5 ± 10.1	61.3 ± 9.2	58.6 ± 10.4
MFCC 15/10	CE + BiLSTM	70.8 ± 14.2	71.7 ± 10.5	61.2 ± 11.6	61.2 ± 12.1
STFT	CE + BiLSTM	**80.4** ± 10.3	**80.4** ± 8.0	**68.3** ± 7.3	**66.0** ± 10.6
Time	CTC + BiLSTM	86.9 ± 9.2	72.0 ± 16.4	76.7 ± 10.8	60.8 ± 18.1
MFCC 15/10	CTC + BiLSTM	79.3 ± 13.5	75.3 ± 14.2	68.3 ± 18.7	66.6 ± 17.2
STFT	CTC + BiLSTM	**94.0** ± 3.5	**87.8** ± 8.9	**83.9** ± 5.9	**77.9** ± 13.3
(b) Words, 15 words + rest, 16 classes, 1.4 s					
Time	CE	57.4 ± 19.9	42.3 ± 12.8	44.6 ± 17.9	33.7 ± 10.7
MFCC 15/10	CE	57.1 ± 15.1	45.7 ± 11.3	43.2 ± 13.3	37.0 ± 10.9
STFT	CE	**67.8** ± 16.0	**55.6** ± 10.8	**51.0** ± 14.2	**43.9** ± 10.2
Time	CE + BiLSTM	67.8 ± 19.8	56.0 ± 15.2	57.0 ± 20.2	46.9 ± 13.4
MFCC 15/10	CE + BiLSTM	63.5 ± 14.5	56.8 ± 11.0	52.1 ± 14.4	48.3 ± 10.1
STFT	CE + BiLSTM	**76.1** ± 11.6	**68.6** ± 8.2	**60.5** ± 11.8	**55.5** ± 11.5
STFT	CTC + BiLSTM	73.5 ± 19.1	67.6 ± 10.6	63.4 ± 15.4	55.8 ± 14.6

Length of the Analysis Window The STFT representation has two constants, and the analysis window is the one that decides the trade-off between time and frequency resolution. It is swept from 32 to 512 ms on the word subset, with the recurrent model and the CTC objective, over the three participants and six sessions (Table 4.3). The five settings share every other element of the configuration. Between 32 and 256 ms, the balanced accuracy across the four evaluation settings fluctuates by no more than 6 points. The 128 ms window achieves the highest accuracy in three of these settings, while 32 ms is optimal for the fourth. Expanding the window to 512 ms causes the balanced accuracy to drop to 59.9 ± 15.5% in the global vocalized setting (10 points below the 70.0 ± 15.9% achieved by the 128 ms window) and to 41.1 ± 13.5% in the inter-session silent protocol (fifteen points below 56.2 ± 7.6%). This performance drop occurs because a half-second window oversmooths the brief muscular bursts that separate individual words. On contrast, a 32 ms window provides insufficient frequency resolution, partitioning the relevant 20 to 250 Hz band into nine bins. Therefore, the chosen 128 ms window offers an effective compromise, providing 33 bins with a 7.8 Hz frequency resolution.

Table 4.3: Balanced accuracy (%) against the length of the STFT analysis window, on the word subset at the 1.4 s window, CTC objective with the recurrent stage, greedy decoding. Each value is the mean over the folds and then over the participants, ± one standard deviation across them, three participants over six sessions. Reported results are evaluated with the global and the inter-session protocol, and the adopted setting is in bold.

		Global evaluation		**Inter-session evaluation**	
Window	**Samples and bins**	**Voc**	**Sil**	**Voc**	**Sil**
32 ms	16 samples, 9 bins	69.7 ± 14.5	65.8 ± 11.6	56.8 ± 16.0	**57.9** ± 9.3
64 ms	32 samples, 17 bins	63.6 ± 26.3	63.4 ± 14.0	54.4 ± 23.2	54.8 ± 15.3
128 ms	64 samples, 33 bins	**70.0** ± 15.9	**65.9** ± 10.3	**57.4** ± 17.6	56.2 ± 7.6
256 ms	128 samples, 65 bins	66.8 ± 15.7	63.0 ± 10.5	53.1 ± 15.0	51.2 ± 9.5
512 ms	256 samples, 129 bins	59.9 ± 15.5	52.7 ± 12.1	45.4 ± 14.6	41.1 ± 13.5

4.4.2 MFCC Representation

Unlike the time domain and STFT, the MFCC requires setting domain-specific hyperparameters rather than directly extracting features from the raw signal. The filter bank is fitted to the EMG bandwidth, with 15 filters and 10 coefficients (Section 3.3). Each filter then groups approximately 8 of the 65 FFT bins and provides the intended reduction in variance. At a 500 Hz sampling rate, the mel scale is essentially linear across the entire 0–250 Hz band, so the filter bank applies no frequency warping and only partitions the band into intervals of equal width. Under CE with the BiLSTM model, the deficit against STFT is 4.8 to 9.6 points,

and a deficit of 6.2 to 7.2 points persists under CTC with the Transformer encoder (Table 4.4). This gap aligns with findings by Wu et al. [25], who attribute such losses to the information discarded by the decorrelating cosine transform. The cross-participant variance is also higher for the MFCC, indicating poorer generalization across subjects. In addition, the STFT incorporates logarithmic compression, preserving fine-grained spectral detail. Consequently, the STFT power spectrogram is selected as the input representation for the remainder of this thesis. Discarding fine spectral detail prior to feature extraction introduces an information loss that downstream sequence models cannot recover.

4.5 BiLSTM versus Transformer at Equal Number of Parameters

The final dimension of the evaluation framework is the sequence model. The BiLSTM and Transformer are compared across all seven participants using the STFT representation, the CTC objective, and a 2.0 s window. Their number of parameters are nearly identical (910 328 and 909 816, respectively), ensuring that any performance differences are architectural and not capacity-driven. The standalone backbone is included as a baseline (Table 4.4), and greedy decoding is applied uniformly, as decoding strategies are evaluated in the subsequent step. The backbone isolates the impact of the additional 660 000 parameters introduced by the sequence models. On its own, it achieves a balanced accuracy of 38.6 ± 17.0% (vocalized) and 26.5 ± 14.7% (silent) under global cross-validation, and 27.9 ± 12.1% and 21.4 ± 12.6% across sessions. Adding the Transformer encoder raises these scores to 94.2 ± 3.9% and 90.7 ± 6.1% under global cross-validation, and to 85.8 ± 5.2% and 82.1 ± 10.5% in the inter-session setting. The sequence stage thus provides a 55 to 64 point improvement, representing the largest performance gain observed in this study. The Transformer outperforms the BiLSTM across all four evaluation settings. Under global cross-validation, it provides a 0.2 point edge in the vocalized condition (94.2 ± 3.9% versus 94.0 ± 3.5%) and a larger 2.9 point advantage in the silent one (90.7 ± 6.1% versus 87.8 ± 8.9%). In the inter-session protocol, the gap increases to 1.9 points (85.8 ± 5.2% versus 83.9 ± 5.9%) and 4.2 points (82.1 ± 10.5% versus 77.9 ± 13.3%), respectively. Furthermore, the Transformer responds better as the difficulty of the task increases: silent speech reduces its accuracy by 3.5 points compared to 6.2 points for the BiLSTM model, and electrode repositioning causes a drop of 8.4 points compared to 10.1 points. This stability derives from the self-attention mechanism, which connects any two frames in a single operation, unlike a recurrent network that must propagate information sequentially through its hidden state. Furthermore, this architectural advantage remains consistent regardless of the decoding strategy applied. Replacing greedy

decoding with the prefix beam search optimized in Section 4.6 shifts the accuracy by at most 0.8 points without altering the hierarchy. Although the recurrent model benefits slightly more from beam search than the encoder does, the Transformer maintains its lead across all settings. For these reasons, the Transformer encoder is the optimal sequence model for the closed-set classification task. The confusion matrices of this configuration, one per participant and per speaking condition, are reported in Appendix C.3.

Table 4.4: Sentence classification balanced accuracy (%), comparing the two sequence models at an equal number of parameters. Data show mean ± standard deviation across all participants (using STFT, CTC, 2.0 s) windows, and greedy decoding). Best scores are in bold. See Table C.1. for per-participant details.

			Global evaluation		**Inter-session evaluation**	
Domain	**Seq. stage**	**Params**	**Voc**	**Sil**	**Voc**	**Sil**
STFT	none	247 542	38.6 ± 17.0	26.5 ± 14.7	27.9 ± 12.1	21.4 ± 12.6
STFT	BiLSTM	910 328	94.0 ± 3.5	87.8 ± 8.9	83.9 ± 5.9	77.9 ± 13.3
STFT	Transformer	909 816	**94.2** ± 3.9	**90.7** ± 6.1	**85.8** ± 5.2	**82.1** ± 10.5
MFCC 15/10	Transformer	909 816	87.0 ± 9.5	84.5 ± 11.9	78.7 ± 12.5	75.3 ± 14.1

4.5.1 Learned Representation

The previous section evaluated the sequence stage based on its contribution to balanced accuracy. This impact can also be observed in the learned representation by analyzing the network activations before and after this stage. The analysis uses the Transformer model from Table 4.4, trained on the STFT representation with the rest class included. Activations are extracted from the fifth cross-validation fold of participant S01 at two locations: the output of the convolutional backbone and the output of the encoder. Once flattened across the time axis, both feature spaces consist of 6 400 dimensions. The nearest-centroid classifier detailed in Section 3.7 evaluates these two representation spaces (Table 4.5). In the global vocalized setting, classification accuracy increases from 66.9% at the backbone output to 88.3% after the encoder. Similar gains occur in the global silent setting (from 64.8% to 87.9%) and across sessions (from 70.8% to 87.6% in the vocalized condition and from 76.5% to 87.5% in the silent one). Finally, the sequence stage provides an 11 to 23 point improvement. Figure 4.3 shows these activations reduced to two dimensions for the fifth inter-session fold. Each panel displays a t-SNE using principal component initialization, a perplexity of 30, and a fixed random seed. Within each panel, the

Table 4.5: Balanced accuracy (%) achieved by a nearest-centroid classifier applied to the full network activations. Centroids are computed using the training and validation windows across 21 classes for the Transformer sequence model. Data corresponds to the fifth cross-validation fold of participant S01. Results are reported for both the global and inter-session protocols.

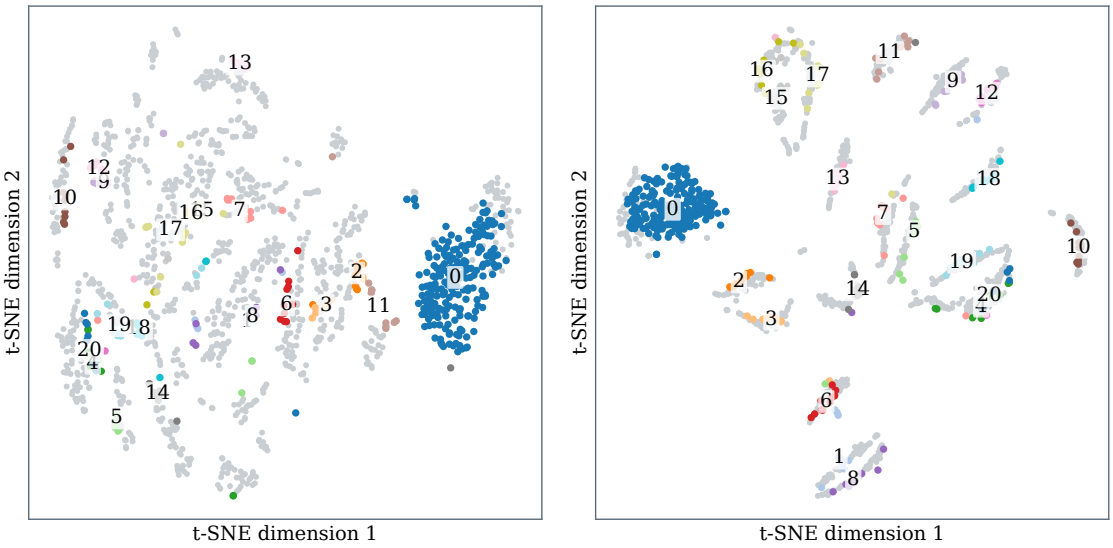
		Backbone output		**Encoder output**	
Protocol	**Cond.**	Train	Test	Train	Test
Global	Voc	75.6	66.9	96.0	**88.3**
Global	Sil	74.9	64.8	95.0	**87.9**
Inter	Voc	74.5	70.8	95.7	**87.6**
Inter	Sil	73.2	76.5	94.0	**87.5**

five training sessions and the held-out session are projected together, allowing the test windows to be compared against the training clusters they are drawn over. At the backbone output, class distributions heavily overlap and the held-out windows scatter between them, with only the *rest* class forming a clearly distinct cluster. At the encoder output, however, the training data resolves into well-separated regions, and most held-out classes map adjacent to or directly over their corresponding training clusters. This residual spatial shift visually demonstrates the inter-session performance drop quantified in Table 4.5, where the nearest-centroid accuracy falls from 95.7% on the training sessions to 87.6% on the held-out session. The same table shows that the encoder recovers class separation in the other conditions as well, maintaining an 87.9% nearest-centroid accuracy during global silent speech and 87.5% across silent sessions.

4.6 Decoding the CTC Classifier

Mapping the per-frame output probabilities to a final predicted class requires two consecutive decision stages. The first stage is character-level decoding, where greedy decoding is compared with a prefix beam search. Neither method incorporates an external lexicon during this phase; both generate unconstrained character sequences, and the closed-set vocabulary constraint is enforced afterward by mapping the generated string to a valid class. To optimize the prefix beam search, a hyperparameter grid search evaluates combinations of beam width, four softmax temperatures, four blank penalties, and four length bonuses. Figure 4.4 visualizes the per-frame posterior probabilities that represent the input for these decoding stages.

Comparing greedy decoding with the optimal prefix beam search configuration, the balanced accuracy of the Transformer increases from 92.5% to 92.8% under global cross-validation and from 84.0% to 84.8% across sessions. Similarly, the BiLSTM model improves from 90.9% to 91.7% and from 80.9% to 82.4% (Table 4.6).

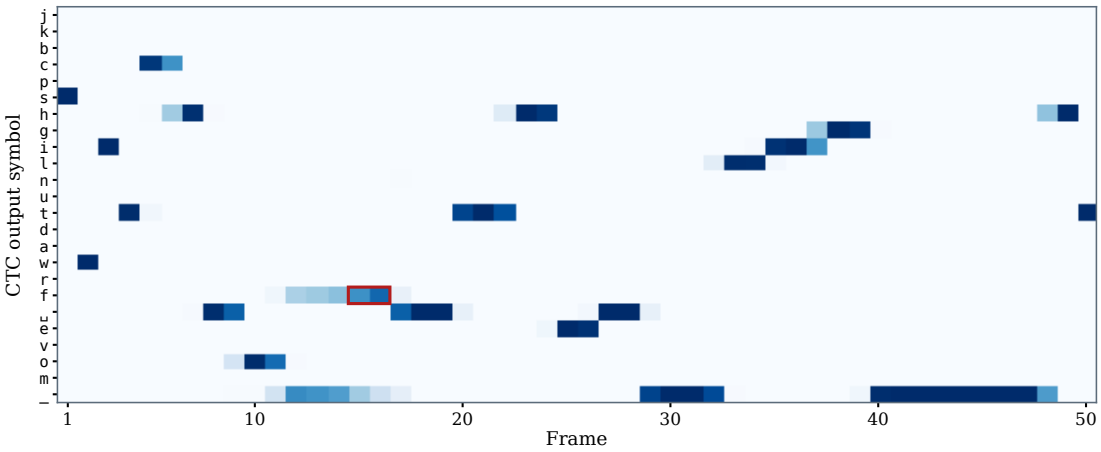


(a) Backbone output. **(b)** Encoder output.

