H.F.R.I.
Hellenic Foundation for Research & Innovation

D3.1 (v3) – The BINGO Benchmarking Framework

BINGO

Brain Imagined-Speech Communication

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

Funded by the European Union
NextGenerationEU

Funded by the
European Union
NextGenerationEU

Greece 2.0
NATIONAL RECOVERY AND RESILIENCE PLAN

H.F.R.I.
Hellenic Foundation for
Research & Innovation

The research project is implemented in the framework of H.F.R.I call “Basic research Financing (Horizontal support of all Sciences)” under the National Recovery and Resilience Plan “Greece 2.0” funded by the European Union – NextGenerationEU (H.F.R.I. Project Number: 15986).

Dissemination level:	Public (PU)
Contractual date of delivery:	Month 24, 27/11/2025
Actual date of delivery:	Month 24, 24/11/2025
Work Package:	WP3 - Benchmarking framework
Task:	T3.2 Data Recording
Type:	DEM
Approval Status:	final
Version:	V3.0
Number of pages:	37
Filename:	D3.1_BINGO_BenchmarkingFramework_v3.docx

Executive Summary: This deliverable investigates EEG-based imagined speech decoding within the BINGO project, aiming to assess decoding performance under varying levels of task complexity and generalization. EEG data were collected from 20 participants over three sessions, using a 26-word vocabulary based on the NATO phonetic alphabet. Multiple evaluation protocols were employed, including within-session and cross-session subject-dependent experiments, as well as subject-independent evaluations using leave-one-subject-out and Monte Carlo cross-validation.

Two deep learning architectures were examined: EEGNet, serving as a compact baseline model, and an EEG Conformer, which integrates convolutional layers with self-attention mechanisms. Results show that decoding performance is highest in within-session settings, decreases in cross-session evaluations, and approaches chance level under subject-independent conditions. Reduced-vocabulary experiments yield more stable performance for both models. Overall, the EEG Conformer consistently matches or slightly outperforms EEGNet, particularly in more challenging evaluation scenarios, while absolute performance highlights the persistent difficulty of imagined speech decoding from EEG signals. The code for the methodologies described in this report are available at the project’s code repository.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided as is and no guarantee or warranty is given that the information is fit for any particular purpose. The user thereof uses the information at its sole risk and liability.

HISTORY

Version	Date	Reason	Revised by
V2.1	3/10/2025	Table of Contents	Kostas Georgiadis
V2.2	17/11/2025	Initial Draft	Kostas Georgiadis
V3.0	24/11/2025	Final	Spiros Nikolopoulos

AUTHOR LIST

Organization	Name	Contact Information
CERTH	Kostas Georgiadis	kostas.georgiadis@iti.gr
CERTH	Fotis P. Kalaganis	fkalaganis@iti.gr
CERTH	Maria Kyrou	mariakyrou@iti.gr
CERTH	Spiros Nikolopoulos	nikolopo@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr

ABBREVIATIONS AND ACRONYMS

BCI	Brain Computer Interface
CNN	Convolutional Neural Network
EEG	ElectroEncephaloGram
NATO	North Atlantic Treaty Organization

Contents

1. INTRODUCTION

Imagined speech, also referred to as inner speech, describes the internal generation of words without any overt articulation. It is a common cognitive process that accompanies thinking, planning, and language-related activities. From an applied perspective, imagined speech has attracted increasing attention in brain–computer interface (BCI) research, particularly in relation to communication support for individuals with severe speech or motor impairments.

Previous studies have shown that imagined speech involves neural activity in frontal, temporal, and sensorimotor areas. This activity is typically distributed both spatially and temporally and is not associated with strong, time-locked responses. As a result, EEG signals recorded during imagined speech tasks tend to have a low signal-to-noise ratio and are sensitive to session-related and inter-subject differences. These properties complicate the design of reliable decoding models.

Recent work has increasingly explored deep learning approaches for EEG-based decoding. Convolutional neural networks have been used to learn temporal and spatial representations directly from EEG signals, reducing the need for handcrafted features. More recently, attention-based architectures have been introduced to capture longer temporal dependencies. While these methods have shown promising results, performance often depends strongly on the evaluation protocol and the degree of subject specificity.

In many studies, evaluation is performed under subject-dependent conditions, where training and testing data originate from the same participant. Subject-independent evaluation, where models are applied to unseen users, remains considerably more difficult and is therefore essential for practical BCI applications. In addition, task complexity, such as vocabulary size, has a direct impact on decoding performance and should be explicitly considered.

In this document, imagined speech decoding is investigated using EEG data collected during a multi-session experiment with a 26-word vocabulary based on the NATO (North Atlantic Treaty Organization) phonetic alphabet. The experimental protocol includes recordings over three days per participant, allowing evaluation under both stable and variable conditions. Several evaluation strategies are employed, including within-session and cross-session subject-dependent experiments, as well as subject-independent settings using leave-one-subject-out and Monte Carlo cross-validation.

Two neural network models are examined. EEGNet is used as a compact baseline architecture designed specifically for EEG classification tasks. In addition, an EEG Conformer model is evaluated, which combines convolutional layers with self-attention mechanisms to model both local and long-range temporal structure. Both models are trained and evaluated using the same preprocessing pipeline and evaluation protocols, enabling a direct comparison of their performance across different experimental conditions.

The aim of this work is to assess imagined speech decoding performance across varying levels of generalization. By considering multiple evaluation paradigms and two complementary architectures, this deliverable provides a structured analysis of the challenges associated with EEG-based imagined speech decoding. It is noted that the code for the methodologies described in this report are available at the project's code repository https://github.com/BINGO-BCI and will constitute a part of project's toolbox.

2. BINGO DATASET

The dataset comprises Electroencephalogram (EEG) recordings collected from 20 participants during a controlled imagined speech experiment conducted in CERTH EEG laboratory (for details you may refer to D3.1 v2). After signing a consent form, each participant attended three experimental sessions on three separate days, following an identical protocol designed to elicit neural activity associated with covert speech production.

The imagined speech vocabulary consisted of 26 words derived from the NATO phonetic alphabet. To manage task complexity and participant fatigue, the vocabulary was distributed across the first two experimental sessions. During the first session, participants were instructed to imagine the pronunciation of the first 13 words (Alpha, Bravo, Charlie, Delta, Echo, Foxtrot, Golf, Hotel, India, Juliett, Kilo, Lima, Mike), with each word repeated 35 times, resulting in 455 trials per participant. The second session followed the same structure and repetition scheme, focusing on the remaining 13 words (November, Oscar, Papa, Quebec, Sierra, Tango, Uniform, Victor, Whiskey, X-Ray, Yankee, Zulu), again yielding 455 trials per participant. In the third session, participants imagined the full vocabulary of 26 words, each repeated 10 times, leading to 260 trials per participant. Across all sessions, each subject contributed a total of 1,170 imagined speech trials.

Each trial followed a fixed temporal structure with a total duration of 4 seconds. At the beginning of the trial, the target word was visually presented on the screen. This was followed by a brief preparation phase, indicated by the sequential appearance of three dots, and subsequently by a fixation cross serving as the cue for task onset. Immediately after the fixation cross, participants were given a 1.5-second interval during which they were instructed to imagine pronouncing the displayed word once internally. Subjects were explicitly instructed to avoid any overt speech, facial movements, or motor actions throughout the experiment, ensuring that the recorded EEG signals predominantly reflected neural processes related to imagined speech rather than muscle activity.

EEG data were acquired using the DSI-24 EEG device by Wearable Sensing [Wearable Sensing, DSI-24]. Recordings were performed at a sampling frequency of 300 Hz using 21 electrodes positioned according to the International 10–20 System. The electrode montage included the channels Fp1, Fp2, Fz, F3, F4, F7, F8, Cz, C3, C4, T3, T4, Pz, P3, P4, T5, T6, O1, O2, A1, and A2. The A1 and A2 electrodes were placed on the mastoid bones and served as reference electrodes. Prior to the experimental procedure, electrode impedances were maintained below 10 kΩ for all channels, and the EEG signals were visually inspected to ensure signal quality and to identify any anomalies before data acquisition.

3. EVALUATION APPROACHES

To evaluate the performance of the imagined speech decoding models under different levels of difficulty and generalization, both subject-dependent and subject-independent evaluation paradigms were adopted. These evaluation settings are commonly used in EEG-based machine learning studies, as they allow performance to be examined under controlled conditions as well as in scenarios that better reflect real-world deployment.

3.1 SUBJECT-DEPENDENT

In the subject-dependent evaluation, training and testing data originate from the same participant, thereby removing inter-subject variability and providing an estimate of the best achievable performance for a given individual. Two complementary subject-dependent scenarios were considered.

WITHIN-SESSION

First, a within-session (within-day) evaluation was carried out using data from either Day 1 or Day 2 independently. Since each of these sessions involved only half of the vocabulary, the task was formulated as a 13-class classification problem. For each subject, 30 trials per class were used for training, and 5 trials per class were reserved for testing. This setting was selected to assess decoding performance under highly stable recording conditions, where temporal variability is minimal, and is typically used as a baseline benchmark in imagined speech and EEG classification studies.

CROSS-SESSION

Second, a cross-session subject-dependent evaluation was performed to investigate how well the models generalize across different recording days. In this configuration, data from Day 1 and Day 2 were used for training, while data from Day 3 were used exclusively for testing. As the third session included the complete vocabulary, this setting resulted in a 26-class classification task. This evaluation scenario was designed to capture session-to-session variability, while still maintaining a subject-specific training framework.

3.2 SUBJECT-INDEPENDENT

The subject-independent evaluation aimed to assess model generalization across participants, which represents a substantially more challenging and practically relevant scenario. In this setting, models are required to decode imagined speech from users whose data were not seen during training, either entirely (as in leave-one-subject-out cross-validation) or partially through subject-independent train-test splits (as in Monte Carlo cross-validation).

LEAVE-ONE-SUBJECT-OUT CROSS-VALIDATION (LOSO-CV)

The primary subject-independent evaluation employed a leave-one-subject-out cross-validation (LOSO-CV) strategy. Data from all days and all but one subject were used for training, while the held-out subject's data were used for testing. This resulted in a 26-class classification problem and provided a rigorous assessment of inter-subject generalization. LOSO-CV is widely regarded as a standard benchmark in EEG decoding studies, as it closely reflects real-world use cases in which subject-specific calibration data may not be available.

MONTE CARLO CROSS-VALIDATION (MC-CV)

In addition, a reduced-vocabulary subject-independent evaluation was conducted using data from Day 1 across all subjects. In this setting, the task involved 13 classes, corresponding to the vocabulary presented on the selected day. For each subject, 500 trials were randomly selected for training, while 85 trials were held out for testing. To ensure robust performance estimation, this random train-test split was repeated 500 times, following a Monte Carlo cross-validation (MC-CV) scheme. This evaluation setup was included to examine subject-independent performance under a reduced vocabulary and to decrease classification complexity.

Taken together, these evaluation scenarios provide a balanced and comprehensive assessment of imagined speech decoding performance, spanning controlled subject-specific conditions, cross-session robustness, and subject-independent generalization, while also accounting for the effect of vocabulary size on classification difficulty.

3.3 A NOVEL EVALUATION METRIC

In traditional classification tasks, accuracy is calculated by dividing the number of correctly classified samples by the total number of samples. However, this approach treats all classes equally, failing to consider that some classes may be more important or difficult than others, which could lead to misleading conclusions when evaluating model performance. For example, in our imagined speech classification experiment, each class corresponds to a letter. However, not all letters are equally significant or difficult to classify. Some classes may be harder to classify (e.g., infrequent letters), or certain classes may hold more importance for the application (e.g., letters that are more commonly used).

In this study, we introduced the Weighted Accuracy as a novel evaluation metric designed to account for the relative importance or contribution of each class in a multi-class classification task. This metric is calculated as the weighted average of the per-class accuracies, where the weight assigned to each class reflects its relative importance or significance in the context of this task (Table 3.1).

The Total Weighted Accuracy is computed as:

$$Total\ Weighted\ Accuracy = \sum_{i=1}^{n} Accuracy_i * Weight_i$$

where n is the number of classes (23 or 13 in this case), $Accuracy$ is the classification accuracy for class i , i is the $Weight$ assigned to class i, representing its importance.

Table 1 Letters and their corresponding significance Weights.

Letter	**Weight**
A	0.0817
B	0.0149
C	0.0278
D	0.0425
E	0.127
F	0.0223
G	0.0202
H	0.0609
I	0.0697
J	0.0015
K	0.0077
L	0.0403
M	0.0241
N	0.0675
O	0.0751
P	0.0193
Q	0.001
R	0.0599
S	0.0633
T	0.0906
U	0.0276
V	0.0098
W	0.0236
X	0.0015
Y	0.0197
Z	0.0007
Sum	1.0002

4. EEGNET

4.1 MOTIVATION

Imagined speech decoding from EEG signals remains a challenging problem, largely due to the subtle and distributed nature of the underlying neural activity. Neurophysiological studies suggest that imagined speech engages a network of frontal, temporal, and motor-related regions, with activity patterns that are distributed across multiple frequency bands rather than confined to a single oscillatory rhythm [Cooney et al., 2018; Martin et al., 2014]. These characteristics motivate the use of learning approaches that can capture spatiotemporal structure while remaining robust to noise and inter-trial variability.

EEGNet [Lawhern et al., 2018] was selected in this work as a suitable baseline architecture as it was explicitly designed for EEG-based classification tasks and incorporates architectural constraints that reflect common properties of EEG signals. Rather than relying on extensive manual feature extraction, EEGNet supports end-to-end learning, allowing task-relevant temporal and spatial patterns to be learned directly from the data. This is particularly appropriate in the context of imagined speech, where the precise neural markers are not yet fully understood and may vary across subjects and sessions.

The temporal convolutional layers employed in EEGNet can be interpreted as data-driven temporal filters, enabling the model to emphasize temporal (spanning across various frequencies) components that are informative for the task. This aligns with prior findings indicating that imagined speech-related activity may involve contributions from multiple frequency ranges, including theta, alpha, beta, and low gamma bands, depending on task design and cognitive strategy [Brigham & Kumar, 2010; Sereshkeh et al., 2019]. By learning such representations implicitly, EEGNet avoids assumptions about fixed frequency bands and allows the model to adapt to the characteristics of the recorded data. In addition, EEGNet employs depthwise spatial convolutions that model relationships across EEG channels while maintaining a relatively small number of trainable parameters. From a methodological perspective, EEGNet is widely used as a reference architecture in EEG-based BCI research. Its inclusion in this study facilitates comparison with prior work and provides a well-established baseline for systematic evaluation under both subject-dependent and subject-independent settings.

