Unpacking the Success of Riemannian Tangent Space Decoding for Speech Imagery

1st Alberto Tates
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
0009-0008-6107-8975

2nd Ana Matran-Fernandez
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
0000-0002-8409-3747

3rd Sebastian Halder
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
0000-0003-1017-3696

4th Ian Daly
School of Computer Science and Electronic Engineering
University of Essex
Colchester, UK
0000-0001-5489-0393

***Abstract*—Speech Imagery (SI) is a highly intuitive Brain-Computer Interface (BCI) paradigm; however, while its popularity continues to increase, its current feasibility remains a subject of debate. In this work, we 'unpack' the performance of Riemannian Tangent Space (TS) projection features classified with Logistic Regression (LR) to determine whether the pipeline's efficiency in decoding SI from electroencephalography (EEG) reflects speech-imagery-specific activity or signals that merely covary with the task. *Methods:* We evaluated this pipeline on two open-access SI datasets that have fundamentally shaped the research field. We introduced a temporal sanity check to distinguish genuine task-related signals from the leakage induced by block-design paradigms. We then examined the spatial distribution of the LR weights, together with the independent components projecting onto the most heavily weighted channels, to assess the physiological origin of the discriminative features, and tested the pipeline's invariance to drastically reduced feature sets to rule out 'shortcut' learning from isolated sources. *Contribution:* We identify two sources of performance inflation that lead to a systematic overestimation of SI feasibility. First, evaluation strategies that ignore trial structure inflate accuracies, and sustained above-chance performance outside the imagery window shows that this leakage can mimic task-related features. Second, in the highest-performing condition the contrasted words differ in duration and in the visual cue presented throughout the trial, so that both muscular and cortical signals separate the classes without reflecting inner speech; a single channel and frequency band suffice to decode above chance. These findings prompt a reevaluation of feasibility claims in SI research.**

***Index Terms*—Speech Imagery, Brain Computer Interfaces, Tangent Space Projection, Riemannian Geometry**

I. Introduction

Speech Imagery (SI) is envisioned as a highly intuitive Brain-Computer Interface (BCI) paradigm and an ideal candidate for speech neuroprostheses. Although it has gathered increasing attention in recent years, a lack of real-time decoding approaches reflects underlying replicability challenges, leaving the actual feasibility of the paradigm under debate [1, 2]. In this work, we approach SI feasibility by investigating the success of Riemannian Tangent Space (TS) projection of covariance matrices of electroencephalography (EEG) signal. This pipeline has not only proven capable of replicating results in motor imagery datasets [3], but has also demonstrated its utility in representing channel relationships on SI signals [2, 4, 5]. Specifically, literature shows that TS feature vectors are effectively discriminable using Logistic Regression (LR)—particularly with L1 regularization [2, 3, 6]. By aggressively discarding non-relevant dimensions, L1-LR finds the simplest hyperplane to isolate discriminative variance. We aim to rigorously audit this pipeline to determine whether these selected features originate from true task-related neural sources. Non-task-related signals, such as temporal drifts or high-frequency muscle artifacts, are present in EEG and frequently correlate with trial structures [7]. Consequently, data-driven algorithms exploit these artifacts as classification shortcuts.

To unpack the TS-LR pipeline, we evaluate its performance on two influential datasets. First is the Nguyen [4] dataset, where the authors first approached SI decoding using Riemannian geometry to obtain high pairwise classification accuracies. This dataset has exerted a strong influence on the field, surpassing 370 citations. Secondly, we analyze the speech imagery dataset made publicly available by the 2020 International BCI Competition [8]. Traditionally, BCI competition datasets have helped standardize decoders [9]. This specific dataset is increasingly used to explore novel decoding approaches [10–12]. Both datasets were recorded using repetition protocols with block designs, which leads to higher scores compared to other experimental paradigms [6]. However, in block-like designs, trials may have temporal correlations with others from the same block; thus, random cross-validation can lead to artificially inflated results [13].

In this work, we audit the TS+LR pipeline along three lines of evidence. First, we ask whether its accuracy survives the absence of the task: a classifier that performs comparably outside the imagery window cannot be attributed to SI intent,

and its features must instead reflect the temporal structure of the block design. Second, we ask whether the features it selects are physiologically plausible as inner speech, or whether they instead carry the spatial and spectral signature of muscular and ocular contamination [14]. Third, we ask whether performance depends on integrating variance across the distributed cortical networks that language processing recruits [15], or whether it can be reproduced from a single isolated feature. Together, these tests 'unpack' the TS-LR classifier and determine if the decoding reflects SI neural activity.

II. Methods

A. Speech Imagery Datasets

BCIComp [8]: Recorded from 15 participants (64-electrode BrainAmp system) for the 2020 International BCI Competition [8], this dataset originally comprised five imagined phrases. Each trial included an auditory cue followed by four 2-s imagery repetitions (60 trials/class). Based on our previous experimental benchmarks [6], we focused exclusively on the two best-performing words: 'hello' and 'thank you'. Critically, the trials were randomized prior to publication, obscuring the session's temporal structure. For full protocol and timing details, see [8].