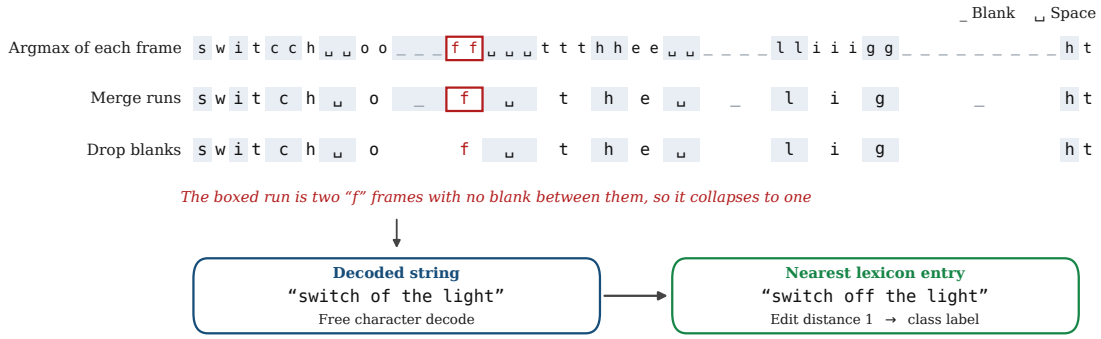
Figure 4.3: Two-dimensional t-SNE of the Transformer activations on the two sides of the sequence stage, for participant S01, vocalized speech and the fifth inter-session fold. Grey points are the windows of the five training sessions, coloured points those of the held-out session. Each cluster is labelled with its class index at the median of its held-out points (Appendix B, rest at index 0).

Because these figures are pooled across participants and speaking conditions, cross-participant variance is omitted. Consequently, the decoding strategy only yields an improvement of 0.3 to 1.5 points. This limited impact is due to the closed-set decision rule, which maps the generated string to the nearest lexicon entry using minimum edit distance. For a transcription improvement to affect classification accuracy, the corrected characters must successfully shift the closest match to the correct target class. However, the unconstrained balanced WER of the Transformer drops from 21.5% to 19.9% globally, and that of the BiLSTM falls from 18.0% to 12.9%.

The hyperparameter grid search highlights three key behaviors. First, increasing the beam width beyond five yields no transcription changes; the balanced accuracy and balanced WER for widths of 5, 10, and 25 agree within 0.1 points. A beam width of one (equivalent to greedy search) degrades balanced accuracy by 0.1 to 0.4 points. This saturation occurs because CTC training produces highly peaked token distributions, concentrating the probability mass along a single alignment path. Therefore, a beam width of five represents the optimal trade-off. Second, among the three scoring terms, only the softmax temperature provides a benefit: increasing it from $T_\theta = 1$ to $T_\theta = 2$ reduces the Transformer's balanced WER from 20.6% to 19.9%. Finally, the blank penalty degrades performance, increasing



(a) Per-frame posterior probabilities (24×50 matrix) produced by the classification head.



(b) Greedy decoding pipeline: frame-level argmax selection, consecutive duplicate merging, and blank token removal, followed by lexicon matching to assign the final class.

Figure 4.4: Mapping from per-frame CTC posteriors to a final class label.

the balanced WER from 20.6% to 23.7% as the penalty scales from zero to four. Similarly, the length bonus is neutral at 0.5 but drops the WER to 24.8% at a value of 2.0.

The hyperparameter configuration that maximizes classification accuracy often diverges significantly from the one that minimizes the error rate. For instance, in the inter-session evaluation of the BiLSTM model, the accuracy-optimal configuration uses a length bonus of 2.0. While this yields a 0.9 point gain in balanced accuracy, it increases the balanced WER from 25.4% to 55.7%. Consequently, all performance metrics throughout this thesis specify the objective function used for their optimization.

Table 4.6: Balanced accuracy and balanced WER (%) comparing decoding strategies on the classification branch. The error rate is the unconstrained one, without the vocabulary mapping of Section 5.4.3. All models use the STFT representation, 2.0 s windows, and include the *rest* condition (21 classes in total). The "plain beam" configuration utilizes default scoring hyperparameters, whereas the "tuned beam" rows report the optimal grid search configurations specific to the stated objective (maximizing accuracy or minimizing WER). Results are reported for both the global and inter-session protocols.

		Bal. accuracy (%)		**Bal. WER (%)**	
Model	**Decoder**	**Global**	**Inter**	**Global**	**Inter**
BiLSTM	Greedy	90.9	80.9	18.0	29.8
BiLSTM	Plain beam, $w = 10$	91.1	81.3	15.5	27.7
BiLSTM	Tuned beam, accuracy-optimal	**91.7**	**82.4**	20.4	55.7
BiLSTM	Tuned beam, WER-optimal	91.3	81.5	**12.9**	**25.4**
Transformer	Greedy	92.5	84.0	21.5	34.0
Transformer	Plain beam, $w = 10$	92.5	84.0	20.6	33.2
Transformer	Tuned beam, accuracy-optimal	**92.8**	**84.8**	21.2	41.3
Transformer	Tuned beam, WER-optimal	92.6	84.1	**19.9**	**32.6**

4.7 Ablation Studies

This section presents two ablation studies performed to assess the quantity and quality of the training by evaluating the number of enrollment sessions and the sliding-window augmentation detailed in Section 3.2.1. Both analyses are conducted on the isolated word subset using a 1.4 s window, evaluated on participants S01, S03, and S04 across all six of their recorded sessions.

4.7.1 Session-Count Scaling

This ablation evaluates how much data is required by retraining the SpeechNet model with the first $1, \ldots, 6$ sessions (Figure 4.5). Balanced accuracy improves as the number of sessions increases in all evaluation settings. Under global cross-validation, vocalized accuracy rises from $25.5 \pm 1.3\%$ with one session to $63.1 \pm 9.9\%$ with six, while silent accuracy increases from $18.7 \pm 8.4\%$ to $50.4 \pm 1.6\%$. The same trend is observed in the inter-session protocol, where accuracy increases from $28.2 \pm 10.2\%$ to $61.5 \pm 13.0\%$ for vocalized speech and from $21.3 \pm 6.3\%$ to $50.1 \pm 4.1\%$ for silent speech. The trend is not strictly monotonic: the vocalized curves decrease at five sessions before recovering at six. With five sessions in the leave-one-session-out protocol, each test fold contains 20% of the available data. Given the three-participant mean, a single unrepresentative session can therefore lead to a decrease being a consequence of the limited sample size and validation split.

Six sessions were the maximum that could be collected from each participant within the time allocated to data collection. The purpose of this ablation was therefore to determine whether this full allocation was necessary or whether a smaller number of sessions could provide comparable performance. The sixth session still produces substantial gains: 4.5 points in the global vocalized setting, 3.9 points in the global silent setting, and 19.9 and 5.1 points in the corresponding inter-session evaluations. Although the curves have not clearly plateaued at six sessions, these results indicate that six sessions are sufficient for the present protocol.

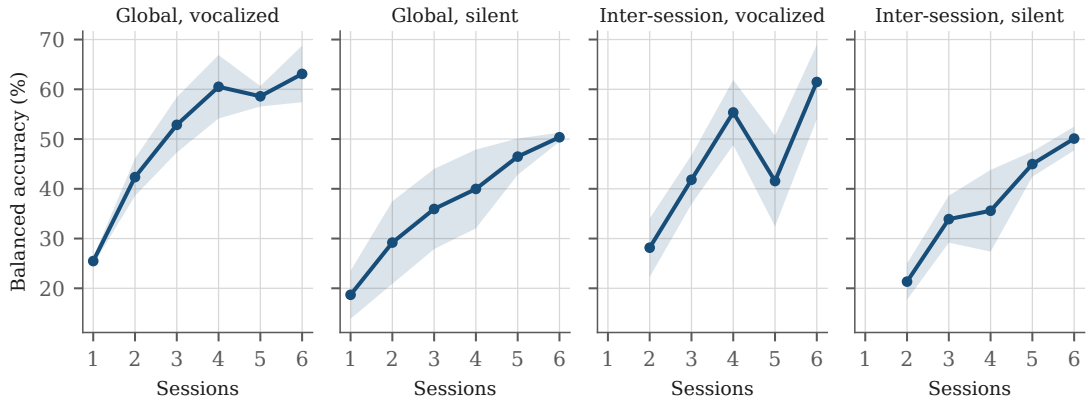


Figure 4.5: Balanced accuracy as a function of the number of enrollment sessions, averaged across participants S01, S03, and S04. Shaded regions denote ± one standard error across the three participants.

4.7.2 Sliding-Window Augmentation

This ablation study investigates the effectiveness of sliding-window augmentation by sweeping the stride (10, 20, 50, and 100 ms) and the number of shifts (2, 5, and 10 per side). The resulting twelve configurations, each trained on six sessions, are evaluated under the two modalities defined in Section 3.2.1. The constant-size modality fixes the total number of training windows to isolate the effect of temporal *placement* (Figure 4.6). In contrast, the augmented-size modality expands the dataset by a factor of $2N + 1$ to measure the impact of data *volume* (Figure 4.7). Both conditions utilize identical participants, raw windows, network architectures, and cross-validation folds. Under the constant-size condition (Table 4.7), augmentation yields no meaningful improvement. The nine variants with a maximum displacement of 0.25 s remain within 3.8 points of the no-augmentation baseline (which achieves 63.1 ± 9.9% and 50.4 ± 1.6% balanced accuracy globally for vocalized and silent speech). When the training set size expands, the trend reverses (Figure 4.7 and Table 4.8). In this case, the optimal configuration in each setting outperforms the baseline by 5.8 to 8.8 points and the number of shifts is not monotonical.

For instance, two shifts per side yield an average +7.0 point gain across the four settings, five shifts provide +4.9 points, and 10 shifts drop to +1.2 points. Stride has little effect at two shifts per side because excessive duplicates add no new information. The constant-size protocol shows that shifting the window has no effect for small shifts and decreases accuracy for large ones. In contrast, the augmented-size protocol shows that adding more data increases accuracy.

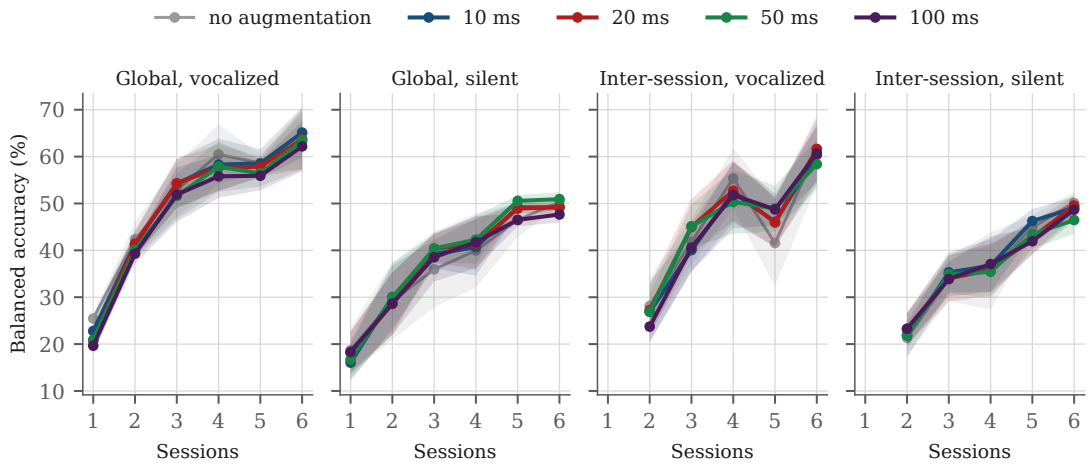


Figure 4.6: Sliding-window augmentation on the single-word subset at constant training-set size. Error bars indicate one standard deviation across the cross-validation folds (per participant) and across the three participants (for the average).

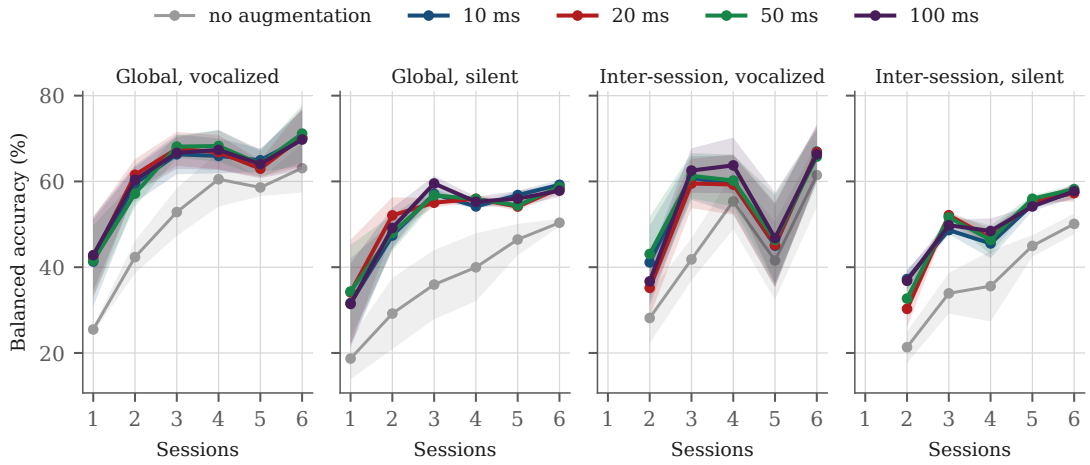


Figure 4.7: Sliding-window augmentation which expands the training set by a factor of $2N + 1$. Layout, participants, variants, and baseline are the same as Figure 4.6.

Table 4.7: Sliding-window augmentation with the pool resampled to the base cardinality. Evaluated on the single-word subset using a 1.4 s window, CE objective on the baseline backbone, at six sessions. Values represent balanced accuracy (%), averaged over the folds and then over the three participants (± one standard deviation across participants). The first row displays the no-augmentation baseline, shared with Table 4.8. The highest accuracy in each column is in bold.