4.2 METHODOLOGY

PREPROCESSING

The EEG signals were initially bandpass filtered between 1 and 145 Hz using a 3rd-order Butterworth filter to remove slow drifts and out-of-band noise. A 50 Hz notch filter was applied to suppress power-line interference. To further improve signal quality, Artifact Subspace Reconstruction (ASR) [Kothe & Jung, 2016] was employed to attenuate transient, high-variance artifacts while preserving the underlying neural activity.

Prior to model training, the EEG data were normalized using z-score standardization in order to reduce inter-channel amplitude variability and improve numerical stability during learning. Normalization was performed in a training-aware manner to prevent information leakage from the test data into the training process. For each evaluation split, the dataset was first divided into training and testing subsets. The normalization parameters were then computed exclusively on the training data. Specifically, for each EEG channel, the mean and variance were calculated across all training trials and all time samples. These channel-wise statistics were subsequently used to standardize both the training and testing data.

Given training data $X_{train} \in \mathbb{R}^{N_{tr}xCxT}$ and testing data $X_{test} \in \mathbb{R}^{N_{te}xCxT}$, where N_{tr} and N_{te} are the number of train and test trials, respectively, C are the electrode channels, and T denotes the time samples, the mean and variance were computed per channel by averaging over the trial and temporal dimensions. The resulting parameters were then applied to normalize the data as:

$$X_{norm} = \frac{X - \mu}{\sqrt{\sigma^2}}$$

where μ and σ^2 denote the channel-wise mean and variance estimated from the training set. By using statistics derived solely from the training data, this preprocessing strategy ensures a fair evaluation and avoids bias in both subject-dependent and subject-independent settings. This normalization procedure was applied consistently across all experiments and evaluation scenarios.

NETWORK ARCHITECTURE

The proposed architecture is an end-to-end trainable deep neural network based on EEGNet (Figure 4.1), specifically designed for efficient decoding of imagined speech from EEG signals. Below, we describe the main components of the network used for imagined speech decoding:

- **Temporal convolutional layer**

The first layer applies convolutional temporal filters implemented as 2D convolutions with kernel size (1,$\frac{fs}{2}$), where fs denotes the sampling frequency. These filters act as learnable band-pass filters, extracting frequency-specific temporal patterns from the EEG signals while preserving the channel dimension. This layer is followed by batch normalization to stabilize training.

- **Depthwise spatial convolution**

To model spatial relationships across EEG channels, a depthwise convolution is applied with kernel size (C, 1), where C is the number of EEG channels. This operation learns spatial filters independently for each temporal feature map, effectively capturing channel-wise dependencies. The depthwise convolution is followed by batch normalization, a non-linear activation function (ReLU or ELU), average pooling with kernel size (1, 4), and dropout with probability 0.5 to reduce overfitting.

- **Separable convolutional layer**

Next, a separable convolution is employed, consisting of a depthwise temporal convolution followed by a pointwise convolution. This layer further refines the extracted features by jointly modeling temporal dynamics and inter-feature interactions, while maintaining a low parameter count. As before, batch normalization, non-linearity, average pooling, and dropout are applied.

- **Feature aggregation and classification**

The resulting feature maps are flattened and passed to a fully connected linear layer that produces the final class scores corresponding to the imagined speech categories. The entire network is trained end-to-end using backpropagation, enabling joint optimization of temporal, spatial, and discriminative features directly from raw EEG signals.

Overall, EEGNet provides an efficient and interpretable architecture that leverages biologically meaningful constraints while remaining computationally lightweight, making it well suited for imagined speech decoding tasks with limited training data.

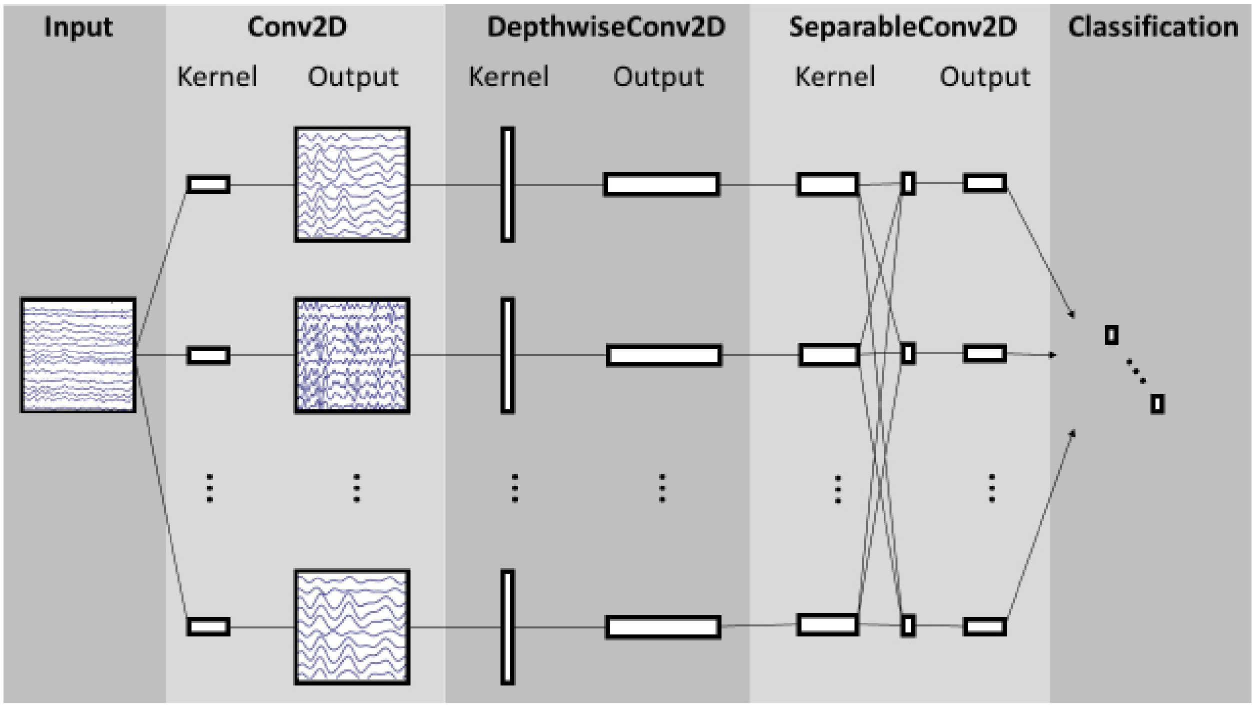

Fig. 4.1 The EEGNet architecture. The network starts with a temporal convolution (second column), then uses a depthwise convolution (middle column). The separable convolution (fourth column) is a combination of a depthwise convolution, followed by a pointwise convolution. Image source: [Lawhern et al., 2018]

4.3 RESULTS

SUBJECT-DEPENDENT

A systematic hyperparameter search was conducted using 5-fold stratified cross-validation on the complete training set, which consisted of 30 trials per class. The goal of this procedure was to identify the optimal learning configuration for the EEGNet model while ensuring robust performance estimation across folds. The hyperparameter search explored a small set of AdamW configurations, focusing on two learning rates and combined with light to no L_2 regularization. For each configuration, a 5-fold stratified cross-validation scheme was applied to preserve class balance across folds. In each fold, the training set was further split into fold-specific training and validation subsets.

An EEGNet architecture was trained using the EEGClassifier framework with a cross-entropy loss function and the AdamW optimizer. Training was performed for a maximum of 200 epochs with a batch size of 64. Early stopping was employed based on the validation loss, with a patience of 10 epochs, to prevent overfitting. Classification accuracy was monitored throughout training. For each fold, the validation accuracy and the number of training epochs until convergence were recorded. After completing the five folds, the mean validation accuracy, standard deviation, and mean number of epochs were computed for each hyperparameter configuration. The configuration yielding the highest mean cross-validation accuracy was selected as the optimal setting. Using the best-performing hyperparameter configuration, a final EEGNet model was trained on the entire training dataset (30 trials per class). The number of training epochs was fixed to the mean number of epochs observed during cross-validation for the selected configuration, ensuring consistency between the optimization and final training stages. The trained model was subsequently evaluated on an independent held-out test set consisting of 5 trials per class. Model performance was quantified using overall classification accuracy, computed as the percentage of correctly predicted labels across all test samples.

WITHIN-SESSION (13-CLASS)

DAY1

Figure 4.2 shows the classification accuracy per subject for EEGNet on Day 1. The overall mean accuracy across subjects is **7.31%**, slightly below the theoretical chance level of 7.7%, with noticeable inter-subject variability. The corresponding weighted accuracy (Figure 4.4) yields an overall mean of **0.32**, indicating that performance is unevenly distributed across classes, with higher-weighted letters contributing more strongly to the final score.

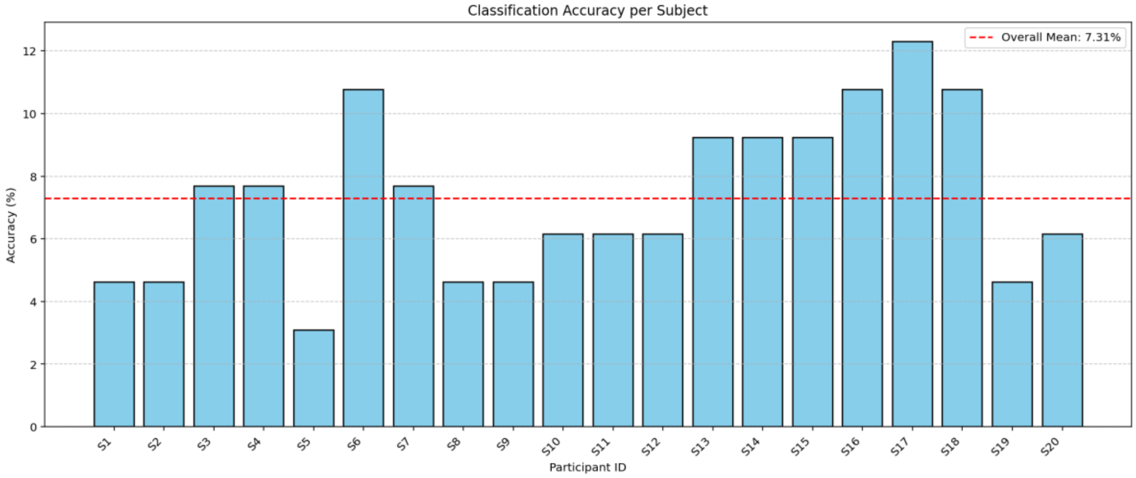

Fig. 4.2 EEGNet classification accuracy per subject for DAY1.

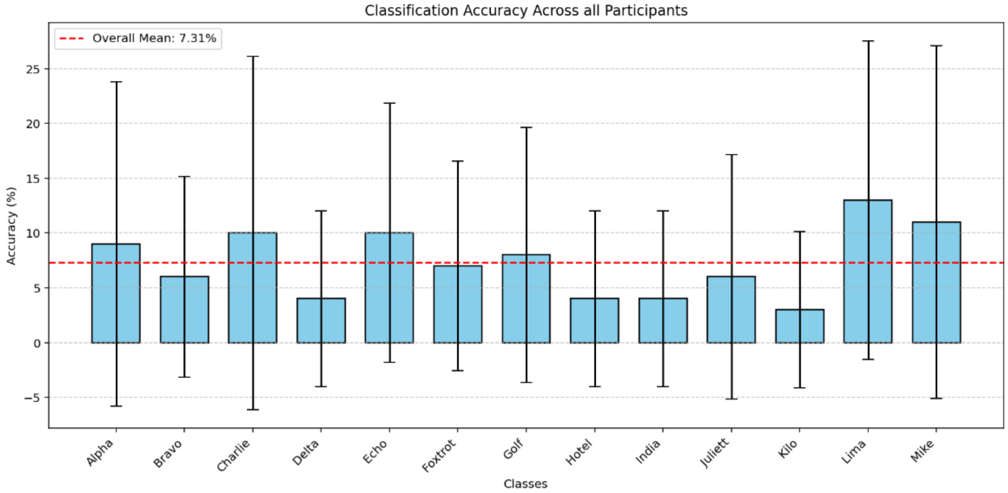

Fig. 4.3 EEGNet classification accuracy per class across all subjects for DAY1.

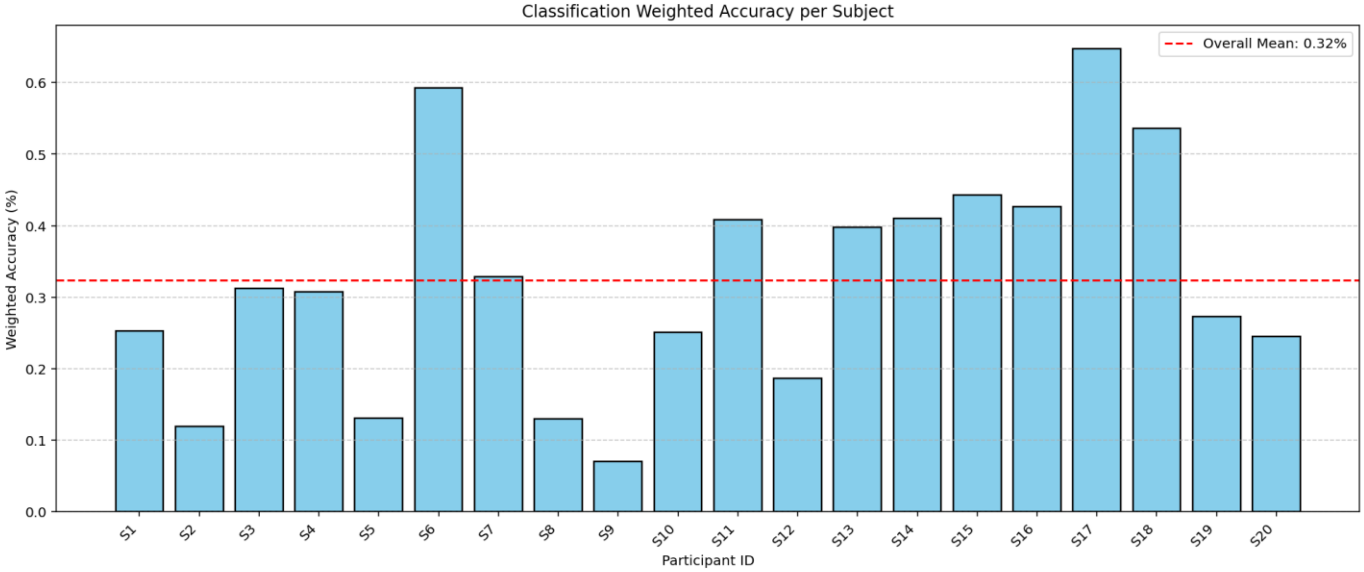

Fig. 4.4 EEGNet classification accuracy weighted by letter importance per subject for DAY1.