Nguyen [4]: This dataset consists of EEG signals from 15 subjects (64-electrode BrainProducts system at 1000,Hz) performing rhythmic speech imagery of vowels and words . Each of the 1000 trials per class was paced by four auditory beeps at 1.4-s intervals. We compared all pair-wise conditions: long versus short words ('in', 'cooperate'), short words ('in', 'out'), vowels (/i/ vs /u/) and long words ('cooperate', 'independent'). Although the trial blocks were scrambled before publication, the internal chronological order of trials within each block was preserved. Refer to [4] for further details on word groupings and rhythmic pacing.

B. Preprocessing

The EEG signals for both datasets were provided as epoched data, which we imported into MNE-Python [16] for subsequent analysis. All epochs underwent visual inspection, and those exhibiting severe motor artifacts were discarded, resulting in a 1–8% rejection rate across subjects. Prior to public release, the Nguyen dataset was preprocessed with a bandpass filter (4–80 Hz) and Independent Component Analysis (ICA) for eye-blink removal [4]. For the BCIComp dataset, we applied a 60 Hz notch filter and utilized the Picard ICA algorithm to isolate ocular artifacts. Based on the power spectral density (PSD) and spatial component distribution, one to two artifact-related components were removed per subject prior to signal reconstruction.

C. Filter-Bank TS+LR Decoding pipeline

The tangent space projection with logistic regression (TS+LR) is a Riemannian geometry-based method that has proven successful in SI replication studies [2]. We extended this approach with a filter bank step. The filter bank is defined with an 10 Hz bandwidth and a 8 Hz step, covering 2–128 Hz for BCIComp and 4-80 Hz for Nguyen. For each frequency band, regularized covariance matrices were computed using Ledoit–Wolf estimation [17]. These matrices were then projected into the tangent space using the PyRiemann library [18]. Each resulting vector has a dimensionality of $n(n+1)/2$, where n is the number of channels. Finally, the tangent space vectors from all frequency bands were concatenated and used to train a Logistic Regression (LR) classifier with an L1 penalty (maximum 600 iterations).

D. Evaluation

We utilized a Group 10-Fold cross-validation strategy. For the Nguyen dataset, we divided the continuous imagery period into three overlapping segments of 1.4 s duration with a 0.1 s stride, replicating the data augmentation protocol originally described by the authors [4]. The grouping procedure ensured that all segments originating from a single trial were assigned together to either the training or testing fold to avoid leakage.

The BCIComp dataset was released with a fully randomized trial order so the original chronological order decoupled prior to publication. Consequently, we implemented a "pseudo-grouping" for BCIComp by assigning four consecutive trials per class to each group, mirroring the original experimental structure. For both datasets, we also performed a standard Stratified 10-Fold cross-validation to verify invariance to the group cross-validation. Throughout this study, we report the distribution of the pooled classification accuracy calculated across all participants.

E. Sanity Check

We evaluated the classifier on the 500 ms post-imagery window, corresponding to the 2.0–2.5 s mark for BCIComp and the 4.5–5.0 s mark for Nguyen.

F. Dual Spatial-Spectral Feature Selection

We performed a dual-axis feature selection using L1 regularization weights. First, the TS+LR pipeline was trained on the full feature set (Channels $\times$ Frequencies) derived from the training set, and the absolute L1 coefficients were extracted. To rank spatial importance, weights were summed across all frequency bands for each channel. To rank spectral importance, weights were summed across all channels for each frequency band. The classifier was then iteratively restricted, for both training and evaluation, to only the top N channels and top M frequency bands.

G. Physiological Validation of Selected Features

To assess whether the discriminative features had a plausible cortical origin, we examined them along two axes.

First, we plotted topographic maps of the absolute L1 weights for each frequency band. L1-LR acts as an aggressive feature selector in SI, zeroing out over 95% of the tangent space [6]; weights concentrated at the extreme lateral or frontal edges of the montage, particularly above 40Hz, were taken as indicative of electromyographic (EMG) contamination rather than of cortical networks [14].