	Global evaluation		**Inter-session evaluation**	
Stride and shifts per side	**Voc**	**Sil**	**Voc**	**Sil**
No augmentation	63.1 ± 9.9	50.4 ± 1.6	61.5 ± 13.0	**50.1** ± 4.1
Two shifts per side				
10 ms	65.1 ± 9.4	49.2 ± 1.1	60.4 ± 10.8	49.1 ± 4.2
20 ms	63.8 ± 11.2	49.3 ± 2.6	**61.6** ± 11.3	49.5 ± 3.9
50 ms	63.5 ± 10.9	**50.9** ± 2.6	58.4 ± 9.2	46.5 ± 5.1
100 ms	62.2 ± 9.2	47.7 ± 3.0	60.7 ± 10.6	48.7 ± 4.0
Five shifts per side				
10 ms	**65.3** ± 10.0	50.8 ± 3.8	60.1 ± 11.9	48.7 ± 4.5
20 ms	62.9 ± 9.3	50.4 ± 3.0	59.0 ± 9.6	49.4 ± 2.7
50 ms	64.1 ± 9.3	49.5 ± 2.4	59.7 ± 11.2	46.3 ± 2.9
100 ms	55.3 ± 9.4	44.1 ± 2.5	52.7 ± 11.4	39.7 ± 5.6
Ten shifts per side				
10 ms	63.5 ± 9.7	49.9 ± 3.6	60.9 ± 11.1	48.8 ± 3.0
20 ms	62.9 ± 10.5	48.1 ± 4.4	59.9 ± 11.2	47.9 ± 2.9
50 ms	53.4 ± 10.9	41.9 ± 5.9	54.7 ± 12.3	41.2 ± 4.1
100 ms	34.3 ± 5.1	30.6 ± 8.7	39.8 ± 11.0	29.1 ± 6.1

4.8 Trigger-Free Speech Detection

All previously reported metrics rely on windows anchored at the trigger. This section evaluates the algorithm detailed in Section 3.8, which autonomously locates speech onsets in the continuous EMG signal without external triggers. It first quantifies the algorithm's yield on the corpus and subsequently evaluates it as a binary speech-against-rest classifier. No classification model in this thesis is trained on these detected windows.

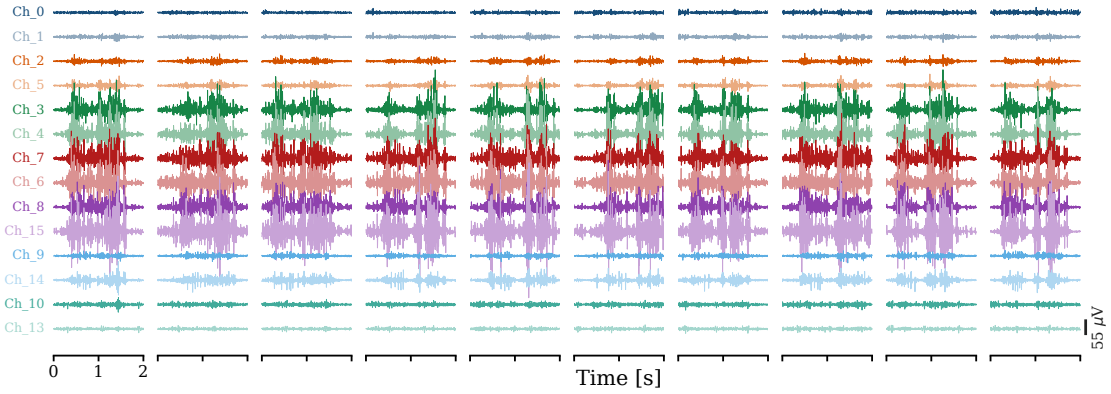
4.8.1 Corpus Yield and Classification Performance

On the sentence recordings, the visual cue box has a median duration of 2.51 s. However, the actual speech burst begins with a median delay of 0.31 s after the cue. A window attached to the trigger must account for this variable delay, while anchoring to the detected speech onset eliminates it (Figure 4.8). Finally, the 95th percentile of utterance durations averages 2.34 s across all recordings. Executed across the 84 recordings of the sentence corpus, the detector identifies 20 448 distinct

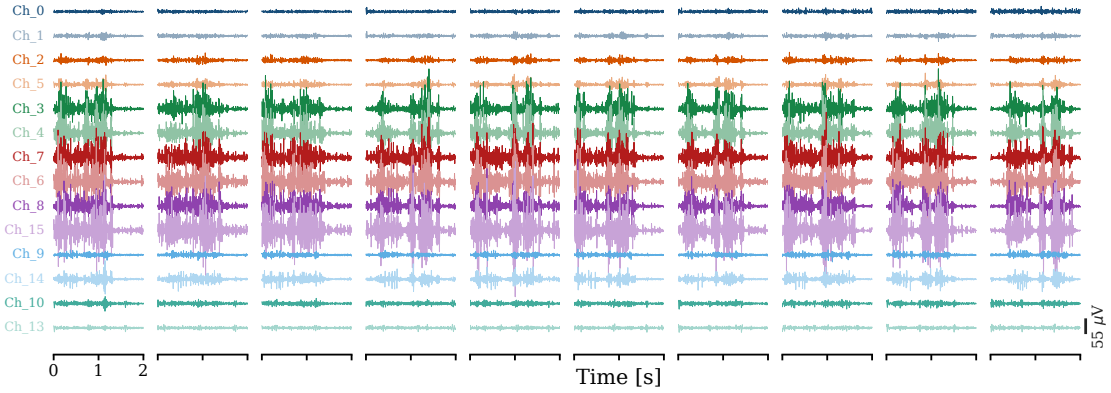
Table 4.8: Sliding-window augmentation considering the augmented data pool. Evaluated on the single-word subset using a 1.4 s window, CE objective on the baseline backbone, at six sessions. Values represent balanced accuracy (%), averaged over the folds and then over the three participants (± one standard deviation across participants). The first row displays the no-augmentation baseline, shared with Figure 4.6. The highest accuracy in each column is in bold.

	Global evaluation		**Inter-session evaluation**	
Stride and shifts per side	**Voc**	**Sil**	**Voc**	**Sil**
No augmentation	63.1 ± 9.9	50.4 ± 1.6	61.5 ± 13.0	50.1 ± 4.1
Two shifts per side				
10 ms	70.1 ± 11.1	**59.2** ± 1.9	66.9 ± 11.1	57.5 ± 2.4
20 ms	70.3 ± 11.9	58.4 ± 2.1	66.8 ± 10.8	57.3 ± 3.4
50 ms	**71.1** ± 11.5	58.7 ± 0.6	65.8 ± 10.8	**58.2** ± 1.6
100 ms	69.8 ± 12.2	57.8 ± 2.6	66.3 ± 10.8	57.7 ± 3.3
Five shifts per side				
10 ms	67.1 ± 9.8	57.8 ± 2.2	65.7 ± 8.8	55.2 ± 4.5
20 ms	67.7 ± 9.2	56.5 ± 1.3	**67.3** ± 10.7	56.3 ± 2.7
50 ms	66.9 ± 10.8	57.2 ± 3.9	66.0 ± 9.9	54.6 ± 1.7
100 ms	66.7 ± 9.3	55.1 ± 4.3	64.4 ± 10.3	54.3 ± 4.3
Ten shifts per side				
10 ms	65.4 ± 10.1	52.6 ± 1.1	61.3 ± 7.4	54.9 ± 3.6
20 ms	64.5 ± 9.4	53.0 ± 3.0	62.2 ± 9.1	53.8 ± 3.3
50 ms	62.7 ± 7.9	51.2 ± 1.3	64.2 ± 9.5	52.9 ± 3.0
100 ms	63.1 ± 8.6	50.1 ± 2.3	58.1 ± 7.3	49.4 ± 4.0

events. Of these, 799 fall outside all cue boxes and account for the false-positive rate, averaging 0.70 occurrences per minute of recording. The remaining 19 343 attributed events cover 16 340 of the 16 800 target utterances, yielding a recall of 97.3%. The poorest recording-level performance is 69.5% (participant S04, silent condition, session 3), whereas all other recordings exceed 91%. Figure 4.9 shows the two segmentations on one recording. Evaluated as a binary speech-against-rest classifier (Table 4.9), the system achieves a specificity of 93.6% and a precision of 93.8%, with an accuracy of 95.4%. Notably, 84 of the 460 missed utterances out of 16 800 occur within the first 5 s of a recording, a warm-up period where the detector lacks sufficient history to establish its dynamic baseline; excluding these initial seconds raises the recall to 97.7%. Consistent with stronger muscular recruitment, the vocalized condition outperforms the silent condition (98.0% versus 96.6% recall). A hyperparameter sweep yields four observations that dictate the system's operating point. First, the optimal confirmation threshold is 2.5 and lower values fragment the utterances. Raising this threshold to 2.8 would eliminate 133 false positives, but at the cost of discarding 14 valid utterances. Second, a reduced onset threshold



(a) Windows anchored at the visual cue.



(b) Windows anchored at the detected onset.

Figure 4.8: Ten repetitions of the command *move forward* (participant S02, vocalized condition, session 1), extracted using two distinct anchoring strategies. The windows in panel (a) begin at the visual cue, while the windows in panel (b) begin at the algorithmically detected onset.

captures earlier boundaries without increasing false positives, allowing the adopted 0.4 ratio to recover around 160 ms of initial articulation. Third, the merge gap is highly sensitive because increasing it to 0.80 s merges consecutive sentences and discards over one-third of the individual utterances. Finally, averaging between two and eight channels per sample yields identical performance, whereas averaging all fourteen drops the detection rate to 86.7%.

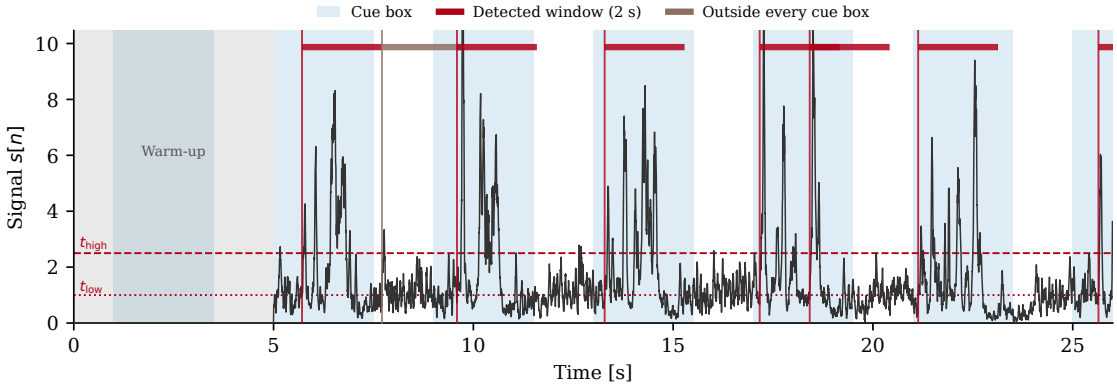


Figure 4.9: The first 26 s of one sentence recording (participant S01, vocalized condition, session 1), with the signal $s[n]$ and the two thresholds of Table 3.4. The shaded spans are the cue boxes of the protocol, the bars along the top are the 2.0 s windows opened by the detected onsets.

Table 4.9: Performance of the trigger-free detection algorithm evaluated as a binary speech-against-rest classifier over the 84 recordings of the sentence corpus. Metrics are computed globally, thus cross-participant variance is not defined.

	Per segment	**Per sample**
Decisions	33 684	34 076 956
Of which speech	16 800	21 228 131
True positives	16 340	10 232 166
False negatives	460	10 995 965
False positives	1 085	830 347
True negatives	15 799	12 018 478
Recall (%)	97.3	48.2
Specificity (%)	93.6	93.5
Precision (%)	93.8	92.5
F_1	0.955	0.634
Balanced accuracy (%)	95.4	70.9
Matthews correlation	0.909	0.432

Chapter 5

Sentence Continuous Recognition: Methods

While Chapters 3 and 4 treat an utterance as a single label from a closed set, this chapter transitions to continuous recognition. By removing the closed-set constraint, the decoder produces a character string from a fixed alphabet. Although the task changes, the pipeline, input representation, backbone, sequence models, CTC objective, and training schedule are shared with Chapter 3. The following sections detail the required modifications and the corresponding evaluation metrics to perform continuous recognition.

5.1 Problem Formulation

The continuous recognition network processes 14 filtered EMG channels to generate a character string. To achieve this, the network outputs a frame-wise probability distribution over a fixed alphabet, which the CTC collapse operation then converts into a final text string. The model trains on the same dataset, transcribing a rest window to the word *rest* and sentence windows to their corresponding texts, thereby yielding twenty-one reference strings. The token set comprises the 26 English letters, a space character, and the CTC blank at index 0, totaling 28 output dimensions. This broader alphabet enables open-vocabulary recognition rather than restricting the output space to the dataset vocabulary. In addition, since every reference text originates from the protocol detailed in Section 2.3.2, this task evaluates the open-vocabulary decoding of a closed set.

5.2 From Classification to Continuous Recognition

Four modifications differentiate this approach from the classification pipeline. First, a fixed 28-symbol alphabet replaces the lexicon-derived characters, widening the projection head. The complete model then has 911 356 parameters with the BiLSTM and 910 332 with the Transformer (Table 3.3). Second, the system outputs the collapsed string without mapping it to a predefined class, bypassing the cascade described in Section 3.4. Consequently, it measures any discrepancy from the text, meaning every frame-level prediction influences the final output. Third, error rates (Section 5.4.3) replace balanced accuracy as the evaluation metric. Finally, the continuous recognition runs select the checkpoint with the lowest validation error rate instead of the lowest validation loss. All other architectural components remain unchanged including the input windows, the STFT representation with spectral augmentation, the convolutional backbone (Table 3.2c), the sequence models (Table 3.3), and the optimization strategy (Section 3.7). The temporal alignment constraint also persists, mapping 50 input frames to a maximum target length of 24 characters.

5.3 Decoding

The continuous recognition pipeline reuses the decoding methods from Section 3.4, specifically the greedy best-path, the prefix beam search, and the components that differentiate the beam from the argmax path, without modification and without a language model. Because the collapsed string is the final prediction, the metric penalizes any inserted or deleted character and an empty prediction results in a deleted utterance. Figure 5.1 illustrates these modified stages, where the search deliberately avoids constraining the output to predefined texts. Applying such constraints to a wide beam would force convergence to a closed-set argmax, thereby reducing the evaluation to a classification measure. Consequently, this task assesses open-vocabulary generation and Section 4.6 details the interpretation of these decodes.

5.4 Evaluation Settings and Metrics

5.4.1 Cross-Validation Protocols

The evaluation repeats the capacity-matched comparison from Section 3.5.4, using the two sequence models detailed in Table 3.3 with the wider output head as the only architectural change. However, the two tasks evaluate the model outputs differently.

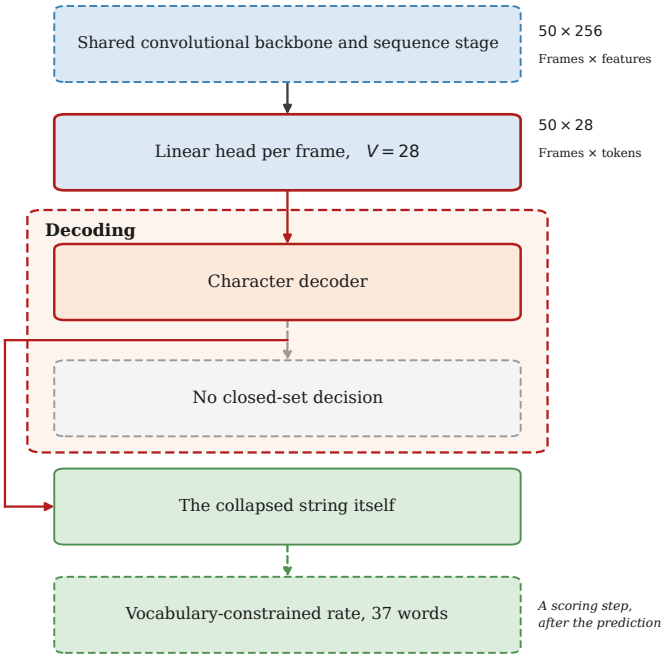


Figure 5.1: The continuous recognition pipeline. The stages preceding the head remain unchanged from Figure 3.3. The head uses a fixed 28-character alphabet. The system skips the closed-set steps.