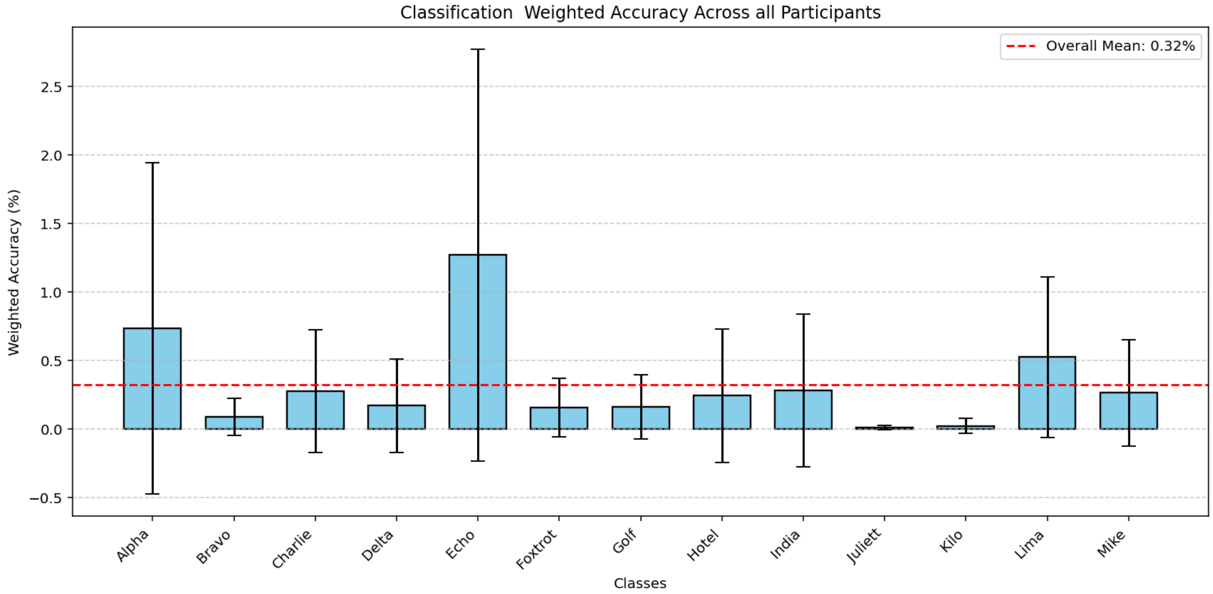

Fig. 4.5 EEGNet classification accuracy weighted by letter importance per class across subjects for DAY1.

DAY 2

As shown in Figure 4.6, the overall mean accuracy on Day 2 decreases to **6.38%**, with a mean weighted accuracy of **0.21** (Figure 4.8). Compared to Day 1, both accuracy and weighted accuracy are reduced, suggesting increased variability or reduced consistency in imagined speech representations across sessions.

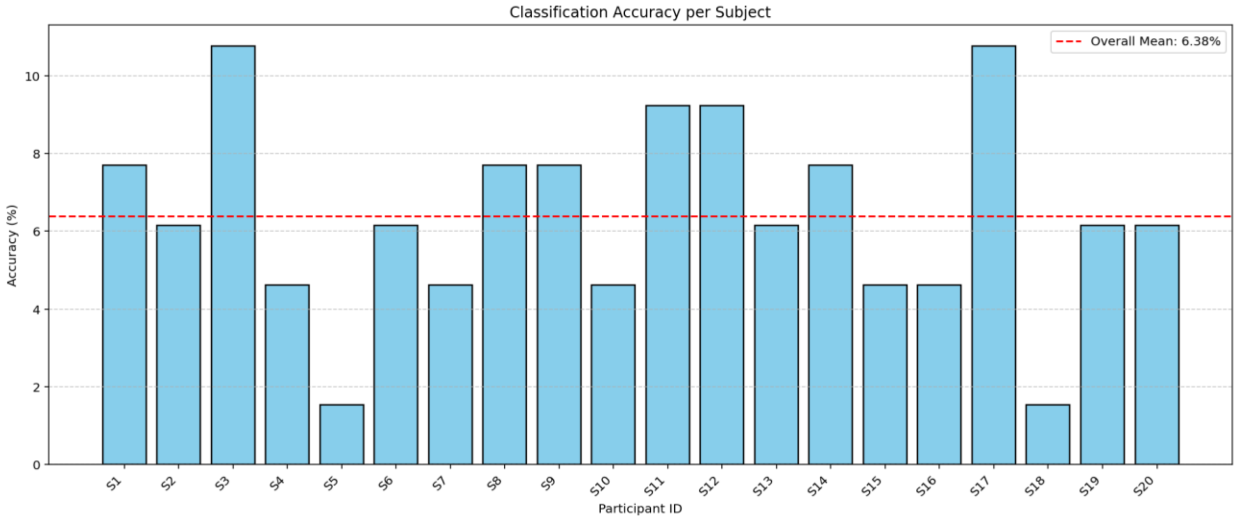

Fig. 4.6 EEGNet classification accuracy per subject for DAY2.

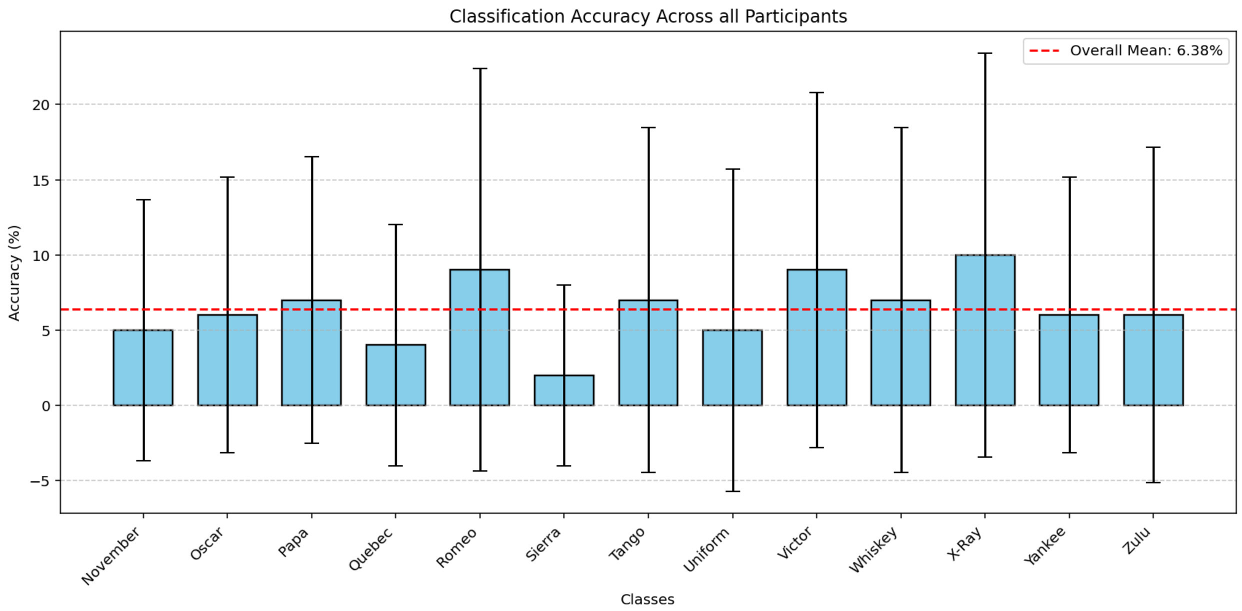

Fig. 4.7 EEGNet classification accuracy per class across subjects for DAY2.

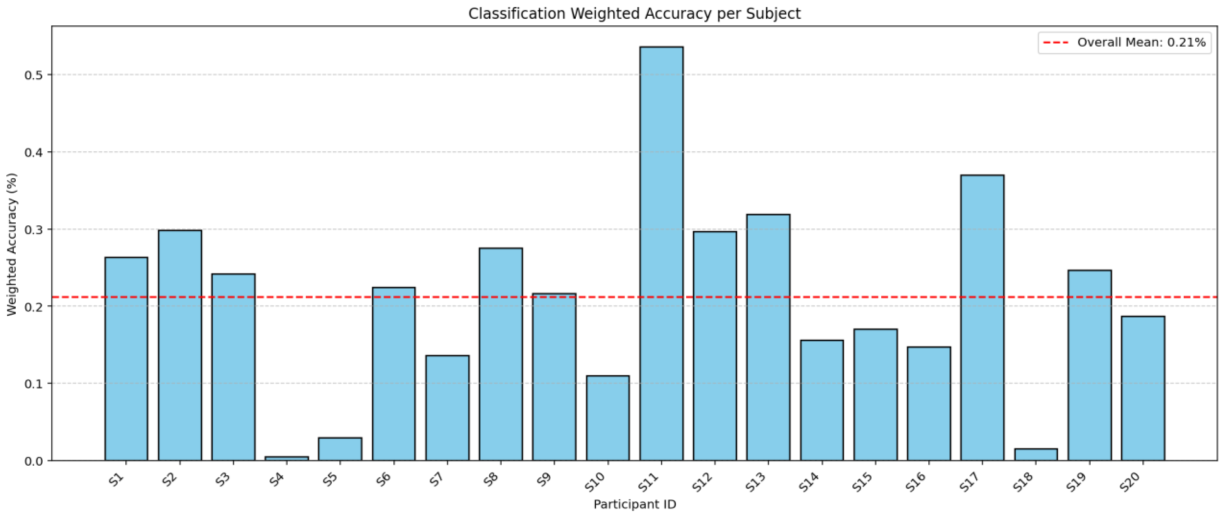

Fig. 4.8 EEGNet classification accuracy weighted by letter importance per subject for DAY2.

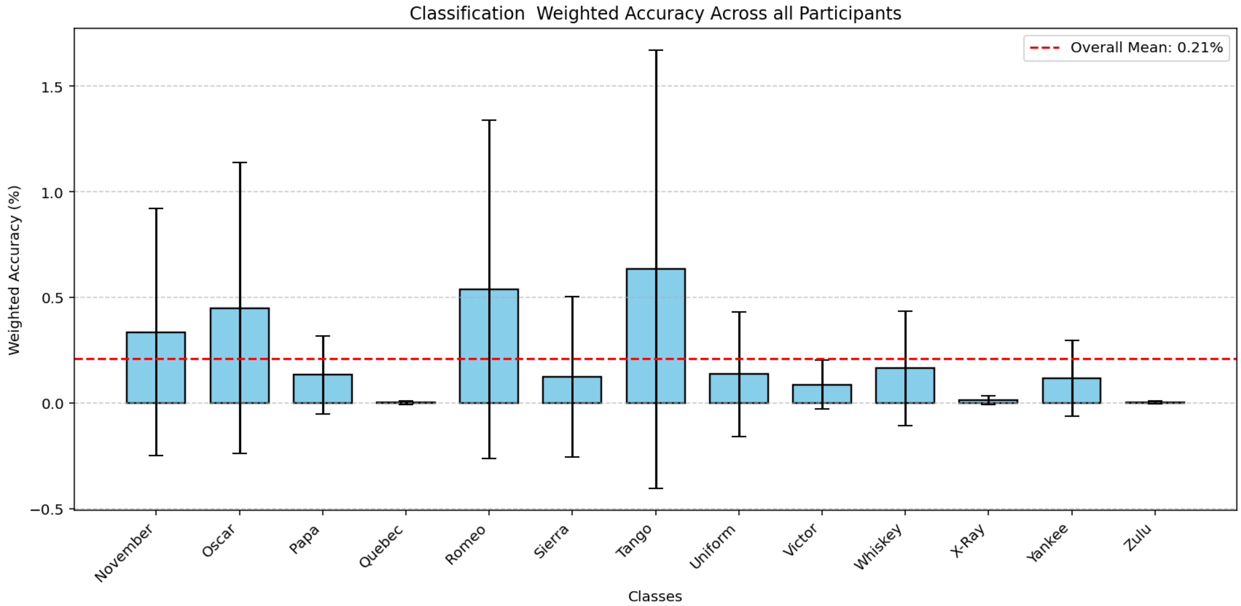

Fig. 4.9 EEGNet classification accuracy weighted by letter importance per class across subjects for DAY2.

CROSS-SESSION (26-CLASS)

Figure 4.10 illustrates EEGNet performance when training on Days 1–2 and testing on Day 3. The overall mean accuracy across subjects is **3.87%**, marginally above the chance level of 3.85%. The corresponding weighted accuracy has an overall mean of **0.14**, reflecting the increased difficulty of generalizing across recording days in a larger vocabulary setting.

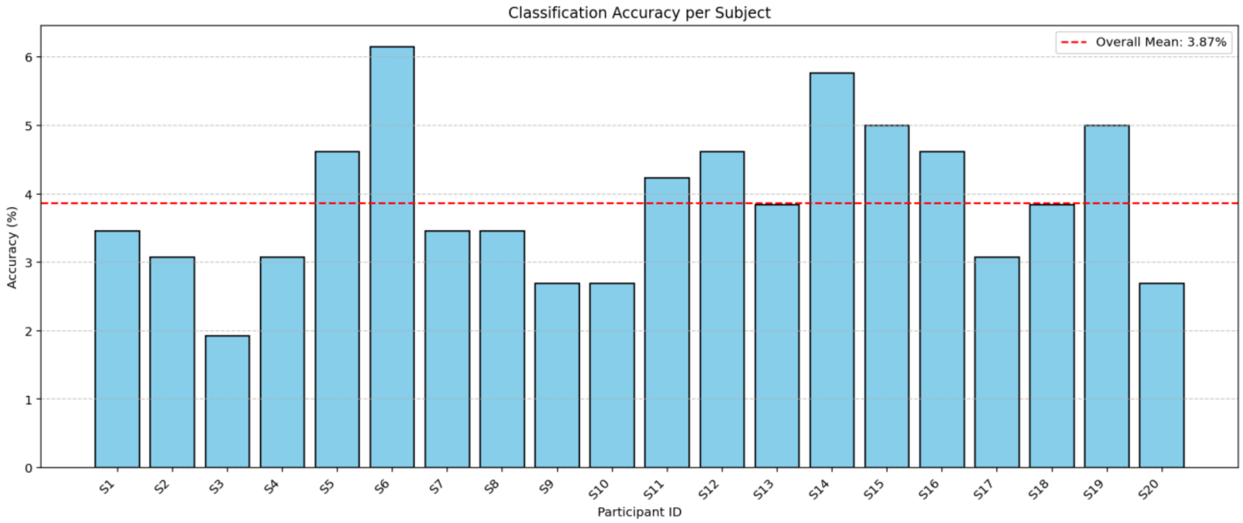

Fig. 4.10 EEGNet classification accuracy per subject, across all sessions (train on DAY1-2, test on DAY3).

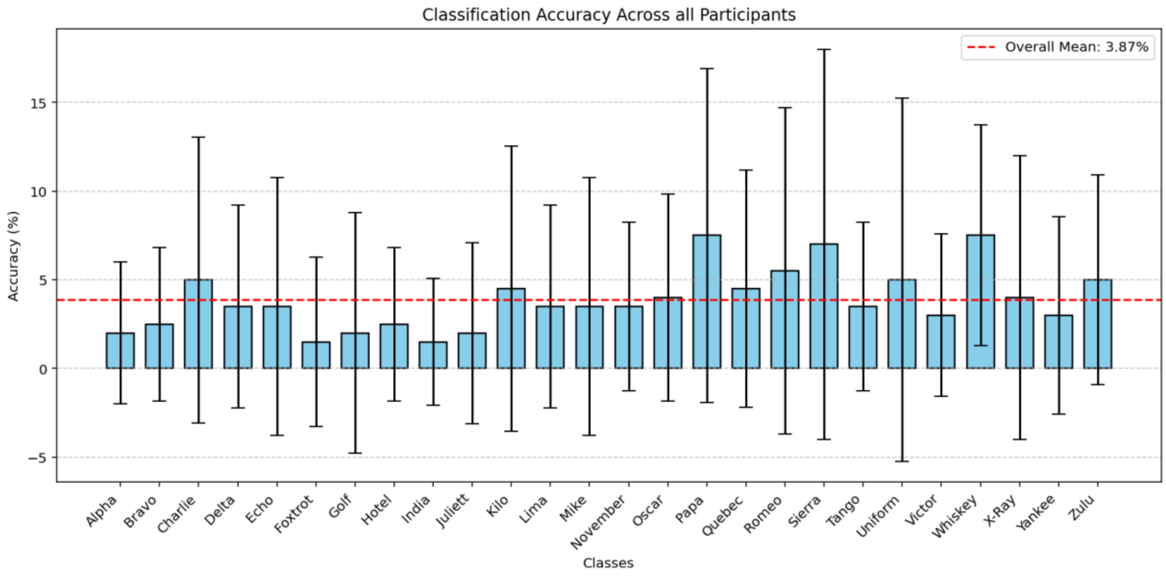

Fig. 4.11 EEGNet classification accuracy per class across subjects, across all sessions (train on DAY1-2, test on DAY3).

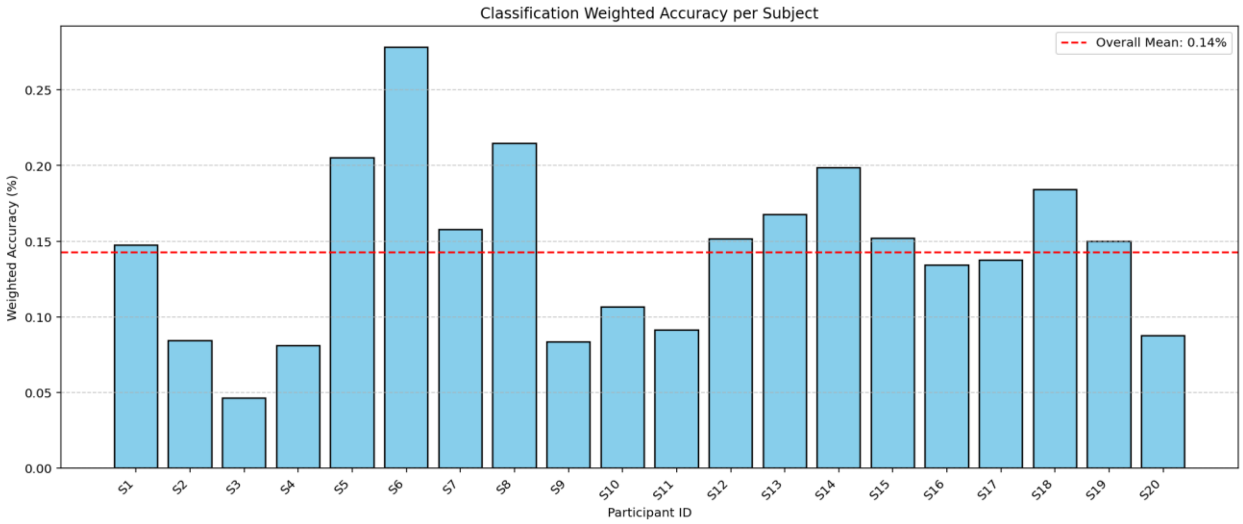

Fig. 4.12 EEGNet classification accuracy weighted by letter importance per subject, across all sessions (train on DAY1-2, test on DAY3).

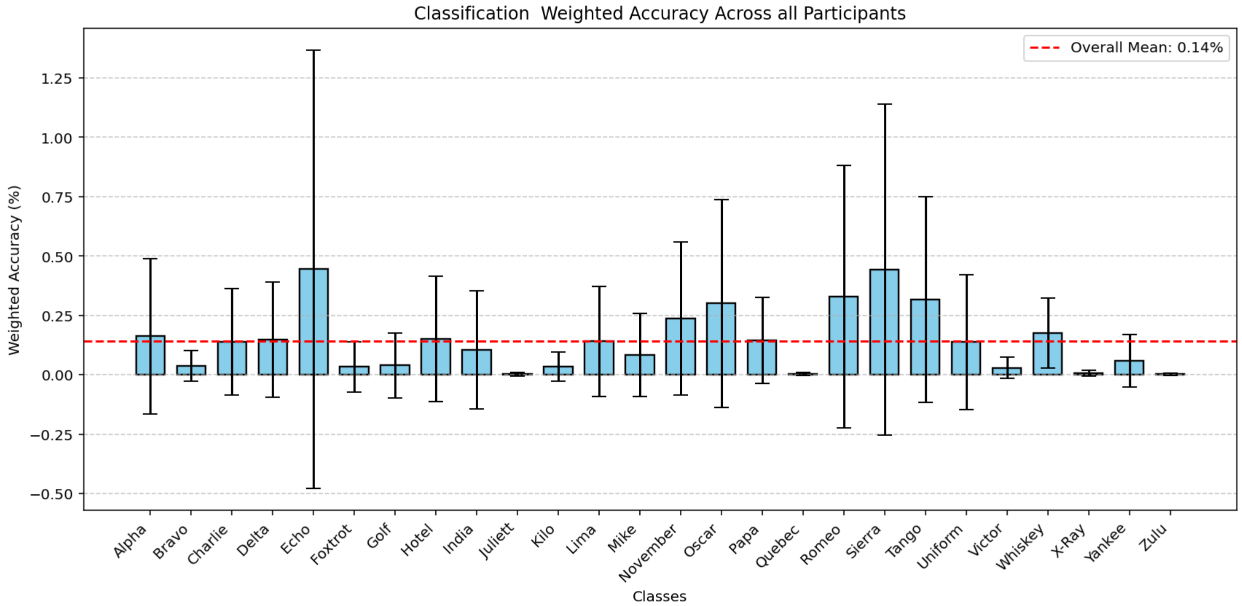

Fig. 4.13 EEGNet classification accuracy weighted by letter importance per class across subjects, across all sessions (train on DAY1-2, test on DAY3).

SUBJECT-INDEPENDENT

LOSO-CV (26-class)

Figure 4.14 presents subject-wise accuracies under LOSO cross-validation. EEGNet achieves an overall mean accuracy of 3.85%, indicating performance at chance level when evaluated on unseen subjects.

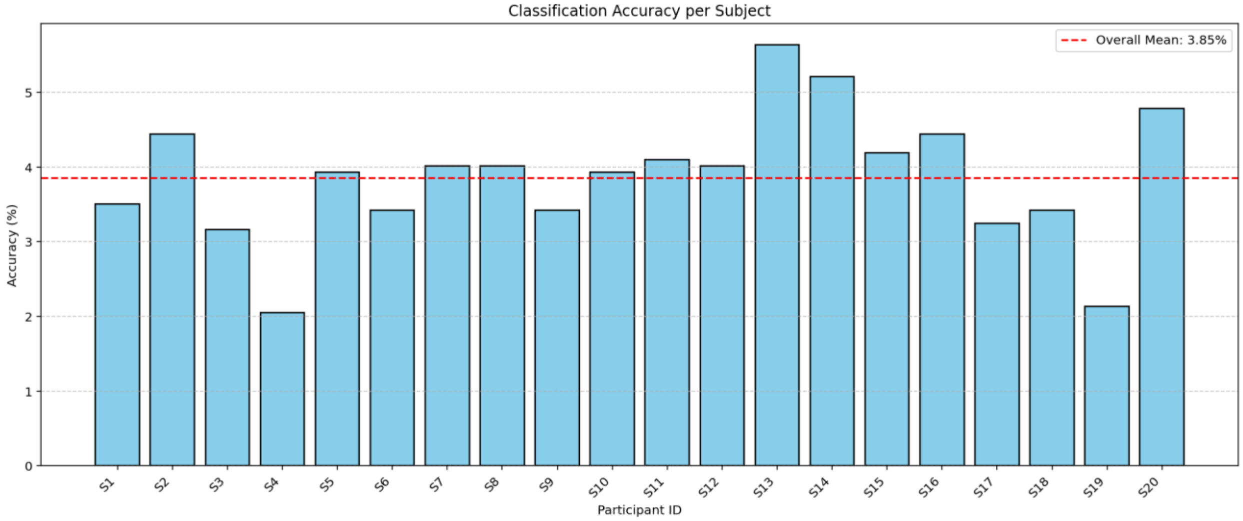

Fig. 4.14 EEGNet classification accuracy per subject, in a Leave-One-Subject-Out manner, across all sessions.

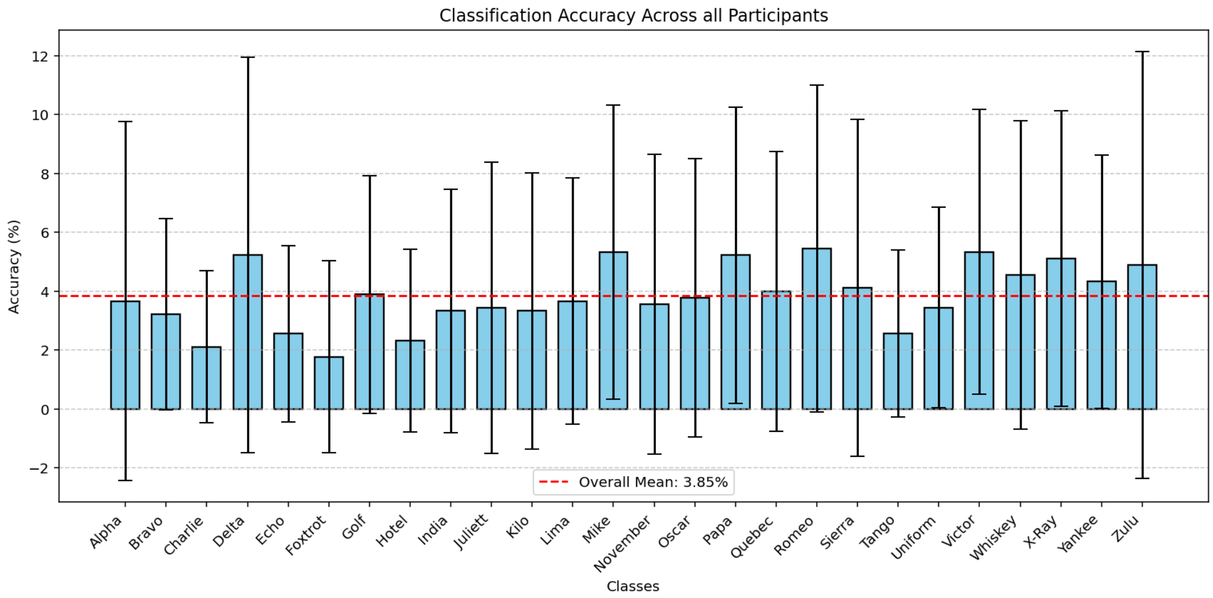

Fig. 4.15 EEGNet classification accuracy per class across all subjects evaluated in a Leave-One-Subject-Out manner, across all sessions.