In the classification pipeline, the per-frame predictions undergo two compression steps before scoring: the CTC collapse forms a string, and the lexicon step maps it to a predefined class. The evaluation employs the same two protocols from Section 3.7: global five-fold stratified cross-validation per participant and inter-session leave-one-session-out validation. Within these frameworks, the pipeline trains subject-specific models for each participant and speaking condition. It handles the rest condition as the classification approach by downsampling it during training and evaluating it fully in the test folds, thereby maintaining a reference set of twenty-one unique texts. Finally, to guarantee that any performance variations can be attributed exclusively to the tested sequence model, all optimization parameters, including the learning rate, warm-up, plateau scheduler, and early-stopping patience, match those in Section 3.7.

5.4.2 Training Schedule and Class Balancing

Checkpoint selection is the only difference with respect to the training strategy of Section 3.7. The error rate is recomputed on the validation split at the end of every epoch. Loss and error rate are not monotonically related: the CTC objective can keep minimizing the negative log-likelihood after the collapsed string has stopped improving.

5.4.3 Evaluation Metrics

The pipeline evaluates predictions using the WER which computes the word-level Levenshtein distance [83] between each prediction i and its corresponding reference

$$\text{WER} \;=\; \frac{\sum_i \Big(S_i + D_i + I_i\Big)}{\sum_i N_{\text{ref}}^{(i)}}, \tag{5.1}$$

where the components for each prediction i are:

- S_i: the number of substitutions;
- D_i: the number of deletions;
- I_i: the number of insertions;
- $N_{\text{ref}}^{(i)}$: the word count of the reference.

By summing these terms across the entire evaluation set, the metric divides the total errors by the overall number of reference words [94]. Three properties define this approach. First, a four-word sentence carries four times the weight of a one-word sentence. Second, it lacks an upper bound, because the reference length does not limit insertions. Third, a word registers as correct only upon a perfect string match, meaning that a single-character mismatch (*lefr* instead of *left*) implies a substitution penalty. Applying this same logic at the character level yields the CER, which penalizes a typo with only a single error instead. Moreover, model selection relies on WER for the sentence corpus and on CER for word-level tasks. To correct for class imbalance, the error rate is also computed independently within each group of data. This grouping is created by merging the distinct reference texts, and the group-level metrics are then averaged with equal weight

$$\text{Balanced WER} \;=\; \frac{1}{|U|} \sum_{u \in U} \text{WER}_u, \tag{5.2}$$

where:

- U: the set of unique reference texts;
- WER_u: the WER computed over the subset of predictions whose reference text is u.

By treating the rest condition as a single, equally weighted group among the unique references, the balanced formulation prevents long sentences from having more impact than shorter ones. A third metric addresses a limitation of the open-vocabulary WER: it heavily penalizes minor errors on a closed corpus. Every

predicted word is mapped to the vocabulary word with the minimum character edit distance. This way, the mapping always replaces an out-of-vocabulary word, regardless of the edit distance magnitude. Recomputing the WER on these adjusted predictions yields the vocabulary-constrained rate, which also features a balanced counterpart.

Chapter 6

Sentence Continuous Recognition: Results

This chapter reports the results of the continuous recognition task performed using the same models, windows, data splits, and training procedure as Chapter 4. The difference is that the decoded string is evaluated directly, without mapping it to a predefined lexicon after the CTC collapse. Unless stated otherwise, all mentioned error rates are vocabulary-constrained and balanced across the twenty-one reference texts, giving each text equal weight. In addition, the unconstrained rate is also reported. Decoding is greedy and uses no language model, except in Section 6.2, which measures the decoder itself.

6.1 Continuous Recognition Performance

Table 6.1 summarizes the continuous recognition results for the recurrent and the Transformer encoder models, the global and inter-session protocols, and both vocalized and speaking conditions, using 2.0 s windows on the sentence dataset. The convolutional backbone without a sequence stage is included as a reference, while the MFCC representation with the Transformer provides the continuous recognition counterpart of the analysis in Section 4.4.2 but the time-domain representation is not considered, as it was selected based only on the classification results. The best configuration, the recurrent stage on the STFT representation under greedy decoding, achieves a vocabulary-constrained balanced WER of 6.9 ± 4.5% for vocalized speech and 12.2 ± 8.5% for silent speech under global cross-validation. Under the inter-session protocol, these values increase to 16.3 ± 5.0% and 22.4 ± 14.1%, respectively. Without a sequence stage, performance decreases with vocabulary-constrained balanced WER ranging from 77.2% to 86.6%. Thus, most of the

transcription capability is provided by the sequence stage, which contains approximately 660 000 parameters. The unconstrained results also reveal errors that are hidden by the vocabulary constraint. For the best configuration, the balanced WER increases from 6.9% to 14.2 ± 7.4% for vocalized speech and from 12.2% to 20.5 ± 11.5% for silent speech under global cross-validation. Under inter-session evaluation, the corresponding increases are from 16.3% to 25.8 ± 6.7% and from 22.4% to 31.8 ± 16.6%. A substantial part of this difference comes from near-miss spellings that are mapped to the correct vocabulary entry. The counts reported below are read from the greedy decodes of the Transformer in the inter-session silent condition. The most frequent example is *rest*: 1 367 of its 8 400 occurrences are decoded incorrectly, with *rst*, *reest*, *ret*, *rt*, *r est*, and *re st* accounting for 491 cases. Among the sentence references, doubled letters are a frequent source of errors. For example, *mission* is often decoded as *mision*, accounting for 59 errors in *start the mission* and 58 in *stop the mission*. Similar errors occur for *proceed* → *proced*, *wall* → *wal*, and *off* → *of*. These errors are consistent with greedy CTC decoding because consecutive identical characters must be separated by a blank to survive the collapse, and the most probable path does not always provide this separation. Prefix beam search can recover some of these cases, as discussed in Section 6.2. After spelling errors are excluded, the remaining errors mainly reflect the same confusions observed in the closed-set task (Section C.3), such as *turn right* versus *turn left* and *start the mission* versus *stop the mission*. In the same setting, these two pairs account for 49 errors, compared with 117 errors caused by doubled-letter collapses in the two mission sentences alone. The recurrent block outperforms the Transformer in three of the four settings. Under global cross-validation, it achieves 6.9 ± 4.5% versus 8.3 ± 5.4% for vocalized speech and 12.2 ± 8.5% versus 13.0 ± 8.9% for silent speech. Under inter-session evaluation, the corresponding values are 16.3 ± 5.0% versus 18.6 ± 6.4% for vocalized speech, while the Transformer performs slightly better for silent speech, at 21.2 ± 12.6% versus 22.4 ± 14.1%. The advantage of the recurrent model is therefore consistent in three settings but small in the silent inter-session condition. Its comparison with the Transformer at matched capacity is discussed in Section 6.3. The MFCC representation also performs worse than STFT. At the Transformer stage, the MFCC representation increases the balanced WER by 6.5 and 5.3 points under global cross-validation, from 8.3 ± 5.4% to 14.8 ± 10.6% for vocalized speech and from 13.0 ± 8.9% to 18.3 ± 12.2% for silent speech. Under inter-session evaluation, the increase is 3.7 and 5.8 points, from 18.6 ± 6.4% to 22.3 ± 11.7% and from 21.2 ± 12.6% to 27.0 ± 14.7%, respectively. The same trend is observed with the recurrent model, where replacing STFT with MFCC increases the error by 13.7-16.8 points. Thus, the loss of spectral information introduced by the cepstral representation affects the model's ability to distinguish between the predefined words.

Table 6.1: Sentence continuous recognition balanced vocabulary-constrained WER (%). All configurations use 2.0 s windows, the 21 references and the rest condition, both with vocabulary-constrained rate and unconstrained score. Values denote the mean ± one standard deviation across the seven participants. Results are provided for both the global and inter-session protocols, with the lowest error rate in each column highlighted in bold.

			Global evaluation		**Inter-session evaluation**	
Domain	**Seq. stage**	**Dec.**	**Voc**	**Sil**	**Voc**	**Sil**
Constrained balanced WER (%),						
MFCC 15/10	BiLSTM	Greedy	23.5 ± 14.4	29.0 ± 17.2	31.7 ± 15.9	36.1 ± 17.7
MFCC 15/10	Transformer	Greedy	14.8 ± 10.6	18.3 ± 12.2	22.3 ± 11.7	27.0 ± 14.7
STFT	none	Greedy	77.2 ± 8.8	83.8 ± 7.1	83.6 ± 6.0	86.6 ± 6.1
STFT	BiLSTM	Greedy	6.9 ± 4.5	12.2 ± 8.5	16.3 ± 5.0	22.4 ± 14.1
STFT	BiLSTM	Beam	**6.6** ± 4.1	**11.5** ± 7.9	**15.9** ± 4.9	21.5 ± 13.7
STFT	Transformer	Greedy	8.3 ± 5.4	13.0 ± 8.9	18.6 ± 6.4	21.2 ± 12.6
STFT	Transformer	Beam	8.2 ± 5.2	12.7 ± 8.7	18.4 ± 6.3	**21.0** ± 12.4
Unconstrained balanced WER (%),						
MFCC 15/10	BiLSTM	Greedy	40.2 ± 16.8	44.1 ± 16.9	45.9 ± 17.7	49.6 ± 18.4
MFCC 15/10	Transformer	Greedy	24.7 ± 15.2	31.2 ± 16.3	34.0 ± 15.3	39.6 ± 16.4
STFT	none	Greedy	97.6 ± 4.1	101.1 ± 3.3	100.8 ± 2.3	102.4 ± 2.9
STFT	BiLSTM	Greedy	14.2 ± 7.4	20.5 ± 11.5	25.8 ± 6.7	31.8 ± 16.6
STFT	BiLSTM	Beam	9.8 ± 5.8	15.2 ± 9.6	21.4 ± 5.8	27.2 ± 15.3
STFT	Transformer	Greedy	17.7 ± 10.8	25.3 ± 13.7	32.0 ± 9.5	35.5 ± 16.8
STFT	Transformer	Beam	16.0 ± 9.7	23.6 ± 13.0	31.0 ± 9.0	34.3 ± 16.0

6.2 Decoder Sweep: Greedy Decoding versus Prefix Beam Search

The decoder grid of Section 4.6 is evaluated in this section using the continuous recognition error rate. The analysis is performed on the saved logits produced by the models of Section 6.1, so no network is retrained (Table 6.2). Overall, the choice of decoder has a limited effect on WER. The tuned beam reduces the error from 9.5% to 9.0% for the recurrent model under global cross-validation and from 19.3% to 18.6% under the inter-session protocol. For the Transformer, the corresponding reductions are from 10.6% to 10.4% and from 19.9% to 19.6%. Although the recurrent model benefits slightly more from beam search, the performance gap between the two architectures remains unchanged. This confirms that the observed ranking is mainly determined by the model outputs rather than by the decoding strategy. Among the three additional scoring terms, temperature has the more effect than the others. The figures of this paragraph are unconstrained. Increasing T_θ from 1 to 2, with the other parameters at their default values, reduces the balanced WER of the Transformer from 20.7% to 19.8% and that of the recurrent

model from 14.7% to 13.7%, with the latter reaching a minimum of 12.5% at $T_\theta = 1.6$. In contrast, increasing the blank penalty worsens performance, with the Transformer reaching 23.8% at a penalty of four. The length bonus has little effect at 0.5 and becomes counterproductive at two, where the error reaches 24.9%. The effect of temperature is consistent with the decoding errors described in Section 6.1: flattening the frame-level distributions increases the contribution of alternative CTC alignments, which can recover doubled characters that greedy decoding misses. The blank penalty has the opposite effect by reducing the probability of the blank required between repeated characters. No configuration produces an empty decode. Expanding the beam width beyond a value of 5 provides no meaningful benefit, as widths of 5, 10, and 25 differ by at most 0.1 points of balanced WER. On the other hand, restricting the width to 1 increases the error by 2.9 points for the recurrent model and 0.7 points for the Transformer. Eventually, while prefix beam search improves in respect with greedy decoding, its impact is secondary to the choice of model architecture and input representation.

Table 6.2: Balanced WER (%) of decoding strategies on the recognition task. Results are split by vocabulary-constrained ("Constr.") and unconstrained ("Unconstr.") rates across global and inter-session protocols. All models use STFT, 2.0 s windows. Hyperparameters: "plain beam" uses $w = 10$; "tuned beam" uses optimal settings $w = 5$, $T_\theta = 1.3$, $\beta = 1$ for BiLSTM; $w = 10$, with $T_\theta = 2.0$ for global CV and $T_\theta = 1.6$ for inter-session for the Transformer. Best results per model are in bold.

		Global evaluation		**Inter-session evaluation**	
Model	**Decoder**	**Constr.**	**Unconstr.**	**Constr.**	**Unconstr.**
BiLSTM	Greedy	9.5	17.3	19.3	28.8
BiLSTM	Plain beam, $w = 10$	9.2	14.7	18.9	26.5
BiLSTM	Tuned beam	**9.0**	**13.7**	**18.6**	**25.0**
Transformer	Greedy	10.6	21.5	19.9	33.8
Transformer	Plain beam, $w = 10$	10.5	20.7	19.8	33.1
Transformer	Tuned beam	**10.4**	**19.8**	**19.6**	**32.6**

6.3 BiLSTM versus Transformer at Equal Parameters

The two sequence models are evaluated at equal number of parameters on the continuous recognition task. Under greedy decoding, the BiLSTM outperforms the Transformer encoder in three of the four experimental settings. It achieves a vocabulary-constrained balanced WER of 6.9 ± 4.5% and 12.2 ± 8.5% under

global cross-validation versus the encoder's 8.3% and 13.0%, and leads 16.3 ± 5.0% against 18.6 ± 6.4% in the inter-session vocalized condition. The encoder recovers only the silent setting between sessions, scoring 21.2 ± 12.6% against 22.4 ± 14.1%. Yet, the recurrent block outperforms the encoder in all four settings on the balanced CER. This discrepancy reveals that the two architectures distribute errors differently: the encoder spreads fewer character errors across more words, while the recurrent block concentrates more character errors on fewer words. At the WER-optimal configuration, the recurrent block maintains a lead of over 7 points on the unconstrained rate. Consequently, the BiLSTM architecture performs better in generating unconstrained text on this corpus. For what concerns performance spread among participants, standard deviations range from 4.1 to 17.7 points of balanced WER. Participants S02, S03, and S06 achieve error rates around 2% in the global vocalized setting, whereas S04 and S05 are near 10% and between 24.7% and 43.9% in the silent conditions. Participant S07 inverts the standard trend, with 14.4% when vocalizing but improving to 4.7% during silent articulation.