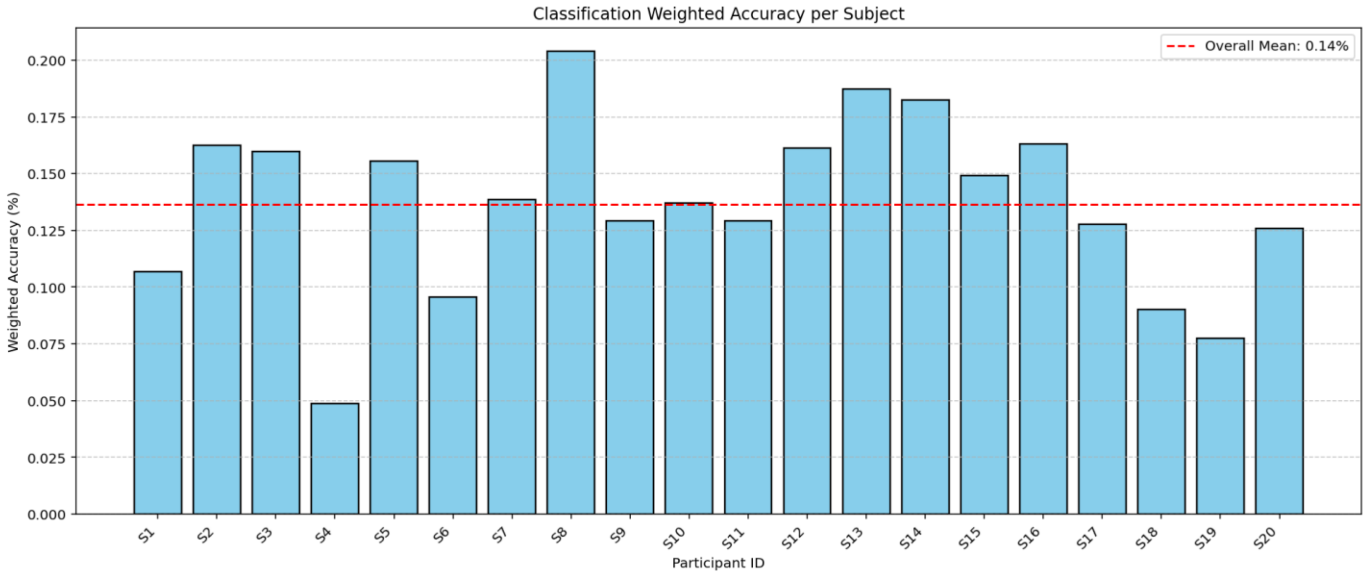

Fig. 4.16 EEGNet classification accuracy weighted by letter importance per subject, in a Leave-One-Subject-Out manner, across all sessions.

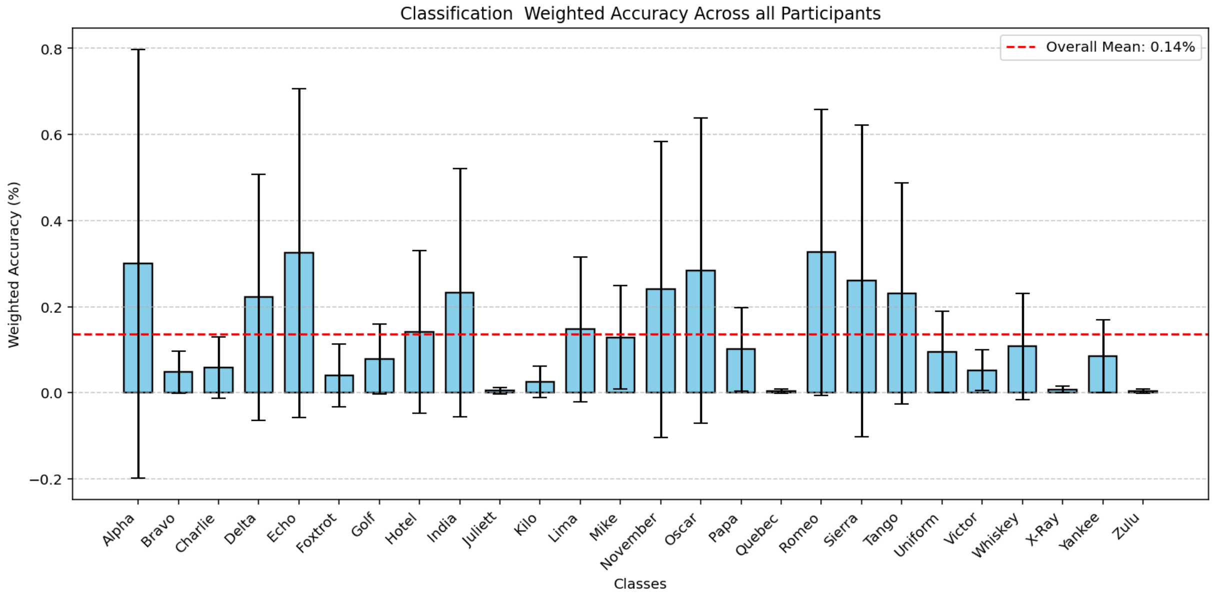

Fig. 4.17 EEGNet classification accuracy weighted by letter importance per class across all subjects evaluated in a Leave-One-Subject-Out manner, across all sessions.

MC-CV (13-class)

Table 4.1 summarize the Monte Carlo cross-validation results. EEGNet achieves a mean accuracy of **7.65%** with a mean weighted accuracy of **0.31**, showing improved stability when both vocabulary size and inter-subject variability are reduced.

Class	**Accuracy %**	**Weighted Accuracy %**
Alpha	7.23	0.59
Bravo	7.59	0.11
Charlie	7.66	0.21
Delta	7.48	0.31
Echo	7.62	0.96
Foxtrot	7.31	0.16
Golf	6.57	0.13
Hotel	9.27	0.56
India	7.02	0.48
Juliet	9.00	0.013
Kilo	6.64	0.051
Lima	7.73	0.31
Mike	8.42	0.20
Mean	**7.65**	**0.31**

Table 4.1 EEGNet classification performance (unweighted and weighted) per class (train on DAY1, test on Day3) evaluated in MC-CV manner.

5. EEG CONFORMER

5.1 MOTIVATION

The EEG Conformer [Song et al., 2022] was employed in this study for inner speech decoding, as it is particularly well suited to the temporal characteristics of EEG signals. Although originally developed for speech-related tasks, the EEG Conformer transfers effectively to EEG analysis through its combination of convolutional layers and attention mechanisms. This architecture enables efficient extraction of short-term EEG features while simultaneously integrating information across longer temporal windows. As a result, it provides a robust framework for modeling inner speech-related neural dynamics and for handling variability across subjects.

5.2 METHODOLOGY

PREPROCESSING

The EEG signals were preprocessed following the exact procedure described in Section 4.2. In brief, the same bandpass and notch filtering, artifact attenuation using ASR, and channel-wise z-score normalization were applied.

Importantly, normalization was performed in a training-aware manner, where channel-wise statistics were computed exclusively on the training data for each evaluation split and subsequently applied to both training and testing sets. This strategy prevents information leakage and ensures a fair evaluation in both subject-dependent and subject-independent experimental settings.

NETWORK ARCHITECTURE

The overall framework is depicted in Fig. 5.1. The architecture comprises three components: a convolution module, a self-attention module, and a fully-connected classifier. In the convolution module, taking the raw two-dimensional EEG trials as the input, temporal and spatial convolutional layers are applied along the time dimension and electrode channel dimensions, respectively. Then, an average pooling layer is utilized to suppress noise interference while improving generalization. Secondly, the spatial-temporal representation obtained by the convolution module is fed into the self-attention module. The self-attention module further extracts the long-term temporal features by measuring the global correlations between different time positions in the feature maps. Finally, a compact classifier consisting of several fully-connected layers is adopted to output the decoding results.

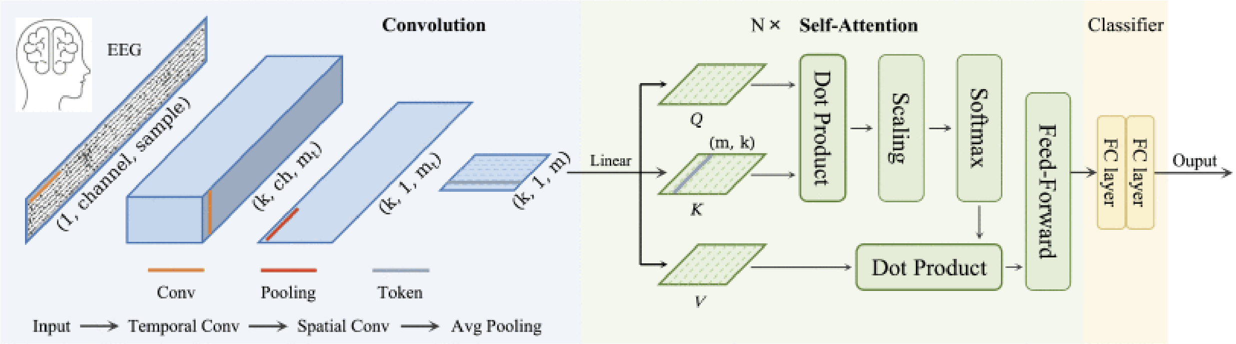

Fig. 5.1. The framework of Convolutional Transformer (Conformer), including a convolution module, a self-attention module, and a classifier module. Image source: [Song et al., 2022]

The proposed architecture is an end-to-end trainable deep neural network based on the EEG Conformer, which integrates convolutional neural networks with Transformer-style self-attention mechanisms. This hybrid design enables joint learning of local temporal–spatial EEG features and global temporal dependencies, making it particularly suitable for imagined speech decoding from short-duration EEG trials. The network configuration employed in this study consists of 40 temporal convolutional filters, a Transformer encoder with 4 layers and 8 attention heads, and a dropout probability of 0.5. The model operates on EEG trials of 1.5 s duration sampled at 300 Hz, resulting in 450 time samples per trial.

- Convolution Module

The convolution module follows an EEG-specific design that decomposes two-dimensional convolutions into sequential temporal and spatial filtering stages. The first stage applies temporal convolutional filters with kernel size (1, 25). These filters act as learnable band-pass filters, capturing short-term oscillatory patterns and transient neural dynamics relevant to imagined speech, while preserving the electrode channel dimension. Subsequently, a spatial convolution is applied across all EEG channels using kernels of size (C, 1), where C denotes the number of electrodes. This layer learns spatial activation patterns and inter-channel dependencies, enabling the network to model distributed cortical activity associated with speech imagery.

Batch normalization and a non-linear activation function are applied to stabilize training and enhance representational capacity. An average pooling operation with kernel size (1, 75) and stride (1, 15) is then used to smooth temporal features and reduce the temporal resolution. Given the short trial duration, this pooling strategy produces a compact sequence of feature representations while retaining sufficient temporal granularity for subsequent processing. Dropout with probability 0.5 is applied to reduce overfitting. Finally, the convolutional feature maps are rearranged such that the temporal dimension is interpreted as a sequence of tokens, and the convolutional feature dimension forms the token embeddings. Each token therefore represents a compact temporal–spatial summary of the EEG signal over a short time window.

- Self-Attention Module

To complement the locally constrained receptive field of the convolution module, a self-attention mechanism is employed to capture long-range temporal dependencies across the entire EEG trial. The token sequence generated by the convolution module is processed by a Transformer encoder consisting of 4 stacked self-attention layers. Each layer employs a multi-head attention strategy with 8 parallel attention heads, allowing the model to attend to different temporal relationships and neural dynamics simultaneously. This design enhances representational diversity and enables robust modeling of contextdependent EEG features. Dropout is applied within the attention layers to improve generalization, especially given the limited duration of the EEG trials and the relatively small number of training samples.

- Classification Module

Following the self-attention module, the learned representations are aggregated and passed to a fully connected classification head. The final layer outputs class probabilities corresponding to the imagined speech categories. The entire network is trained end-to-end using a cross-entropy loss function.

5.3 RESULTS

SUBJECT-DEPENDENT

A hyperparameter search was conducted for the EEG Conformer model using 5-fold stratified cross-validation on the training dataset. Three AdamW configurations were evaluated, combining two learning rates and light L2 regularization. Stratification ensured balanced class distributions across folds. For each configuration, the model was trained within each fold using with cross-entropy loss, a batch size of 64, and a maximum of 2000 epochs. Early stopping based on validation loss (patience = 10) was applied. For every fold, the validation accuracy and the number of epochs until convergence were recorded. Mean validation accuracy, standard deviation, and mean convergence epochs were then computed across folds, and the configuration with the highest mean accuracy was selected. Using the best hyperparameter setting, a final model was trained on the full training dataset for the mean number of epochs observed during cross-validation. Model performance was finally evaluated on an independent held-out test set, reporting overall classification accuracy.

Within-session (13-class)

<u>DAY1</u>

Figure 5.2 shows that the EEG Conformer attains an overall mean accuracy of 7.38% on Day 1, comparable to EEGNet. The corresponding weighted accuracy (Figure 5.4) has an overall mean of 0.29, indicating a similar class-dependent performance pattern.

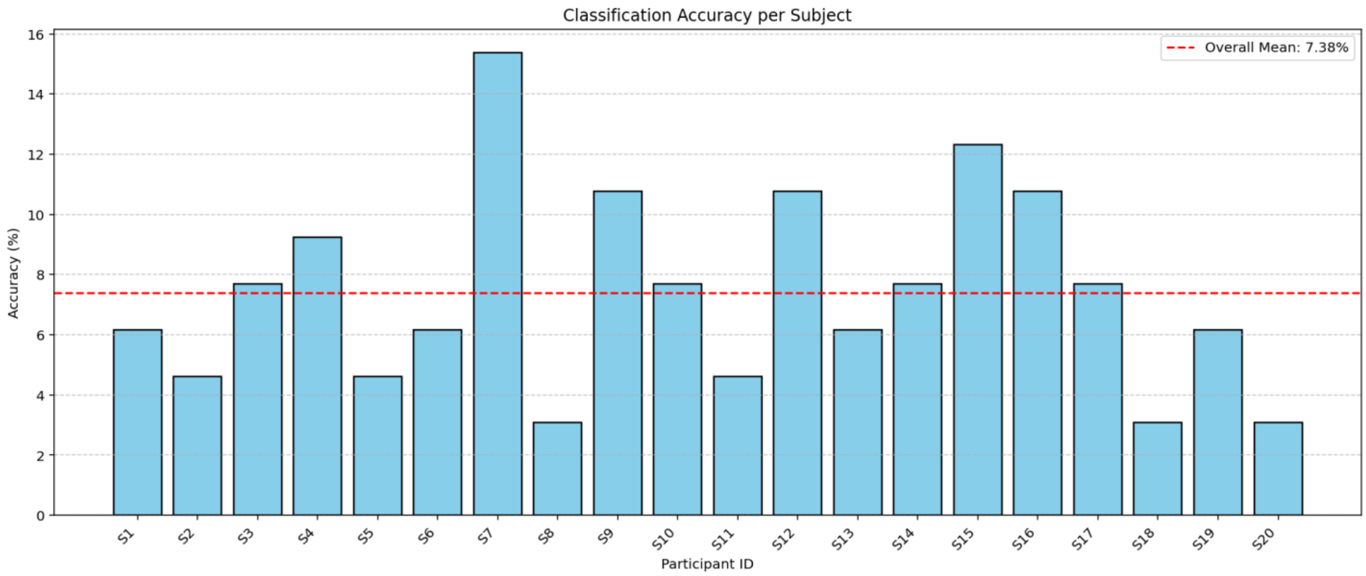

Fig. 5.2 EEGConformer classification accuracy per subject for DAY1.