Chapter 7

Conclusion

This thesis demonstrates sentence-level silent speech recognition using a wearable, fully dry surface EMG neckband, operating within the computational limits of an edge device. The work offers four primary contributions, with all reported margins of error (±) representing one standard deviation across the seven participants. First, we introduce a multimodal corpus with 29 400 utterances and nearly 39 h 40 min of recording. To capture natural sensor variability, devices were removed and repositioned between the six sessions. Each session contains repetitions of isolated words and twenty sentences, articulated both aloud and silently, simultaneously recorded via an EMG neckband, an EEG headband, and synchronized microphones. Hardware characterization guarantees synchronization, bounding the stream-to-microphone delay to 7.7 ms and the inter-device offset to 0.12 ms. Second, we present an empirically validated decoding pipeline. We show that the CTC objective outperforms CE for sentence-level tasks. For input representation, the STFT power spectrogram exceeds both the raw time signal and the MFCC, since the cepstral transform discards EMG discriminative evidence. Finally, introducing a sequence modeling stage yields the largest single performance gain, improving accuracy by up to 64 points over a convolutional baseline. Third, we compare recurrent and attention-based sequence models at equal parameter number. The optimal architecture depends on the evaluation task. The Transformer dominates closed-set classification, achieving 94.2 ± 3.9% (vocalized) and 90.7 ± 6.1% (silent) balanced accuracy under global cross-validation, and 85.8 ± 5.2% and 82.1 ± 10.5% under the inter-session protocol. However, for continuous transcription, the BiLSTM performs better under global cross-validation, reaching a vocabulary-constrained balanced WER of 6.9 ± 4.5% (vocalized) and 12.2 ± 8.5% (silent), and in the vocalized inter-session setting, at 16.3 ± 5.0%. The Transformer is ahead in the silent inter-session setting, at 21.2 ± 12.6% against 22.4 ± 14.1%. In addition, decoding uses no language model. Because the models distribute errors differently, the choice of architecture must align with the intended application. Fourth, we

design a causal onset detector to identify speech boundaries from the EMG stream which is a necessity for a continuous, trigger-free system. Using a fixed set of hyperparameters across all participants and conditions, it achieves 97.3% recall and 93.6% specificity, generating a false positive only 0.70 times per minute.

These results are subject to four limitations. First, the closed corpus limits vocabulary size; thus, our recognition metrics evaluate the unconstrained decoding of a closed set and cannot be directly compared to open-vocabulary literature. Second, the models are subject-specific, and inter-subject variability exceeds any architectural difference. Third, the offline evaluation utilizes cue-anchored windows, providing an upper bound for true streaming performance. Finally, while the network sizes fit the target wearable budget, they have not yet been benchmarked on the physical edge accelerator.

Future work directly addresses these limitations. Incorporating a language model during decoding will mitigate residual errors, as many remaining mistakes involve single-word substitutions within otherwise correct sentences. Deploying the pipeline requires quantizing the models to eight bits and executing them on the target ultra-low-power RISC-V system to measure true latency and energy consumption. Furthermore, the onset detector must be integrated online, training the models directly on detected boundaries rather than predefined visual cues. Data labels can also be refined: the synchronized microphone audio serves as a precise ground truth for speech boundaries during vocalized sessions, and its timing can be transferred to the paired silent recordings with 7.7 ms accuracy. Additionally, the parallel EEG serves as source signal for sensor fusion, particularly to improve inter-session generalization where cortical signals may resist the electrode displacement issues affecting the EMG.

In conclusion, this thesis proves that entire sentences and not just isolated commands, can be reliably decoded from a fully dry fourteen-channel neckband. It achieves a practical operating point while respecting the constraints of edge hardware. Alongside the multimodal dataset, this work establishes an empirically validated framework to guide the next generation of wearable silent speech interfaces.

Appendix A

Data-Collection Protocol

This appendix presents the complete data-collection protocol. It details the corpus, the preliminary lexical analysis, trial and session timing, randomization and balancing criteria, expected dataset size, and scheduling rules. While Section 2.3 outlines the theoretical rationale behind these choices, the following text acts as the operational manual for the operator. Predefined targets are indicated as nominal quantities, whereas the actual dimensions of the collected dataset are reported in Section 2.3.2.

A.1 Corpus

A.1.1 Phase A: Words

Phase A composes of fifteen words divided into two subsets (Table A.1). Group A contains eight HMI keywords that correspond to the SilentWear command vocabulary [4], each of which appears in at least one of the Phase B sentences detailed in Section A.1.2. In contrast, Group B consists of seven additional words that do not appear in any of the Phase B sentences.

A.1.2 Phase B: Sentences

Table A.2 presents the twenty sentences of Phase B with their class identifiers, grouped by function. The "Keyword" column identifies sentences containing any of the eight Phase A Group A commands, while the "Source" column details their origin. Items labeled as *Generated* were authored for this study. The remaining sentences are drawn from the specified corpora; the *adapted* tag indicates modifications applied to satisfy the length and domain constraints outlined in Section 2.3.1.

Table A.1: The fifteen words of Phase A. Group A words are HMI keywords that also occur inside Phase B sentences; Group B words are synonyms that occur in no Phase B sentence. "Syl." is the number of syllables. The "#" column is the class identifier written into the recordings, rest being class 0 (Section 2.2).

#	Word	Syl.	Context
Group A: HMI keywords, present in Phase B			
1	Up	1	Sentence 13, *Pick up the object*; spatial particle
2	Down	1	Sentence 14, *Put the object down*; spatial particle
3	Left	1	Sentence 2, *Turn left*; directional
4	Right	1	Sentence 3, *Turn right*, directional; also sentence 18, *Stop right there*, as an adverb
5	Forward	2	Sentence 1, *Move forward*; directional
6	Backward	2	Sentence 8, *Move backward*; directional
7	Start	1	Sentences 19 and 20; control
8	Stop	1	Sentences 4 and 18; control
Group B: interchangeable synonyms, absent from Phase B			
9	Advance	2	Synonym of *forward* / *go* / *proceed*
10	Reverse	2	Synonym of *backward*
11	Rotate	2	Synonym of *turn*
12	Halt	1	Synonym of *stop*
13	Begin	2	Synonym of *start*
14	Grab	1	Synonym of *pick up*
15	Place	1	Synonym of *put down*

The sentence set comprises 5 two-word, 6 three-word, and 9 four-word sentences, yielding a mean length of 3.2 words. At the character level, the 20 sentences span 22 distinct letters in addition to the space character. Letters q, x, y, and z are omitted.

A.2 Lexical Analysis

Table A.3 details the frequency of every word type across the twenty sentences. Thirteen words appear in multiple sentences, whereas twenty-three occur once. Three sentences (*Turn right*, *Start the mission*, *Stop the mission*) contain no unique words. Furthermore, seven sentence pairs differ by one word position: 1–6, 1–8, 2–3, 4–20, 5–7, 6–8 and 19–20.

Table A.2: The 20 sentences of Phase B, with word count, HMI keyword annotation and provenance. The "#" column is the class identifier written into the recordings, rest being class 0 (Section 2.2); the rows are grouped by function, so the identifiers are not contiguous within a group.

#	Sentence	Words	Keyword	Source
Navigation				
1	Move forward	2	Forward	SayTap (adapted) [57]
2	Turn left	2	Left	AVE Speech [50]
3	Turn right	2	Right	AVE Speech [50]
5	Go to the center	4	—	SCOUT (adapted) [55]
6	Move inside	2	—	Generated
7	Go to the wall	4	—	SCOUT [55]
8	Move backward	2	Backward	Generated
9	Proceed to the door	4	—	SCOUT (adapted) [55]
10	Back to initial position	4	—	SCOUT (adapted) [55]
11	Enter the room	3	—	SCOUT [55]
12	Proceed two meters	3	—	SCOUT (adapted) [55]
Object and assistive interaction				
13	Pick up the object	4	Up	BridgeData V2 (adapted) [58]
14	Put the object down	4	Down	BridgeData V2 (adapted) [58]
15	Turn on the light	4	—	AVE Speech [50]
16	Switch off the light	4	—	Generated
17	Turn off the volume	4	—	Generated
Control commands				
4	Stop the mission	3	Stop	Generated
18	Stop right there	3	Stop	Generated
19	Start the task	3	Start	Generated
20	Start the mission	3	Start	Generated

Table A.3: Content-word frequency across the Phase B sentences. The "Count" column denotes the number of occurrences, "Sentences" lists the sentences containing the word, and "Note" indicates its contextual part of speech.

Word	Count	Sentences	Note
Types occurring in more than one sentence			
the	12	4, 5, 7, 9, 11, 13, 14, 15, 16, 17, 19, 20	Definite article
turn	4	2, 3, 15, 17	Verb, imperative
to	4	5, 7, 9, 10	Preposition, directional
go	2	5, 7	Verb, imperative
move	3	1, 6, 8	Verb, imperative
light	2	15, 16	Noun, common
mission	2	4, 20	Noun, common
object	2	13, 14	Noun, common
off	2	16, 17	Adverbial particle

Table A.3: Content-word frequency across the Phase B sentences. (continued)

Word	Count	Sentences	Note
proceed	2	9, 12	Verb, imperative
right	2	3, 18	Directional adverb in 3, intensifying adverb in 18; semantic ambiguity
start	2	19, 20	Verb, imperative; Phase A Group A keyword
stop	2	4, 18	Verb, imperative; Phase A Group A keyword
Types occurring exactly once			
back	1	10	Directional adverb
backward	1	8	Directional adverb; Phase A Group A keyword
center	1	5	Noun, common
door	1	9	Noun, common
down	1	14	Adverbial particle; Phase A Group A keyword
enter	1	11	Verb, imperative
forward	1	1	Directional adverb; Phase A Group A keyword
initial	1	10	Adjective, attributive
inside	1	6	Directional adverb
left	1	2	Directional adverb; Phase A Group A keyword
meters	1	12	Noun, common, plural
on	1	15	Adverbial particle
pick	1	13	Verb, imperative
position	1	10	Noun, common
put	1	14	Verb, imperative
room	1	11	Noun, common
switch	1	16	Verb, imperative
task	1	19	Noun, common
there	1	18	Locative adverb
two	1	12	Cardinal numeral
up	1	13	Adverbial particle; Phase A Group A keyword
volume	1	17	Noun, common
wall	1	7	Noun, common

A.3 Trigger Module Configuration

The trigger module of BioGUI (Section 2.2) is configured through a JSON file defining the labels and the trial timing. The two configurations, encode the 15 words of Phase A and the 20 sentences of Phase B, respectively. Their parameters appear in Table A.4 and implement the trial structure detailed in Section A.4. At the beginning of the acquisition, a list incorporates `nReps` repetitions of each

Table A.4: Trigger configurations for the two acquisition phases. Durations are expressed in milliseconds, following the JSON specification.

Parameter	Phase A	Phase B
`triggers`	15	20
`nReps`	10	10
`durationStart`	5 000	5 000
`durationTrigger`	1 500	2 500
`durationRest`	1 500	1 500
`shuffle`	true	true

label and is subsequently shuffled. A viewer window displays a START screen for `durationStart` milliseconds, after which item and rest screens alternate. An item screen presents the label as centered text; a rest screen displays a countdown alongside the upcoming label, allowing the participant to read the item before articulation. Class identifiers follow the declaration order of the labels in the JSON file. The rest condition is encoded as zero. The processing pipeline adopts this mapping, with the corresponding texts listed in Section 2.3.1 and Section A.1. The state identifier resets to zero at the beginning of every rest interval.

The stop marker. The string `last_stop` is appended as the final element of the shuffled list. When the sequence reaches this item, the module displays the word `stop` for `durationTrigger` milliseconds and broadcasts the rest value, 0, to all streaming controllers. Following this interval, the list concludes. The module closes the viewer, halts the timers. This marker signals the conclusion of the block to the participant and terminates the trigger generator without operator intervention.

A.4 Trial and Session Timing

Table A.5 gives the timing parameters of both phases together with the configuration key that sets each of them. The utterance window is longer for sentences than for words, since a multi-word utterance takes longer to articulate; the ITI is identical in both phases, so the trial structure differs in one parameter only. Table A.6 details the within-session block structure. The devices are repositioned by the operator between sessions only; within a session, and in particular across the rests and the phase break, it remains fixed.

A.5 Shuffling and Balance Rules

1. **Repetitions per block.** Each item appears ten times per block, interleaved throughout the recording rather than presented in consecutive sequences. This

Table A.5: Timing parameters of the acquisition protocol, per phase.

Parameter	Words (Phase A)	Sentences (Phase B)
Utterance window	1.5 s	2.5 s
ITI	1.5 s	1.5 s
Time per trial	3.0 s	4.0 s
Items per block	15	20
Repetitions per item	10	10
Utterances per block	150	200
Block duration	450 s (7 min 30 s)	800 s (13 min 20 s)
Blocks per session	2	2
Inter-block rest	300 s (5 min)	300 s (5 min)
Phase duration	1 200 s (20 min)	1 900 s (31 min 40 s)
Phase break from A to B	300 s (5 min)	
Recorded time / session	3 400 s (56 min 40 s)	
Slot length / session	≈ 67 min	

Table A.6: Within-session block structure. Each session contains four recorded blocks, two per phase, alternating the vocalized and silent conditions.

Block	Condition	Items	Reps	Duration
Phase A, words				
W1	Vocalized	15 words	10	7 min 30 s
—	Rest	—	—	5 min
W2	Silent	15 words	10	7 min 30 s
—	Phase break	—	—	5 min
Phase B, sentences				
S1	Vocalized	20 sentences	10	13 min 20 s
—	Rest	—	—	5 min
S2	Silent	20 sentences	10	13 min 20 s
Session total: 56 min 40 s recorded, ≈ 67 min per subject slot				

distribution ensures that intra-block fatigue and electrode drift affect all classes uniformly.

2. **Shuffle independence.** A pseudo-random presentation order is generated independently for every block, session, and participant, constrained such that no two consecutive blocks share the same sequence.

3. **Vocalized-silent symmetry.** Across the complete set of sessions, every item yields an identical number of vocalized and silent repetitions. This ensures that performance comparisons between the two speaking conditions remain unaffected by imbalances in data volume.

A.6 Nominal Dataset Size

Table A.7 summarizes the expected dataset size. These nominal values are derived from the protocol parameters, assuming all seven participants complete their six assigned sessions.

Table A.7: Nominal dataset yield, assuming seven participants complete six recording sessions each.

Content	Condition	Sessions	Reps / block	Per session	Total
Words (15 items)	Vocalized	6	10	150	6 300
Words (15 items)	Silent	6	10	150	6 300
Sentences (20 items)	Vocalized	6	10	200	8 400
Sentences (20 items)	Silent	6	10	200	8 400
Total	—	—	—	700	29 400

In recording time this amounts to 56 min 40 s per session, roughly 5 h 40 min per subject over six sessions, and approximately 39 h 40 min over seven subjects.

A.7 Scheduling Rules and Realized Schedule

Assuming a slot duration of approximately 67 minutes and a capacity of four slots per day, the 42 sessions across seven participants require a minimum of eleven recording days. This total extends to twelve days under a round-robin schedule, where all participants must complete session s before anyone advances to session $s + 1$. This calculation serves as a logistical estimate rather than a fixed calendar, and was used primarily to allocate operator time and laboratory availability. The scheduling constraints enforced during data collection are defined below without reference to absolute dates, accommodating variable participant availability. These rules constitute the criteria determining the schedule.

1. **One session per subject per day.** No subject undergoes more than one recording session per day. Consequently, no two sessions for a given subject share the same sensor application or baseline skin condition.

2. **Minimum 24-hour interval.** A minimum of 24 hours must elapse between consecutive sessions for the same subject. This ensures that every session boundary encompasses a full sleep cycle, introducing variations in skin impedance and overall physiological state.

3. **Device removal and repositioning.** The devices are removed at the end of every session and repositioned by the operator at the start of the next. Since the hardware is never moved during a recording, electrode placement remains constant within a session but varies between sessions.

4. **A fixed acquisition chain.** The exact same neckband, headband, BioGAP-Ultra unit [16], electrode type, sampling rate, gain, and BioGUI configuration [47] are used for all sessions of all subjects. This ensures that no hardware or software changes can influence the data across sessions.

5. **Fixed within-session order.** Phase A always precedes Phase B. Within each phase, the vocalized block always comes before the silent one. Keeping this order strictly constant across all sessions and subjects ensures that the sequence of tasks does not act as a hidden variable.

Although the realized schedule satisfies all five invariants, it deviates from the plan in two respects. First, sessions for individual participants were not recorded on consecutive days; instead, intervals between successive sessions ranged from a single day to several weeks, depending on availability. Second, participant enrollment was not parallelized. The recording campaign spanned several months, from May to July 2026, resulting in disjoint recording time spans across different subjects.

Appendix B

Word and Sentence Lexicon

The CTC branch processes the reference text of each class from a lexicon file, allocating one file per task and one line per class. Each line contains the text, its character-level tokenization, and the word-boundary token `|`:

```
rest r e s t |
move forward m o v e f o r w a r d |
back to initial position b a c k t o i n i t i a l p o s i t i o n |
```

The line order defines the class indices used throughout this thesis with the rest condition occupying index 0, followed by the items sorted by their trigger identifier. The word file comprises rest and the fifteen Phase A words following the sequence presented in Table A.1. The sentence file comprises rest and the twenty Phase B sentences listed in Table B.1; Table A.2 groups the sentences by function, yielding non-consecutive identifiers, and therefore does not reflect the internal file structure.

Table B.1: The entries of the sentence lexicon, sorted by file sequence. The index corresponds to the cue identifier from Table A.2.

#	**Text**	#	**Text**
0	rest	11	enter the room
1	move forward	12	proceed two meters
2	turn left	13	pick up the object
3	turn right	14	put the object down
4	stop the mission	15	turn on the light
5	go to the center	16	switch off the light
6	move inside	17	turn off the volume
7	go to the wall	18	stop right there
8	move backward	19	start the task
9	proceed to the door	20	start the mission
10	back to initial position		

Appendix C

Per-Participant Results

This section details the individual per-participant results and each task is reported at its optimal configuration and evaluated across global and inter-session protocol.

C.1 Sentence Closed-Set Classification, per Participant

Table C.1 reports the best closed-set configuration of the thesis, the Transformer encoder on the STFT representation with the CTC objective, under greedy decoding and the lexicon constraint.

Table C.1: Per-participant balanced accuracy (%) of the sentence classifier, Transformer sequence stage, STFT representation, 2.0 s windows, CTC objective, 21 classes. Each value is the mean over the folds of the protocol, and the **All** column is the mean over the seven participants, ± one standard deviation across them.

Cond.	**S01**	**S02**	**S03**	**S04**	**S05**	**S06**	**S07**	**All**
Inter-session protocol								
Voc	88.7	92.8	84.4	79.8	80.7	92.8	81.7	85.8 ± 5.2
Sil	82.5	78.8	86.7	67.6	68.6	94.0	96.6	82.1 ± 10.5
Global protocol								
Voc	95.9	98.9	96.2	91.3	89.5	98.8	89.0	94.2 ± 3.9
Sil	89.4	91.9	92.8	80.1	84.7	98.2	97.8	90.7 ± 6.1

C.2 Sentence Continuous Recognition, per Participant

Table C.2 reports the best recognition configuration of the thesis, the recurrent sequence stage on the STFT representation, under greedy decoding.

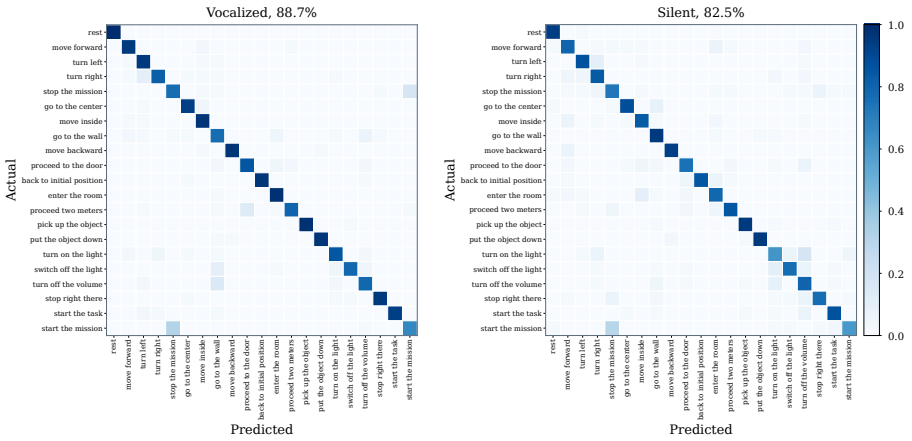
Table C.2: Per-participant balanced vocabulary-constrained WER (%) for continuous recognition, BiLSTM sequence stage, STFT representation, 2.0 s windows, 21 classes. Both blocks decode with greedy strategy. Each value is the mean over the folds of the protocol, and the **All** column is the mean over the seven participants, ± one standard deviation across them.

Cond.	**S01**	**S02**	**S03**	**S04**	**S05**	**S06**	**S07**	**All**
Inter-session protocol								
Voc	14.4	7.1	17.3	22.0	20.9	12.3	20.0	16.3 ± 5.0
Sil	18.5	24.6	11.7	41.9	43.9	6.8	9.0	22.4 ± 14.1
Global protocol								
Voc	6.7	1.9	2.9	10.3	10.0	2.0	14.4	6.9 ± 4.5
Sil	10.9	10.4	4.8	25.2	24.7	4.3	4.7	12.2 ± 8.5

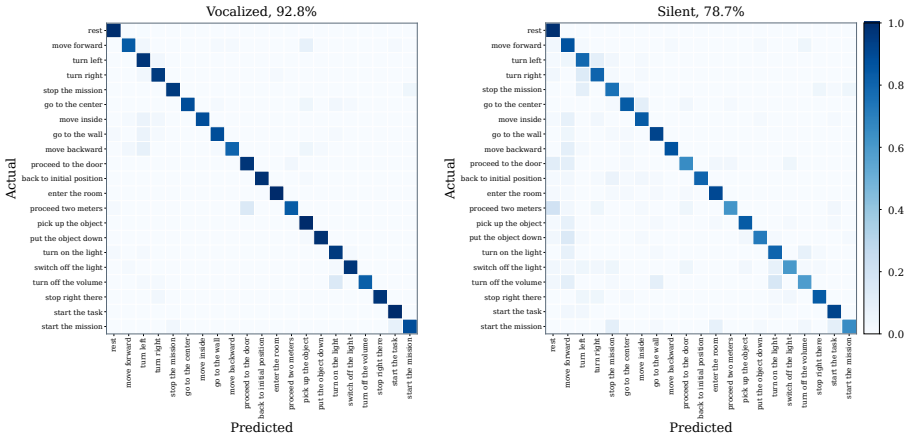
C.3 Per-Participant Confusion Matrices

Confusion matrices for each participant across vocalized and silent conditions with inter-session protocol. The setup consists of the Transformer encoder, STFT representation, 2.0 s windows, and 21 classes, with greedy decoding.

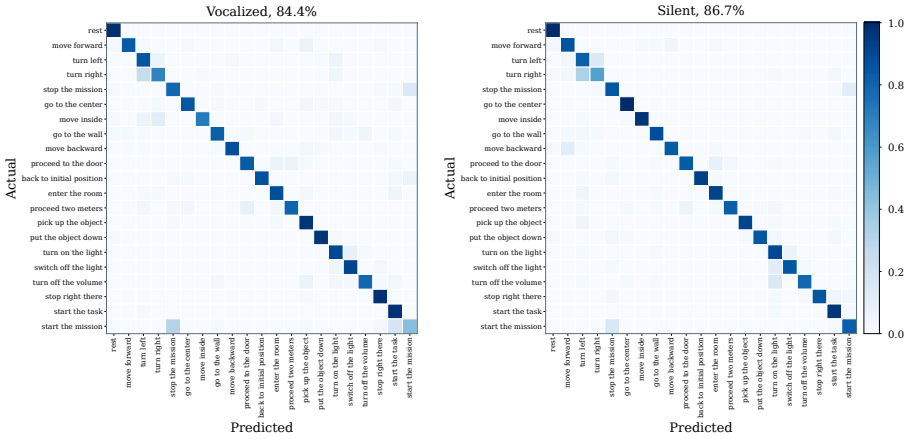
Participant S01



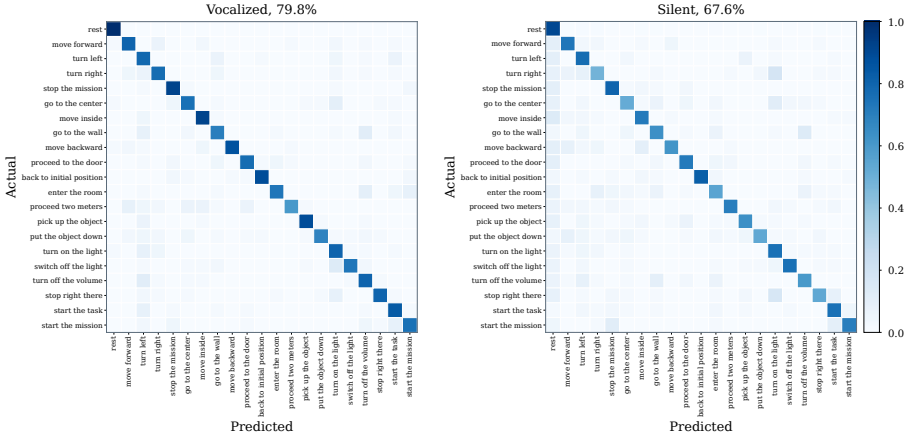
Participant S02



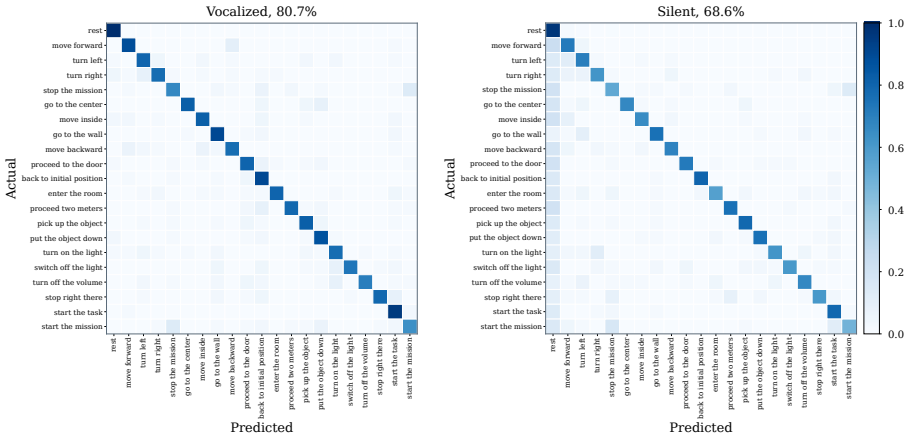
Participant S03



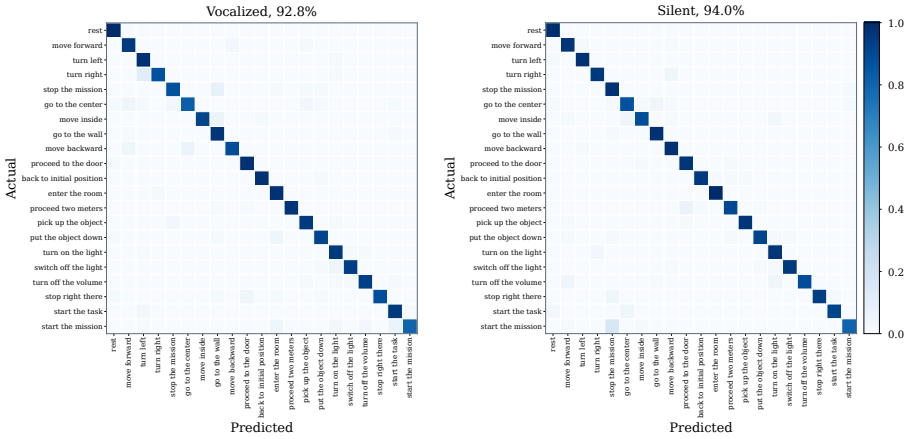
Participant S04



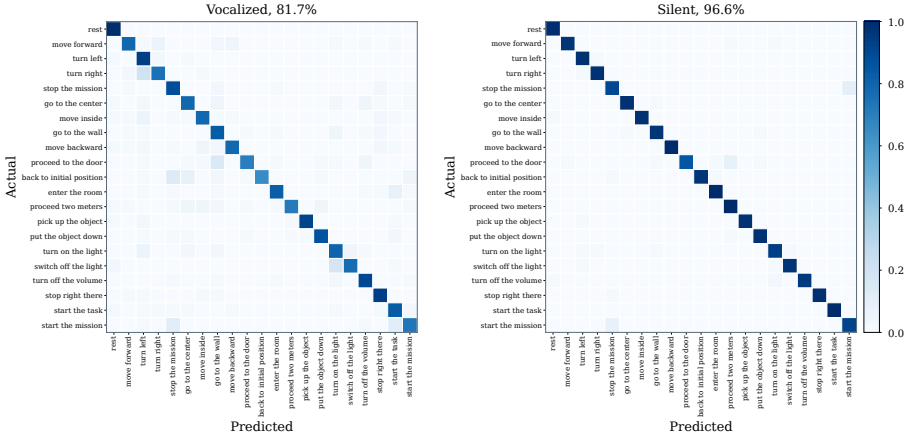
Participant S05



Participant S06



Participant S07



Bibliography

[1] Arnav Kapur, Shreyas Kapur, and Pattie Maes. «AlterEgo: A Personalized Wearable Silent Speech Interface». In: *Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI)*. 2018, pp. 43–53. DOI: 10.1145/3172944.3172977. URL: https://dl.acm.org/doi/10.1145/3172944.3172977 (cit. on pp. 1, 3–8).