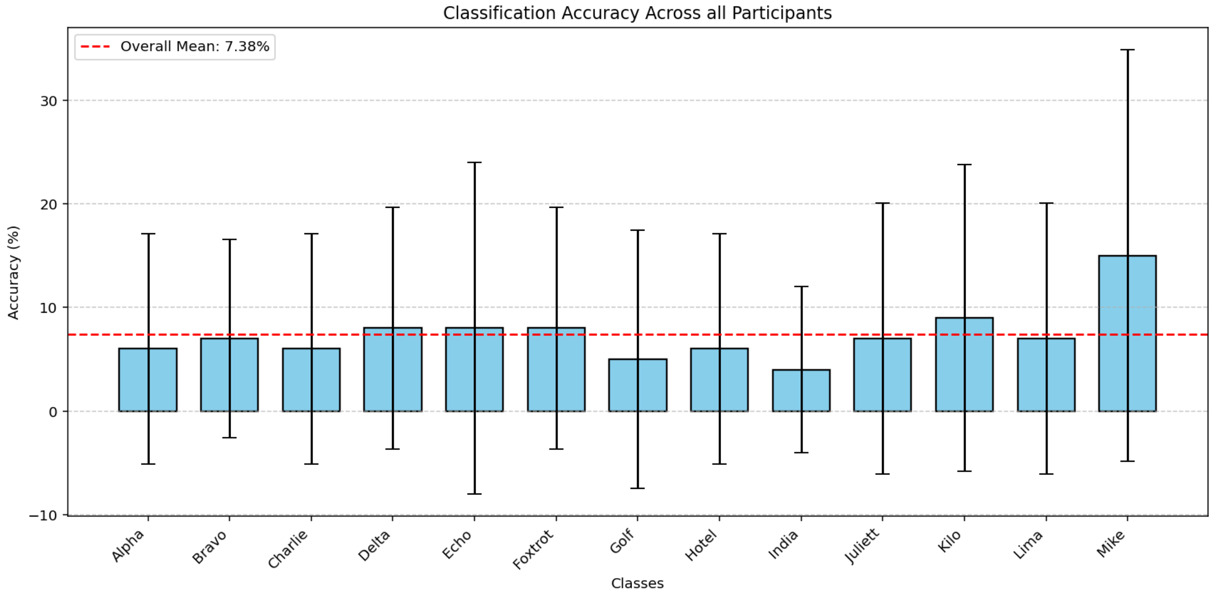

Fig. 5.3 EEGConformer classification accuracy per class across all subjects for DAY1.

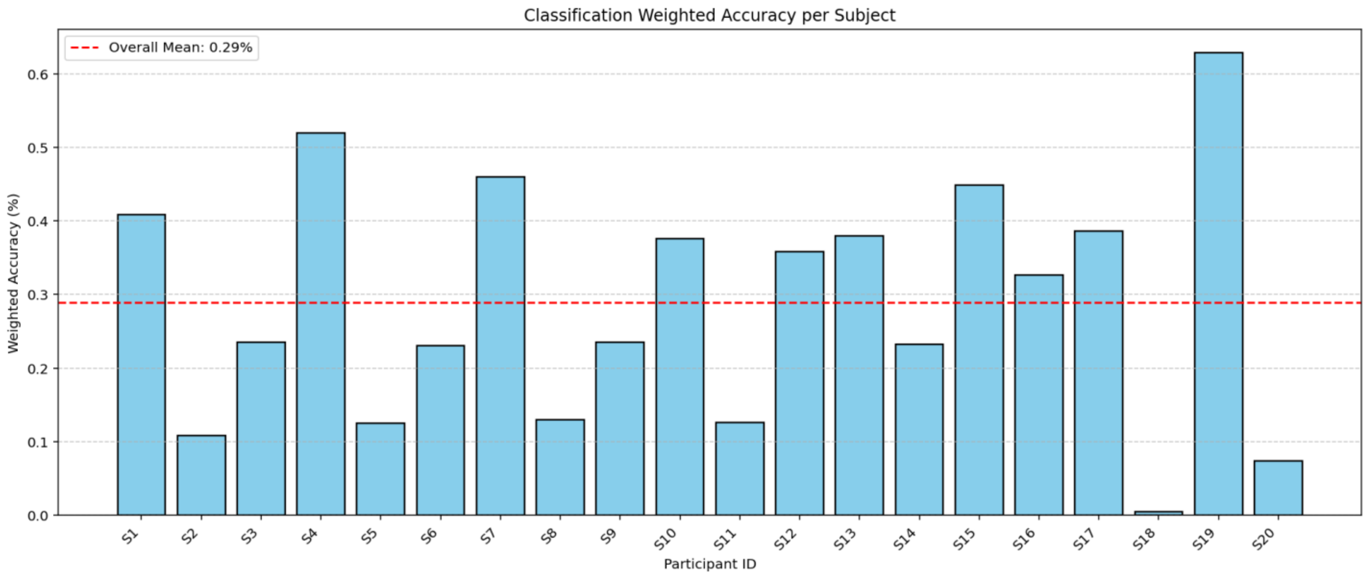

Fig. 5.4 EEGConformer classification accuracy weighted by letter importance per subject for DAY1.

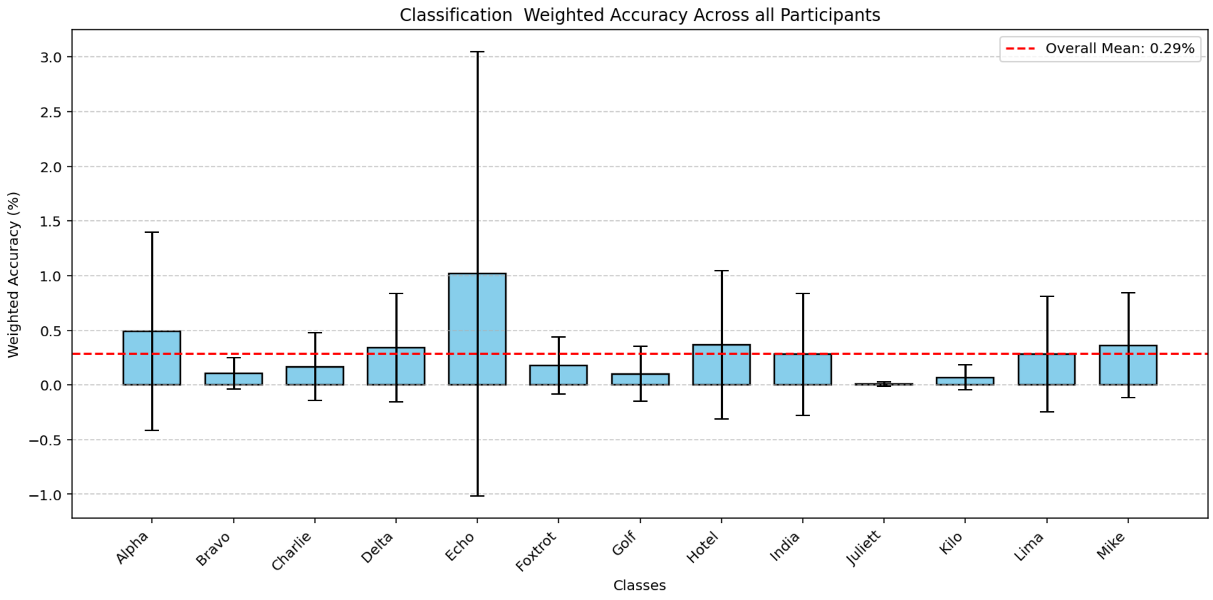

Fig. 5.5 EEGConformer classification accuracy weighted by letter importance per class across subjects for DAY1.

DAY2

As illustrated in Figure 5.6, performance improves on Day 2, with the overall mean accuracy increasing to **8.15%**. This represents the highest within-session accuracy observed across both models.

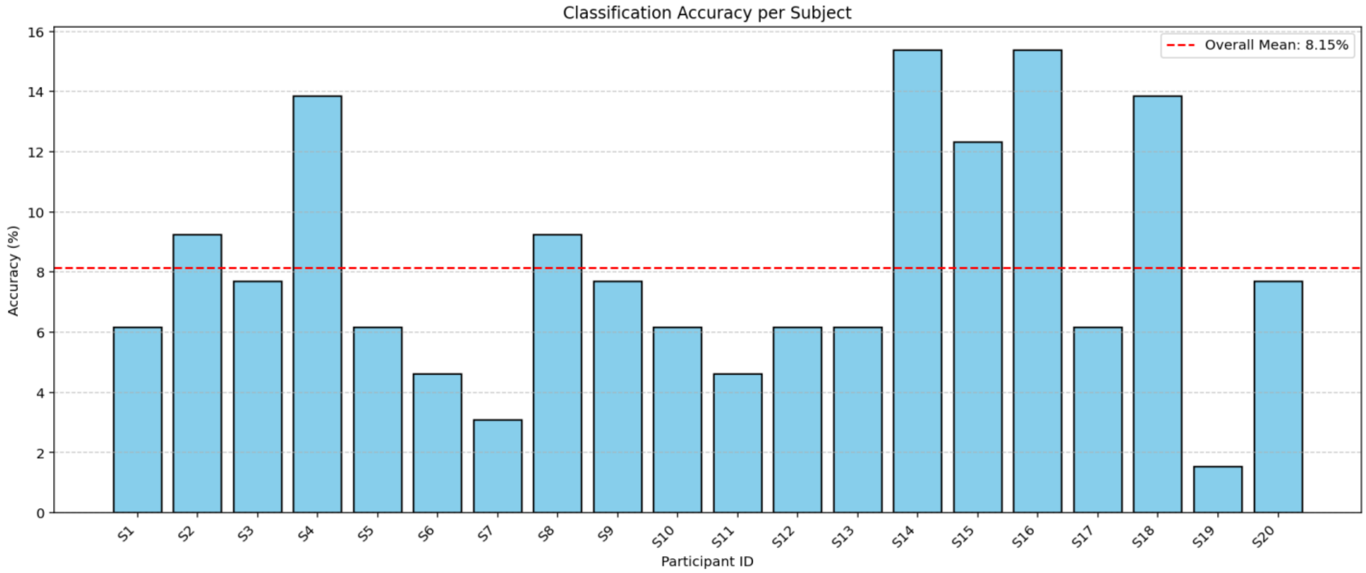

Fig. 5.6 EEGConformer classification accuracy per subject for DAY2.

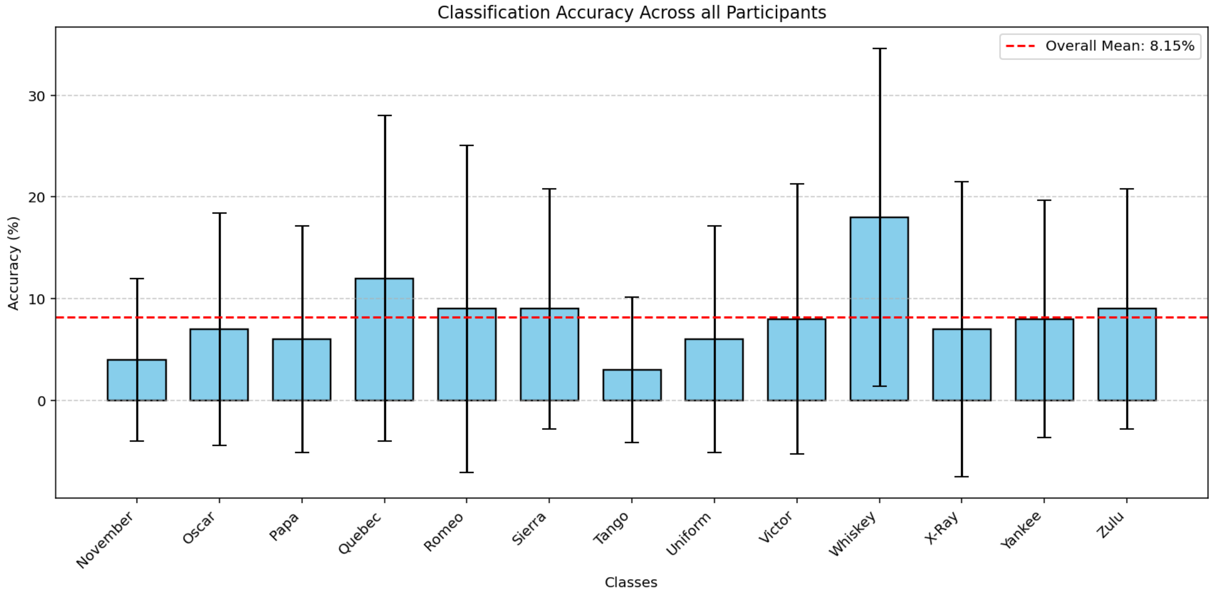

Fig. 5.7 EEGConformer classification accuracy per class across subjects for DAY2.