[2] Adiba Tabassum Chowdhury, Mehrin Newaz, Purnata Saha, Mohannad Natheef AbuHaweeleh, Sara Mohsen, et al. «Decoding silent speech: a machine learning perspective on data, methods, and frameworks». In: *Neural Computing and Applications* 37 (2025), pp. 6995–7013. DOI: 10.1007/s00521-024-10456-z. URL: https://link.springer.com/10.1007/s00521-024-10456-z (cit. on pp. 1, 3, 9).

[3] Fiona Meier, Giusy Spacone, Sebastian Frey, Luca Benini, and Andrea Cossettini. *A Parallel Ultra-Low Power Silent Speech Interface based on a Wearable, Fully-dry EMG Neckband*. arXiv preprint arXiv:2509.21964. 2025. URL: http://arxiv.org/abs/2509.21964 (cit. on pp. 1, 2, 4–6, 12, 13).

[4] Giusy Spacone, Sebastian Frey, et al. *SilentWear: an Ultra-Low Power Wearable System for EMG-based Silent Speech Recognition*. arXiv preprint arXiv:2603.02847. 2026. URL: https://arxiv.org/abs/2603.02847 (cit. on pp. 1–3, 5–8, 12, 13, 18, 20, 24, 30, 33, 37–39, 43–45, 74, 91).

[5] Youhua Wang, Tianyi Tang, et al. «All-weather, natural silent speech recognition via machine-learning-assisted tattoo-like electronics». In: *npj Flexible Electronics* 5 (2021). DOI: 10.1038/s41528-021-00119-7. URL: https://www.nature.com/articles/s41528-021-00119-7 (cit. on pp. 1, 4–8).

[6] Chenyu Tang, Josée Mallah, Dominika Kazieczko, Wentian Yi, Tharun Reddy Kandukuri, et al. «Wireless Silent Speech Interface Using Multichannel Textile EMG Sensors Integrated Into Headphones». In: *IEEE Transactions on Instrumentation and Measurement* 74 (2025), pp. 1–10. DOI: 10.1109/TIM.2025.3583386. URL: https://ieeexplore.ieee.org/document/11052721/ (cit. on pp. 1, 3, 5–8).

[7] Qiang Cui, Xingyu Zhang, Yakun Zhang, Changyan Zheng, Liang Xie, et al. «A simplified adversarial architecture for cross-subject silent speech recognition using electromyography». In: *Journal of Neural Engineering* 21.5 (2024), p. 056001. DOI: 10.1088/1741-2552/ad7321. URL: https://iopscience.iop.org/article/10.1088/1741-2552/ad7321 (cit. on pp. 1, 5, 7).

[8] Huiyan Li et al. «Hybrid Silent Speech Interface Through Fusion of Electroencephalography and Electromyography». In: *Proc. Interspeech*. 2023. URL: https://www.isca-archive.org/interspeech_2023/li23l_interspeech.html (cit. on pp. 1, 3, 8–10).

[9] Jacob Tryon and Ana Luisa Trejos. «Evaluating Convolutional Neural Networks as a Method of EEG–EMG Fusion». In: *Frontiers in Neurorobotics* 15 (2021). DOI: 10.3389/fnbot.2021.692183. URL: https://www.frontiersin.org/articles/10.3389/fnbot.2021.692183/full (cit. on pp. 1, 3, 10).

[10] Nicholas Satterlee, Xiaowei Zuo, Kee Moon, Sung Q. Lee, Matthew Peterson, and John S. Kang. «Sentence-Level Silent Speech Recognition Using a Wearable EMG/EEG Sensor System with AI-Driven Sensor Fusion and Language Model». In: *Sensors* 25.19 (2025), p. 6168. DOI: 10.3390/s25196168. URL: https://pubmed.ncbi.nlm.nih.gov/41094990/ (cit. on pp. 1, 3, 9, 10, 20).

[11] Allen D. Hillel. «The study of laryngeal muscle activity in normal human subjects and in patients with laryngeal dystonia using multiple fine-wire electromyography». In: *The Laryngoscope* 111.S97 (2001), pp. 1–47. DOI: 10.1097/00005537-200104001-00001. URL: https://pubmed.ncbi.nlm.nih.gov/11464801/ (cit. on pp. 2, 18).

[12] Jordan J. Bird, Jhonatan Kobylarz, Diego R. Faria, Anikó Ekárt, and Eduardo P. Ribeiro. «Cross-Domain MLP and CNN Transfer Learning for Biological Signal Processing: EEG and EMG». In: *IEEE Access* 8 (2020), pp. 54789–54801. DOI: 10.1109/ACCESS.2020.2979074. URL: https://ieeexplore.ieee.org/document/9027853 (cit. on pp. 3, 8–10).

[13] Carlo J. De Luca. *Surface Electromyography: Detection and Recording*. DelSys Incorporated, technical tutorial. https://www.delsys.com/downloads/TUTORIAL/semg-detection-and-recording.pdf. 2002 (cit. on pp. 3, 4, 6).

[14] Tyler Benster, Guy Wilson, Reshef Elisha, Francis R. Willett, and Shaul Druckmann. *A Cross-Modal Approach to Silent Speech with LLM-Enhanced Recognition*. arXiv preprint arXiv:2403.05583. 2024. URL: http://arxiv.org/abs/2403.05583 (cit. on pp. 3, 6–9).

[15] Chenyu Tang, Liang Qi, Shuo Gao, Zibo Zhang, Wentian Yi, Muzi Xu, Edoardo Occhipinti, Yu Pan, and Luigi G. Occhipinti. «Sensing technologies for silent speech interfaces». In: *Nature Sensors* 1.1 (2026), pp. 16–26. DOI: 10.1038/s44328-025-00043-0 (cit. on pp. 3, 6).

[16] Sebastian Frey, Giusy Spacone, Andrea Cossettini, Marco Guermandi, Philipp Schilk, Luca Benini, and Victor Kartsch. «BioGAP-Ultra: A Modular Edge-AI Platform for Wearable Multimodal Biosignal Acquisition and Processing». In: *IEEE Transactions on Biomedical Circuits and Systems* (2026), pp. 1–17. DOI: 10.1109/TBCAS.2026.3652501. URL: https://ieeexplore.ieee.org/document/11346484 (cit. on pp. 3, 6, 7, 12, 13, 17, 81).

[17] Dario Farina, Roberto Merletti, and Roger M. Enoka. «The extraction of neural strategies from the surface EMG». In: *Journal of Applied Physiology* 96.4 (2004), pp. 1486–1495. DOI: 10.1152/japplphysiol.01070.2003. URL: https://journals.physiology.org/doi/full/10.1152/japplphysiol.01070.2003 (cit. on p. 4).

[18] Roberto Merletti and Silvia Muceli. «Tutorial. Surface EMG detection in space and time: Best practices». In: *Journal of Electromyography and Kinesiology* 49 (2019), p. 102363. DOI: 10.1016/j.jelekin.2019.102363. URL: https://www.sciencedirect.com/science/article/pii/S1050641119302536 (cit. on pp. 4, 40).

[19] Roberto Merletti and Giacinto Luigi Cerone. «Tutorial. Surface EMG detection, conditioning and pre-processing: Best practices». In: *Journal of Electromyography and Kinesiology* 54 (2020), p. 102440. DOI: 10.1016/j.jelekin.2020.102440 (cit. on p. 4).

[20] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. «Silent Speech Recognition as an Alternative Communication Device for Persons With Laryngectomy». In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 25.12 (2017). Eight sensors; per-channel mel cepstra concatenated into a 112-dimensional vector, the filter bank reduced to fifteen filters for the narrower EMG bandwidth., pp. 2386–2398. DOI: 10.1109/TASLP.2017.2740000. URL: https://ieeexplore.ieee.org/document/8114357/ (cit. on pp. 5, 27).

[21] Geoffrey S. Meltzner, James T. Heaton, Yunbin Deng, Gianluca De Luca, Serge H. Roy, and Joshua C. Kline. «Development of sEMG Sensors and Algorithms for Silent Speech Recognition». In: *Journal of Neural Engineering* 15.4 (2018), p. 046031. DOI: 10.1088/1741-2552/aac965. URL: https://iopscience.iop.org/article/10.1088/1741-2552/aac965 (cit. on p. 5).

[22] Yuta Kurotaki, Shusuke Yamakoshi, Reitaro Yoshida, Yutaka Isoda, Tamami Takano, Yuji Isano, Yusuke Miyake, Kentaro Kuribayashi, and Hiroki Ota. «Soft Active Electromyography Interface for Machine Learning-Enabled Silent Speech Recognition». In: *Advanced Intelligent Systems* (2026). Four perioral channels at 1 kHz; mel cepstra supplied to a dense network, thirty words. DOI:

10.1002/aisy.70440. URL: https://advanced.onlinelibrary.wiley.com/doi/10.1002/aisy.70440 (cit. on p. 5).

[23] Lena Maier-Hein, Florian Metze, Tanja Schultz, and Alex Waibel. «Session Independent Non-Audible Speech Recognition Using Surface Electromyography». In: *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. Reports spectral coefficients superior to cepstral and linear-predictive ones on surface EMG. San Juan, Puerto Rico, 2005, pp. 331–336. DOI: 10.1109/ASRU.2005.1566521 (cit. on p. 5).

[24] Szu-Chen Jou, Tanja Schultz, Matthias Walliczek, Florian Kraft, and Alex Waibel. «Towards Continuous Speech Recognition Using Surface Electromyography». In: *Proc. Interspeech*. Six channels at 600 Hz; the spectral baseline attains 86.8% WER against the 32.0% of the EMG-specific time-domain features. Pittsburgh, PA, 2006, pp. 573–576. DOI: 10.21437/Interspeech.2006-212. URL: https://www.isca-archive.org/interspeech_2006/jou06_interspeech.html (cit. on p. 5).

[25] Jinghan Wu, Yakun Zhang, Liang Xie, Ye Yan, Xu Zhang, Shuang Liu, Xingwei An, Erwei Yin, and Dong Ming. «A Novel Silent Speech Recognition Approach Based on Parallel Inception Convolutional Neural Network and Mel Frequency Spectral Coefficient». In: *Frontiers in Neurorobotics* 16 (2022). Six channels at 1 kHz, 100 Mandarin phrases; mel filter bank energies 90.76%, mel cepstra 82.67%, plain spectrogram 73.31%., p. 971446. DOI: 10.3389/fnbot.2022.971446. URL: https://www.frontiersin.org/articles/10.3389/fnbot.2022.971446/full (cit. on pp. 5, 50).

[26] Matteo Fasulo, Giusy Spacone, Thorir Mar Ingolfsson, Yawei Li, Luca Benini, and Andrea Cossettini. *TinyMyo: a Tiny Foundation Model for Flexible EMG Signal Processing at the Edge*. arXiv preprint arXiv:2512.15729. 2026. URL: http://arxiv.org/abs/2512.15729 (cit. on pp. 5, 7).

[27] Michael Wand and Tanja Schultz. «Towards real-life application of EMG-based speech recognition by using unsupervised adaptation». In: *Proc. Interspeech*. 2014. URL: https://www.isca-archive.org/interspeech_2014/wand14_interspeech.html (cit. on pp. 5, 7).

[28] Anyuan Zhang, Ying He, Haowen Zheng, and Qunchao Lu. «Data Augmentation with Fine-Tuning Strategy for Enhanced Myoelectric Pattern Recognition Using Deep Learning». In: *Proc. 18th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*. 2025, pp. 1–6. DOI: 10.1109/CISP-BMEI68103.2025.11259289. URL: https://ieeexplore.ieee.org/document/11259289/ (cit. on pp. 5, 7, 24).

[29] Michael Wand, Szu-Chen Stan Jou, Arthur R. Toth, and Tanja Schultz. «Impact of different speaking modes on EMG-based speech recognition». In: *Proc. Interspeech*. 2009. URL: https://www.isca-archive.org/interspeech_2009/wand09_interspeech.html (cit. on p. 5).

[30] David Gaddy and Dan Klein. «Digital Voicing of Silent Speech». In: *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2020. URL: https://aclanthology.org/2020.emnlp-main.445/ (cit. on pp. 6–9, 27).

[31] Abhishek Bharali, Bidyut Bikash Borah, Uddipan Hazarika, and Soumik Roy. «Enabling the Translation of Electromyographic Signals Into Speech: A Neural Network Based Decoding Approach». In: *SN Computer Science* (2024). DOI: 10.1007/s42979-024-03457-1. URL: https://link.springer.com/10.1007/s42979-024-03457-1 (cit. on pp. 6, 7, 11).

[32] David Gaddy and Dan Klein. «An Improved Model for Voicing Silent Speech». In: *Proc. Annual Meeting of the Association for Computational Linguistics (ACL)*. 2021. URL: https://aclanthology.org/2021.acl-short.23/ (cit. on pp. 6–9).

[33] Payal Mohapatra et al. *Can LLMs Understand Unvoiced Speech? Exploring EMG-to-Text Conversion with LLMs*. arXiv preprint arXiv:2506.00304. 2025. URL: http://arxiv.org/abs/2506.00304 (cit. on pp. 6, 7, 9).

[34] Peter Wu et al. «Towards EMG-to-Speech with Necklace Form Factor». In: *Proc. Interspeech*. Textile EMG neckband based on the WANDmini acquisition platform; cited by SilentWear [4] as the neckband form-factor inspiration. 2024, pp. 402–406. DOI: 10.21437/Interspeech.2024-1568. URL: https://www.isca-archive.org/interspeech_2024/wu24k_interspeech.html (cit. on pp. 6–8, 91).

[35] Andy Zhou, Samantha R. Santacruz, Benjamin C. Johnson, George Alexandrov, Ali Moin, Fred L. Burghardt, Jan M. Rabaey, Jose M. Carmena, and Rikky Muller. «A wireless and artefact-free 128-channel neuromodulation device for closed-loop stimulation and recording in non-human primates». In: *Nature Biomedical Engineering* 3.1 (2019). WANDmini acquisition platform, used by the neckband of [34], pp. 15–26. DOI: 10.1038/s41551-018-0323-x. URL: https://www.nature.com/articles/s41551-018-0323-x (cit. on pp. 6, 7).

[36] Patrick Kaifosh and Thomas R. Reardon. «A generic non-invasive neuromotor interface for human-computer interaction». In: *Nature* 645.8081 (2025), pp. 702–711. DOI: 10.1038/s41586-025-09255-w. URL: https://www.nature.com/articles/s41586-025-09255-w (cit. on p. 7).

[37] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. «Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks». In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*. 2006, pp. 369–376. DOI: 10.1145/1143844.1143891 (cit. on pp. 7, 28, 32, 36, 37).

[38] Xiaodan Chen, Xiaoxue Gao, Mathias Quoy, Alexandre Pitti, and Nancy F. Chen. *Confidence-Based Self-Training for EMG-to-Speech: Leveraging Synthetic EMG for Robust Modeling*. arXiv preprint arXiv:2506.11862. 2025. URL: http://arxiv.org/abs/2506.11862 (cit. on p. 7).