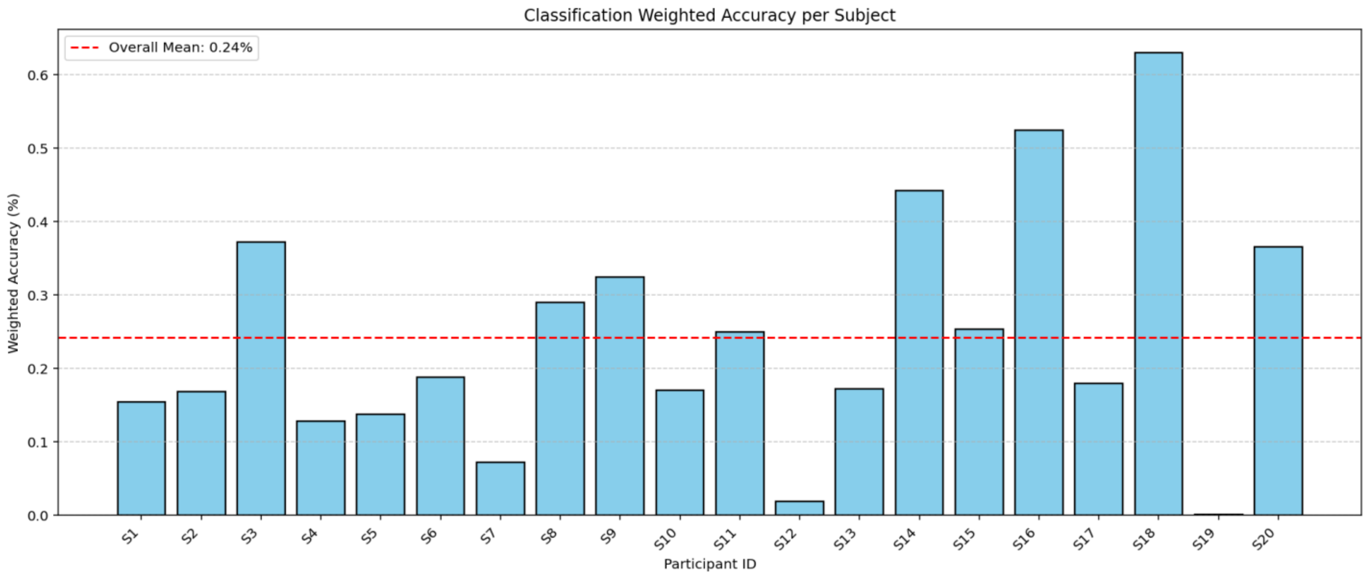

Fig. 5.8 EEGConformer classification accuracy weighted by letter importance per subject for DAY2.

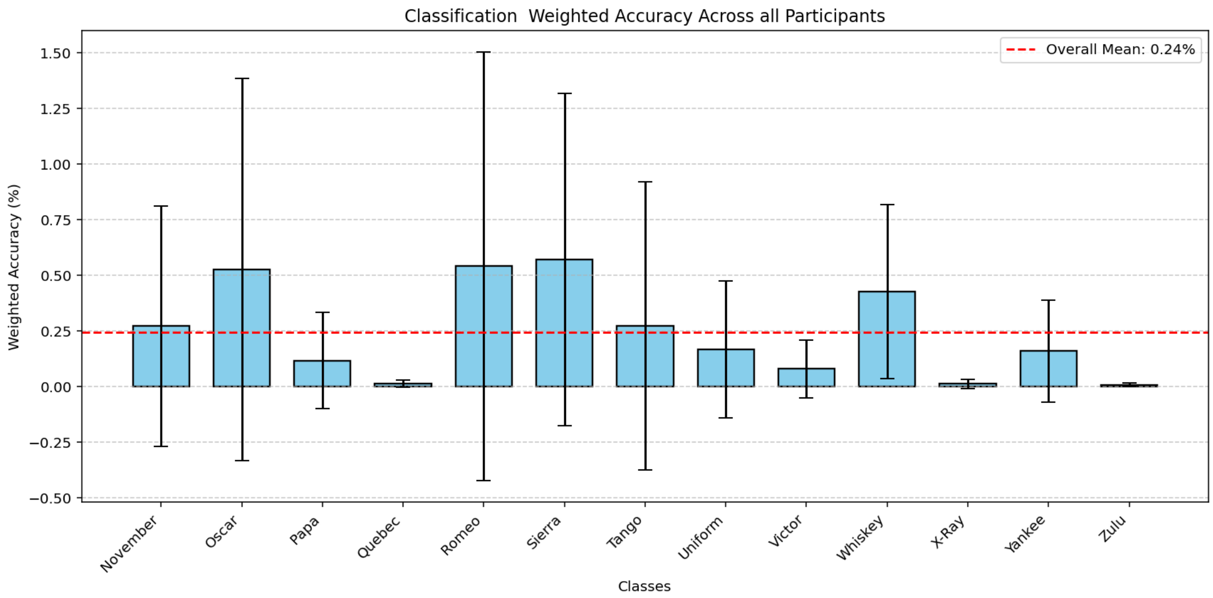

Fig. 5.9 EEGConformer classification accuracy weighted by letter importance per class across subjects for DAY2.

CROSS-SESSION (26-CLASS)

Figure 5.10 depicts EEG Conformer performance in the cross-session setting. The overall mean accuracy reaches **4.15%**, exceeding the corresponding EEGNet result. The weighted accuracy also increases to **0.24**, suggesting improved robustness to session-to-session variability.

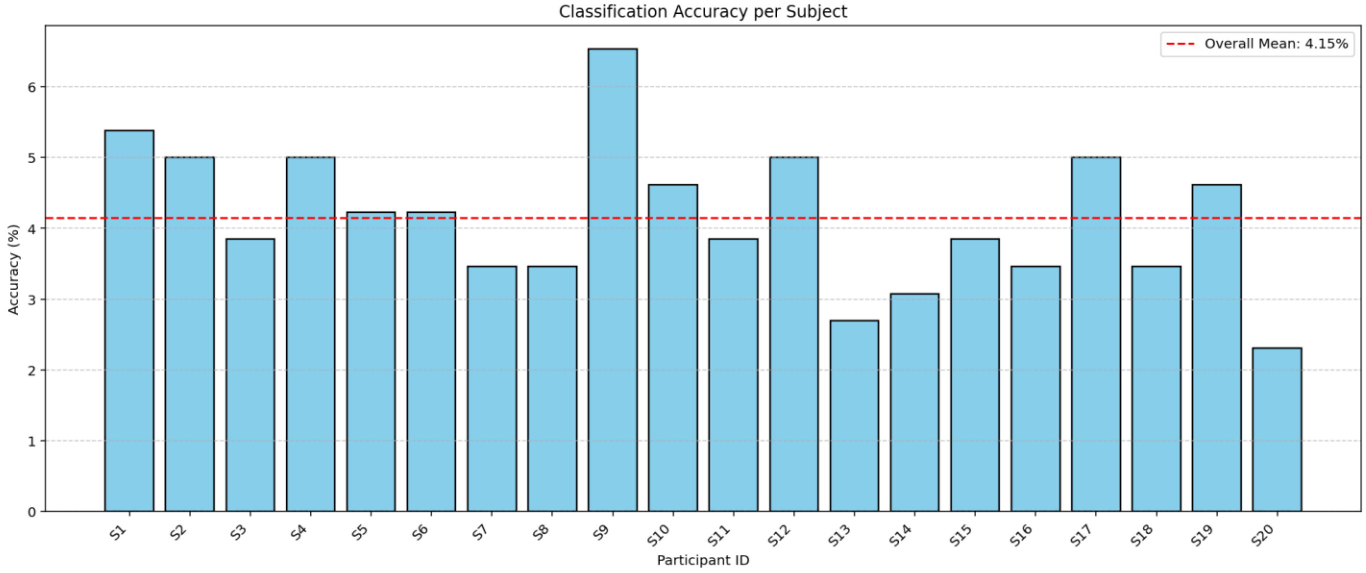

Fig. 5.10 EEGConformer classification accuracy per subject, across all sessions (train on DAY1-2, test on DAY3).

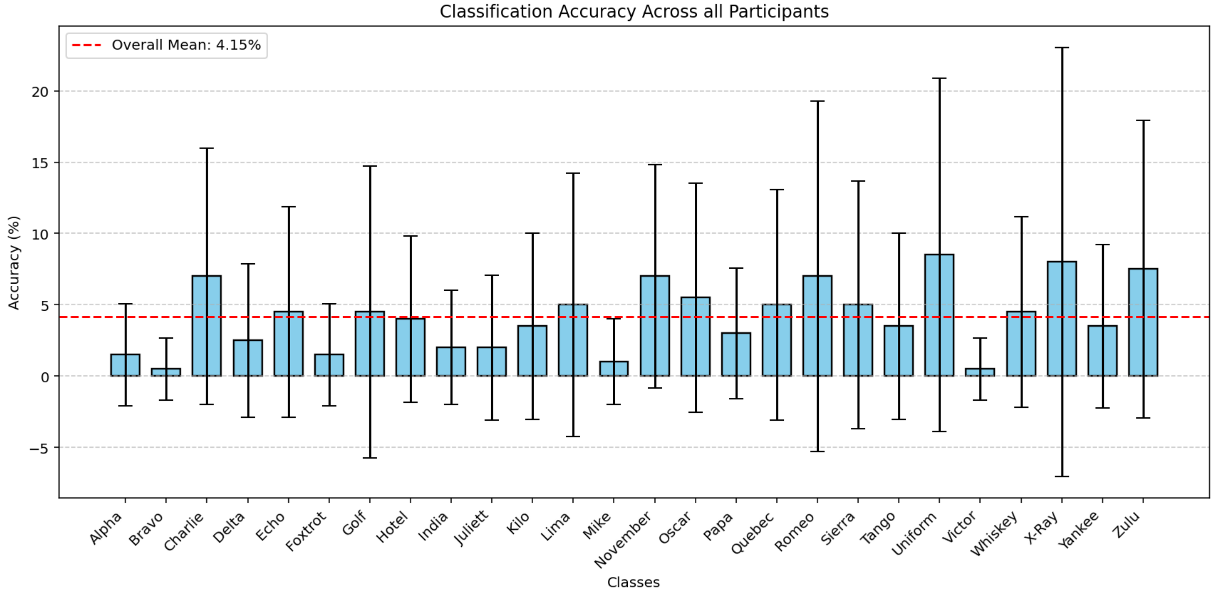

Fig. 5.11 EEGConformer classification accuracy per class across subjects, across all sessions (train on DAY1-2, test on DAY3).

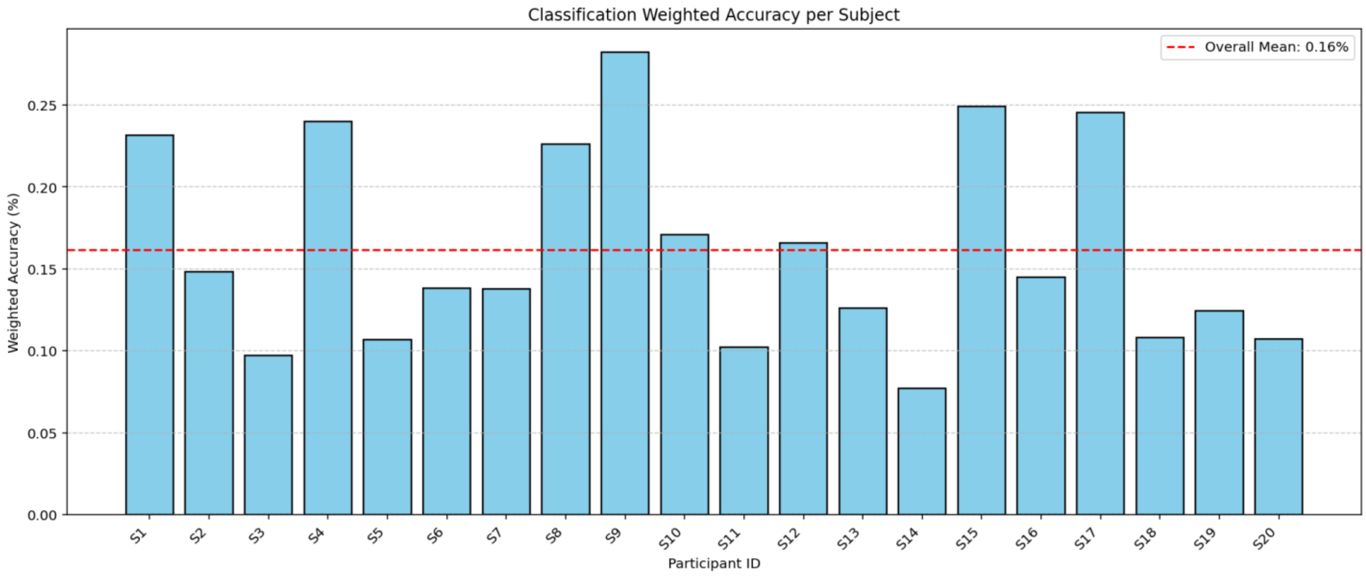

Fig. 5.12 EEGConformer classification accuracy weighted by letter importance per subject, across all sessions (train on DAY1-2, test on DAY3).

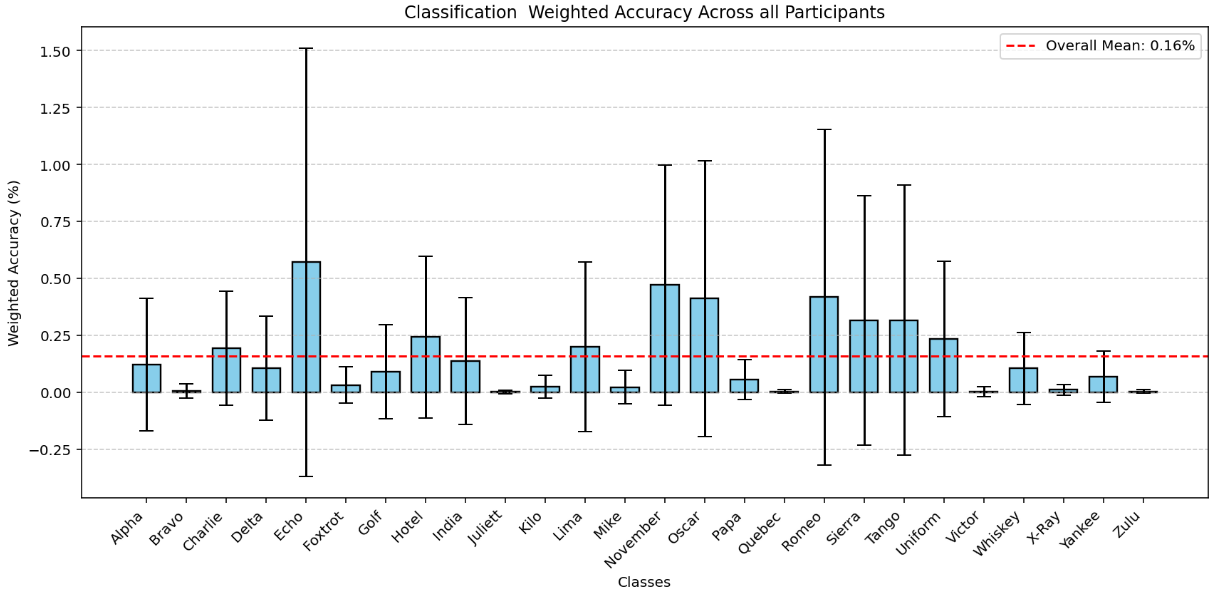

Fig. 5.13 EEGConformer classification accuracy weighted by letter importance per class across subjects, across all sessions (train on DAY1-2, test on DAY3).