[39] Xi Chen, Yuanfei Xia, Yong Sun, Le Wu, Xiang Chen, et al. «Silent Speech Recognition Based on High-Density Surface Electromyogram Using Hybrid Neural Networks». In: *IEEE Transactions on Human-Machine Systems* 53.2 (2023), pp. 335–345. DOI: 10.1109/THMS.2022.3226197. URL: https://ieeexplore.ieee.org/document/10006378/ (cit. on p. 7).

[40] Jennifer J. Newson and Tara C. Thiagarajan. «EEG Frequency Bands in Psychiatric Disorders: A Review of Resting State Studies». In: *Frontiers in Human Neuroscience* 12 (2019), p. 521. DOI: 10.3389/fnhum.2018.00521. URL: https://www.frontiersin.org/articles/10.3389/fnhum.2018.00521/full (cit. on pp. 9, 10).

[41] Thomas Donoghue et al. «Parameterizing neural power spectra into periodic and aperiodic components». In: *Nature Neuroscience* 23.12 (2020), pp. 1655–1665. DOI: 10.1038/s41593-020-00744-x (cit. on p. 9).

[42] Motoshige Sato, Ilya Horiguchi, Masakazu Inoue, Kenichi Tomeoka, Eri Hatakeyama, Yuya Kita, Atsushi Yamamoto, Ippei Fujisawa, and Shuntaro Sasai. *A 1000-hour EEG–EMG–audio dataset of Japanese speech production*. arXiv preprint arXiv:2606.01264. 2026. DOI: 10.48550/arXiv.2606.01264. URL: https://arxiv.org/abs/2606.01264 (cit. on pp. 9, 11).

[43] Hanbeot Park, Yunjeong Cho, and Hunhee Kim. *EEG-to-voice decoding of spoken and imagined speech using non-invasive EEG*. arXiv preprint arXiv:2512.22146. 2025. DOI: 10.48550/arXiv.2512.22146. URL: https://arxiv.org/abs/2512.22146 (cit. on pp. 9, 10).

[44] Masakazu Inoue, Eri Hatakeyama, Yuya Kita, and Shuntaro Sasai. «Large-scale training data enhances silent speech decoding with around-ear EEG». In: *Journal of Neural Engineering* (2026). DOI: 10.1088/1741-2552/ae54d0. URL: https://iopscience.iop.org/article/10.1088/1741-2552/ae54d0 (cit. on p. 9).

[45] Zachary M. Fogg, Nicholas S. Card, Maitreyee Wairagkar, Aparna Srinivasan, Tyler Singer-Clark, Xianda Hou, et al. *A generalizable speech neuroprosthesis*. bioRxiv preprint. 2026. DOI: 10.64898/2026.07.23.739430. URL: https://www.biorxiv.org/content/10.64898/2026.07.23.739430v1 (cit. on p. 10).

[46] Datwyler. *SoftPulse Dry Electrodes*. Product documentation. Brush and Dome geometries for dry biopotential contact. URL: https://datwyler.com/company/innovation/softpulse/products/ (cit. on p. 13).

[47] Mattia Orlandi, Pierangelo Maria Rapa, Marcello Zanghieri, Sebastian Frey, Victor Kartsch, Luca Benini, and Simone Benatti. «Real-Time Motor Unit Tracking From sEMG Signals With Adaptive ICA on a Parallel Ultra-Low Power Processor». In: *IEEE Transactions on Biomedical Circuits and Systems* 18.4 (2024). Reference publication of the BioGUI acquisition software., pp. 771–782. DOI: 10.1109/TBCAS.2024.3410840 (cit. on pp. 14, 21, 24, 81).

[48] Victor J. Boucher and Tareck Ayad. «Physiologic Features of Vocal Fatigue: Electromyographic Spectral-Compression in Laryngeal Muscles». In: *The Laryngoscope* 116.6 (2006), pp. 959–965. DOI: 10.1097/01.MLG.0000216824.07244.00 (cit. on p. 18).

[49] Ran Zhao, Yanru Bai, Shuming Zhang, Jianing Zhu, Hongxing Liu, and Guangjian Ni. «An open dataset of multidimensional signals based on different speech patterns in pragmatic Mandarin». In: *Scientific Data* 12 (2025), p. 1934. DOI: 10.1038/s41597-025-06213-z. URL: https://www.nature.com/articles/s41597-025-06213-z (cit. on pp. 18, 20).

[50] Dongliang Zhou, Yakun Zhang, Jinghan Wu, Xingyu Zhang, Liang Xie, and Erwei Yin. *AVE Speech: A Comprehensive Multi-Modal Dataset for Speech Recognition Integrating Audio, Visual, and Electromyographic Signals*. arXiv preprint arXiv:2501.16780. 2025. DOI: 10.48550/arXiv.2501.16780. URL: https://arxiv.org/abs/2501.16780 (cit. on pp. 18, 20, 27, 76).

[51] Chenyu Tang et al. «Wearable intelligent throat enables natural speech in stroke patients with dysarthria». In: *Nature Communications* 17 (2026), p. 293. DOI: 10.1038/s41467-025-68228-9. URL: https://www.nature.com/articles/s41467-025-68228-9 (cit. on p. 20).

[52] Rosanna Turrisi, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga, and Leonardo Badino. *EasyCall corpus: a dysarthric speech dataset*. arXiv preprint arXiv:2104.02542. 2021. DOI: 10.48550/arXiv.2104.02542. URL: https://arxiv.org/abs/2104.02542 (cit. on p. 20).

[53] Antonio Andriella, Raquel Ros, Sharon Gannot, and Séverin Lemaignan. «Dataset and Evaluation of Automatic Speech Recognition for Multi-lingual Intent Recognition on Social Robots». In: *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24)*. Boulder, CO, USA: ACM, 2024. DOI: 10.1145/3610977.3637473. URL: https://dl.acm.org/doi/10.1145/3610977.3637473 (cit. on p. 20).

[54] Snehesh Shrestha, Yantian Zha, Saketh Banagiri, Ge Gao, Yiannis Aloimonos, and Cornelia Fermüller. *NatSGD: A Dataset with Speech, Gestures, and Demonstrations for Robot Learning in Natural Human-Robot Interaction*. arXiv preprint arXiv:2403.02274. 2024. DOI: 10.48550/arXiv.2403.02274. URL: https://arxiv.org/abs/2403.02274 (cit. on p. 20).

[55] Stephanie M. Lukin et al. *SCOUT: A Situated and Multi-Modal Human-Robot Dialogue Corpus*. arXiv preprint arXiv:2411.12844. 2024. DOI: 10.48550/arXiv.2411.12844. URL: https://arxiv.org/abs/2411.12844 (cit. on pp. 20, 76).

[56] Rutav Shah, Roberto Martín-Martín, and Yuke Zhu. *MUTEX: Learning Unified Policies from Multimodal Task Specifications*. arXiv preprint arXiv:2309.14320; Conference on Robot Learning (CoRL). 2023. DOI: 10.48550/arXiv.2309.14320. URL: https://arxiv.org/abs/2309.14320 (cit. on p. 20).

[57] Yujin Tang, Wenhao Yu, Jie Tan, Heiga Zen, Aleksandra Faust, and Tatsuya Harada. *SayTap: Language to Quadrupedal Locomotion*. arXiv preprint arXiv:2306.07580; Conference on Robot Learning (CoRL). 2023. DOI: 10.48550/arXiv.2306.07580. URL: https://arxiv.org/abs/2306.07580 (cit. on pp. 20, 76).

[58] Homer Walke et al. *BridgeData V2: A Dataset for Robot Learning at Scale*. arXiv preprint arXiv:2308.12952; Conference on Robot Learning (CoRL). 2023. DOI: 10.48550/arXiv.2308.12952. URL: https://arxiv.org/abs/2308.12952 (cit. on pp. 20, 76).

[59] Alan V. Oppenheim and Ronald W. Schafer. *Discrete-Time Signal Processing*. 3rd. Pearson, 2010 (cit. on pp. 24, 26).

[60] Fredrik Gustafsson. «Determining the Initial States in Forward-Backward Filtering». In: *IEEE Transactions on Signal Processing* 44.4 (1996), pp. 988–992. DOI: 10.1109/78.492552 (cit. on p. 24).

[61] Jont B. Allen and Lawrence R. Rabiner. «A Unified Approach to Short-Time Fourier Analysis and Synthesis». In: *Proceedings of the IEEE* 65.11 (1977), pp. 1558–1564. DOI: 10.1109/PROC.1977.10770 (cit. on p. 26).

[62] Steven B. Davis and Paul Mermelstein. «Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences». In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28.4 (1980), pp. 357–366. DOI: 10.1109/TASSP.1980.1163420 (cit. on p. 27).

[63] Nasir Ahmed, T. Natarajan, and Kamisetty R. Rao. «Discrete Cosine Transform». In: *IEEE Transactions on Computers* C-23.1 (1974), pp. 90–93. DOI: 10.1109/T-C.1974.223784 (cit. on p. 27).

[64] Stanley S. Stevens, John Volkmann, and Edwin B. Newman. «A Scale for the Measurement of the Psychological Magnitude Pitch». In: *The Journal of the Acoustical Society of America* 8.3 (1937), pp. 185–190. DOI: 10.1121/1.1915893 (cit. on p. 27).

[65] Douglas O'Shaughnessy. *Speech Communication: Human and Machine*. Addison-Wesley, 1987 (cit. on p. 27).

[66] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. «SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition». In: *Proceedings of Interspeech*. 2019, pp. 2613–2617. DOI: 10.21437/Interspeech.2019-2680 (cit. on p. 28).

[67] Min Lin, Qiang Chen, and Shuicheng Yan. *Network In Network*. arXiv preprint arXiv:1312.4400. 2013. DOI: 10.48550/arXiv.1312.4400 (cit. on p. 32).

[68] Sepp Hochreiter and Jürgen Schmidhuber. «Long Short-Term Memory». In: *Neural Computation* 9.8 (1997), pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 32).

[69] Mike Schuster and Kuldip K. Paliwal. «Bidirectional Recurrent Neural Networks». In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681. DOI: 10.1109/78.650093 (cit. on p. 32).

[70] Alex Graves and Jürgen Schmidhuber. «Framewise Phoneme Classification with Bidirectional LSTM and Other Neural Network Architectures». In: *Neural Networks* 18.5–6 (2005), pp. 602–610. DOI: 10.1016/j.neunet.2005.06.042 (cit. on p. 32).

[71] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. «Dropout: A Simple Way to Prevent Neural Networks from Overfitting». In: *Journal of Machine Learning Research* 15.56 (2014), pp. 1929–1958 (cit. on p. 32).

[72] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. «Attention Is All You Need». In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2017, pp. 5998–6008 (cit. on pp. 33–35).

[73] Linhao Dong, Shuang Xu, and Bo Xu. «Speech-Transformer: A No-Recurrence Sequence-to-Sequence Model for Speech Recognition». In: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 5884–5888. DOI: 10.1109/ICASSP.2018.8462506 (cit. on p. 34).

[74] Anmol Gulati et al. «Conformer: Convolution-augmented Transformer for Speech Recognition». In: *Proceedings of Interspeech*. 2020, pp. 5036–5040. DOI: 10.21437/Interspeech.2020-3015 (cit. on p. 34).

[75] Ruibin Xiong et al. «On Layer Normalization in the Transformer Architecture». In: *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 2020, pp. 10524–10533 (cit. on p. 34).

[76] Dan Hendrycks and Kevin Gimpel. *Gaussian Error Linear Units (GELUs)*. arXiv preprint arXiv:1606.08415. 2016. DOI: 10.48550/arXiv.1606.08415 (cit. on p. 35).

[77] Xavier Glorot and Yoshua Bengio. «Understanding the Difficulty of Training Deep Feedforward Neural Networks». In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*. 2010, pp. 249–256 (cit. on p. 35).

[78] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. *Layer Normalization*. arXiv preprint arXiv:1607.06450. 2016. DOI: 10.48550/arXiv.1607.06450 (cit. on p. 35).

[79] Alex Graves and Navdeep Jaitly. «Towards End-to-End Speech Recognition with Recurrent Neural Networks». In: *Proceedings of the 31st International Conference on Machine Learning (ICML)*. 2014, pp. 1764–1772 (cit. on p. 36).

[80] Awni Y. Hannun, Andrew L. Maas, Daniel Jurafsky, and Andrew Y. Ng. *First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs*. arXiv preprint arXiv:1408.2873. 2014. DOI: 10.48550/arXiv.1408.2873 (cit. on p. 36).

[81] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. *Distilling the Knowledge in a Neural Network*. arXiv preprint arXiv:1503.02531. 2015. DOI: 10.48550/arXiv.1503.02531 (cit. on p. 36).

[82] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. «On Calibration of Modern Neural Networks». In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. 2017, pp. 1321–1330 (cit. on p. 36).

[83] Vladimir I. Levenshtein. «Binary Codes Capable of Correcting Deletions, Insertions and Reversals». In: *Soviet Physics Doklady* 10.8 (1966), pp. 707–710 (cit. on pp. 36, 65).

[84] Ron Kohavi. «A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection». In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*. 1995, pp. 1137–1143 (cit. on p. 37).

[85] Ilya Loshchilov and Frank Hutter. «Decoupled Weight Decay Regularization». In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. 2019 (cit. on p. 37).

[86] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. «On the Difficulty of Training Recurrent Neural Networks». In: *Proceedings of the 30th International Conference on Machine Learning (ICML)*. 2013, pp. 1310–1318 (cit. on p. 38).

[87] Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. *Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour*. arXiv preprint arXiv:1706.02677, 2017. DOI: 10.48550/arXiv.1706.02677 (cit. on p. 38).

[88] Diederik P. Kingma and Jimmy Ba. «Adam: A Method for Stochastic Optimization». In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. 2015 (cit. on p. 38).

[89] Kay H. Brodersen, Cheng Soon Ong, Klaas E. Stephan, and Joachim M. Buhmann. «The Balanced Accuracy and Its Posterior Distribution». In: *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*. 2010, pp. 3121–3124. DOI: 10.1109/ICPR.2010.764 (cit. on p. 38).

[90] Marina Sokolova and Guy Lapalme. «A Systematic Analysis of Performance Measures for Classification Tasks». In: *Information Processing and Management* 45.4 (2009), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002 (cit. on p. 39).

[91] Gerhard Staude, Claus Flachenecker, Martin Daumer, and Werner Rau. «Onset Detection in Surface Electromyographic Signals: A Systematic Comparison of Methods». In: *EURASIP Journal on Applied Signal Processing* 2001.2 (2001), pp. 67–81. DOI: 10.1155/S1110865701000191 (cit. on p. 40).

[92] Paul W. Hodges and Bang H. Bui. «A Comparison of Computer-Based Methods for the Determination of Onset of Muscle Contraction Using Electromyography». In: *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control* 101.6 (1996), pp. 511–519. DOI: 10.1016/S0921-884X(96)95190-5 (cit. on pp. 40, 41).

[93] Paolo Bonato, Tommaso D'Alessio, and Marco Knaflitz. «A Statistical Method for the Measurement of Muscle Activation Intervals from Surface Myoelectric Signal During Gait». In: *IEEE Transactions on Biomedical Engineering* 45.3 (1998), pp. 287–299. DOI: 10.1109/10.661154 (cit. on p. 41).

[94] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2nd. Prentice Hall, 2009 (cit. on p. 65).