LOSO CV (26-CLASS)

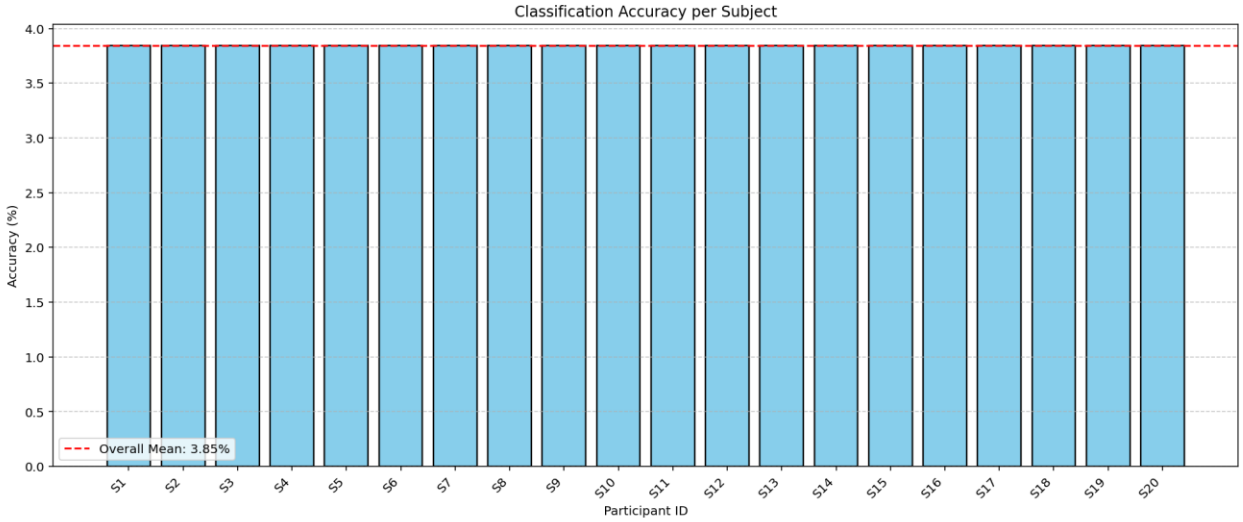

Fig. 5.14 EEGConformer classification accuracy per subject, in a Leave-One-Subject-Out manner, across all sessions.

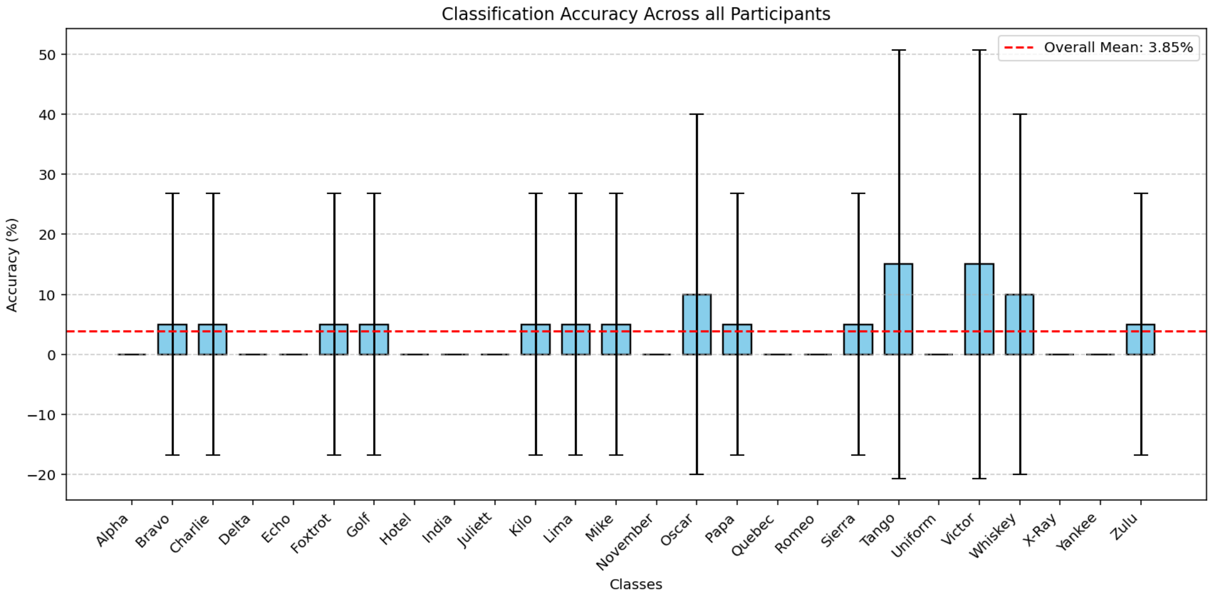

Fig. 5.15 EEGConformer classification accuracy per class across all subjects evaluated in a Leave-One-Subject-Out manner, across all sessions.

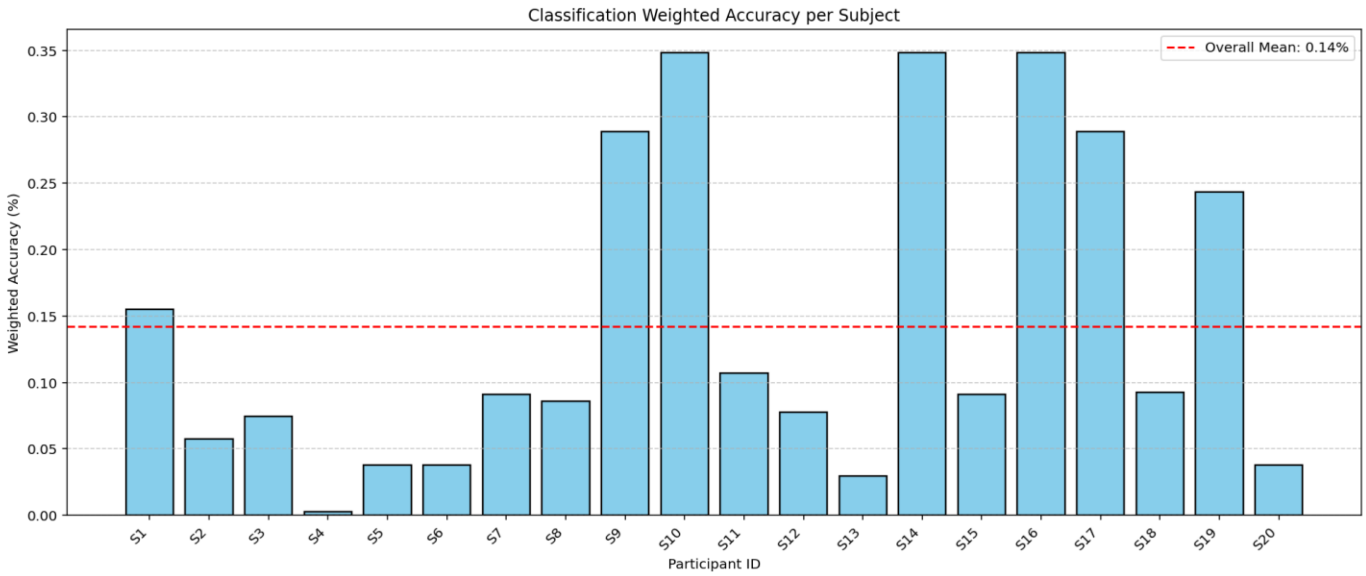

Fig. 5.16 EEGConformer classification accuracy weighted by letter importance per subject, in a Leave-One-Subject-Out manner, across all sessions.

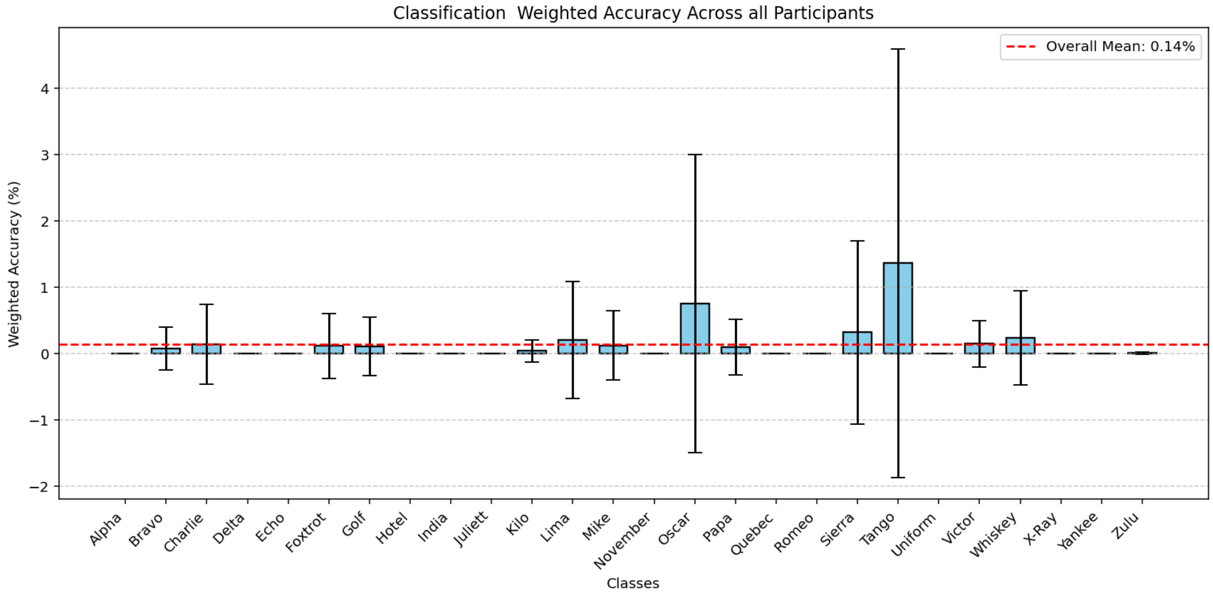

Fig. 5.17 EEGConformer classification accuracy weighted by letter importance per class across all subjects evaluated in a Leave-One-Subject-Out manner, across all sessions.

MC-CV (13-CLASS)

As shown in Table 5.1, the EEG Conformer reaches a mean accuracy of **7.70%** with a weighted accuracy of **0.29**, closely matching EEGNet under the same reduced-vocabulary setting.

Class	Accuracy %	Weighted Accuracy %
Alpha	6.6	0.53
Bravo	7.60	0.11
Charlie	8	0.22
Delta	7.8	0.33
Echo	5.8	0.73
Foxtrot	10	0.22
Golf	8	0.16
Hotel	6.97	0.42
India	6.21	0.43
Juliet	8.2	0.01
Kilo	8.4	0.06
Lima	8.8	0.35
Mike	7.6	0.18
Mean	**7.7**	**0.29**

Table 5.1 EEGConformer classification performance (unweighted and weighted) per class (train on DAY1, test on Day3) evaluated in MC-CV manner.

6. CONCLUSIONS AND FUTURE WORK

This work presented a structured evaluation of imagined speech decoding from EEG signals, with emphasis on understanding how model architecture and evaluation protocol influence performance. Two deep learning approaches, EEGNet and EEG Conformer, were systematically assessed under within-session, cross-session, and subject-independent settings using a multi-session dataset collected within the BINGO project.

The results confirm that imagined speech decoding remains a highly challenging task. Performance is highest under within-session subject-dependent conditions, degrades when models are required to generalize across sessions, and approaches chance level in fully subject-independent evaluations. The EEG Conformer consistently matches or slightly outperforms EEGNet, particularly in cross-session and subject-independent settings, suggesting that attention-based mechanisms offer advantages in modeling temporally distributed neural activity. Nevertheless, the absolute performance levels highlight the strong impact of inter-subject variability and the limited transferability of learned representations.

Future work will extend in two complementary directions. First, we will further investigate the dataset collected within this project, leveraging its multi-session and multi-subject structure to study learning effects, session-dependent variability, and individual imagined speech strategies. Such analyses are expected to provide deeper insight into the neural dynamics of imagined speech.

Second, future methodological efforts will move toward adaptive and incremental learning paradigms, enabling models to evolve as new data are acquired rather than relying on static training. Continual learning, subject-aware adaptation, and representation learning strategies that promote invariance across subjects and sessions could also be explored.

REFERENCES

Brigham, K., & Kumar, B. V. (2010). Imagined speech classification with EEG signals for silent communication: a preliminary investigation into synthetic telepathy. In 2010 4th International Conference on Bioinformatics and Biomedical Engineering (pp. 1-4). IEEE.

Cooney, C., Folli, R., & Coyle, D. (2018). Neurolinguistics research advancing development of a direct-speech brain-computer interface. IScience, 8, 103-125.

Kothe, C. A. E., & Jung, T. P. (2016). U.S. Patent Application No. 14/895,440.

Lawhern, V. J., Solon, A. J., Waytowich, N. R., Gordon, S. M., Hung, C. P., & Lance, B. J. (2018). EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. Journal of Neural Engineering, 15(5).

Martin, S., Brunner, P., Holdgraf, C., Heinze, H.-J., Crone, N.E., Rieger, J., Schalk, G., Knight, R.T. and Pasley, B.N. (2014) Decoding spectrotemporal features of overt and covert speech from the human cortex. Front. Neuroeng. 7:14.

Sereshkeh, A. R., Trott, R., Bricout, A., & Chau, T. (2017). EEG classification of covert speech using regularized neural networks. ACM Transactions on Audio, Speech, and Language Processing. IEEE.

Song, Y., Zheng, Q., Liu, B., & Gao, X. (2022). EEG conformer: Convolutional transformer for EEG decoding and visualization. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 31, 710-719.

Wearable Sensing, DSI-24. Available: https://wearablesensing.com/dsi-24/ [Accessed: 5-Jan-2026].