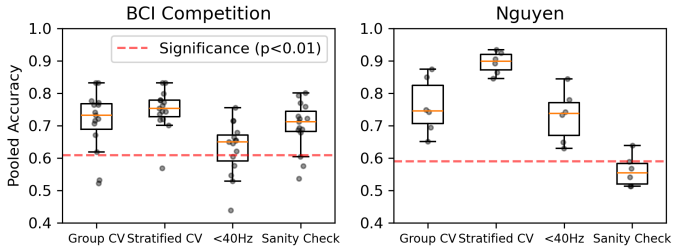

Fig. 1. **Evaluation Strategies and Inter-trial Leakage.** Boxplots display the pooled classification accuracies across all participants for the BCI Competition (left) and Nguyen (right) datasets. The red dashed line indicates the theoretical significance threshold ($p < 0.01$) derived from a binomial distribution. **Group CV** reflects the actual performance on block-like designs, contrasted with standard **Stratified CV**, which exhibits artificially inflated accuracies due to inter-trial leakage. BCIComp performance remains invariant across these two conditions, reflecting the missing trial order and presence of session-wide leakage. The **<40 Hz** condition represents the Clinical Baseline spectral configuration (encompassing Theta, Alpha, and Beta bands), demonstrating performance comparable to the unrestricted baseline for Nguyen and lower scores for BCIComp. The **Sanity Check** evaluates the immediate post-imagery resting window; sustained above-chance accuracy in the BCI Competition dataset further evidences leakage, whereas the Nguyen dataset appropriately degrades to near-chance levels in the absence of the task.

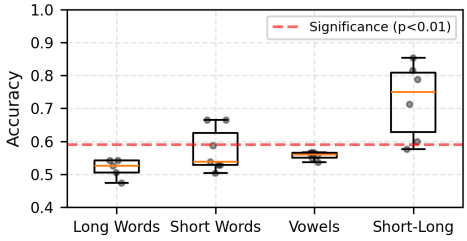

Fig. 2. **Decoding performance comparison across speech imagery conditions in Nguyen dataset** Classification accuracies for Long Words, Short Words, and Vowels tasks remain clustered near chance levels, whereas the Short-Long Words condition exhibits significantly elevated performance (median 0.75)

Second, we inspected the independent components (ICs) projecting most strongly onto the highest-weighted channels. For each participant we ranked channels by summed absolute weight and selected the components whose maximal absolute mixing-matrix coefficient fell on one of the top-ranked channels. For each selected component we examined its scalp projection and its class-wise spectrum, treating focal edge-of-montage topographies and broadband high-frequency power differences between classes as signatures of non-neural origin.

III. Results

A. Impact of Evaluation Strategies and Temporal Leakage

Figure 1 illustrates the classification accuracies across the different evaluation strategies and constraints. For the BCI Competition dataset, performance remained relatively invariant between the Group CV (median: 0.73) and the standard Stratified CV (median: 0.75), though the slight difference was statistically significant ($p < 0.05$). Notably, one participant exhibited a drastic drop in accuracy under the grouped condition. In contrast, the Nguyen dataset demonstrated a pronounced reduction in accuracy when applying the Group CV (median: 0.74) compared to the Stratified CV (median: 0.89), yielding a highly significant difference ($p < 0.01$). Evaluating the immediate post-imagery resting window (Sanity Check) revealed differing outcomes between the two datasets. For BCIComp, the median accuracy slightly decreased to 0.71. While this was significantly lower than the Stratified CV ($p < 0.05$), there was no significant difference when compared to the Group CV ($p > 0.05$). Conversely, the Nguyen dataset exhibited a substantial and highly significant drop during the sanity check (median: 0.55; $p < 0.01$ vs. Group CV), falling below chance level. Finally, the results indicate a clear distinction regarding the spectral locations of the utilized features. When restricting the classifier to the < 40 Hz condition (FS1), BCIComp displayed a highly significant drop in performance (median: 0.65; $p < 0.01$ vs. Group CV), with accuracies falling even below those of the sanity check. In contrast, the Nguyen dataset maintained comparable performance in the lower-frequency condition (median: 0.74), showing only a slight, non-significant change relative to its Group CV baseline ($p > 0.05$).

Comparison of decoding performance across the speech imagery conditions in the Nguyen dataset is illustrated in Figure 2. The short-long words condition achieved the highest mean pooled accuracy, with a median of approximately 0.75, which directly aligns with the findings of the original authors who identified this specific condition as the highest performing in their study. In contrast, the average performance for the other conditions remained clustered below the statistical significance threshold ($p < 0.01$).

B. Non-Neurological Source Evidence

Figure 3 (top row) illustrates the topographic distribution of the spatial weights assigned by the classifier for the two highest-performing participants in the Nguyen dataset. The maps reflect the aggressive feature selection of the L1 regularization, with the majority of electrodes yielding near-zero values across all frequency bands. The highest-weighted features are predominantly localized at the edges of the montage, such as the right temporal and frontal regions for Participant 14, and posterior-occipital locations for Participant 10.

The middle row displays the three independent components projecting most strongly onto the highest-weighted channels for each participant (channels [C6, F8, F4] for Participant 14; [PO8, P8, F8] for Participant 10), together with their class-wise power spectra for the ”cooperate” and ”in” conditions. Grey bands mark the frequency ranges belonging to clusters significant at $p < 0.05$. For Participant 14, the components are localized to frontal and lateral sites; IC1 shows significant clusters in broadly after 20Ḣz, IC2 and IC9 around 40Ḣz. For Participant 10, the selected components are localized to posterior sites; IC1 and IC6 show significant clusters confined at 5–10Ḣz, whereas IC14 additionally shows a cluster at 10Ḣz. In all six components, power for the ”cooperate” condition

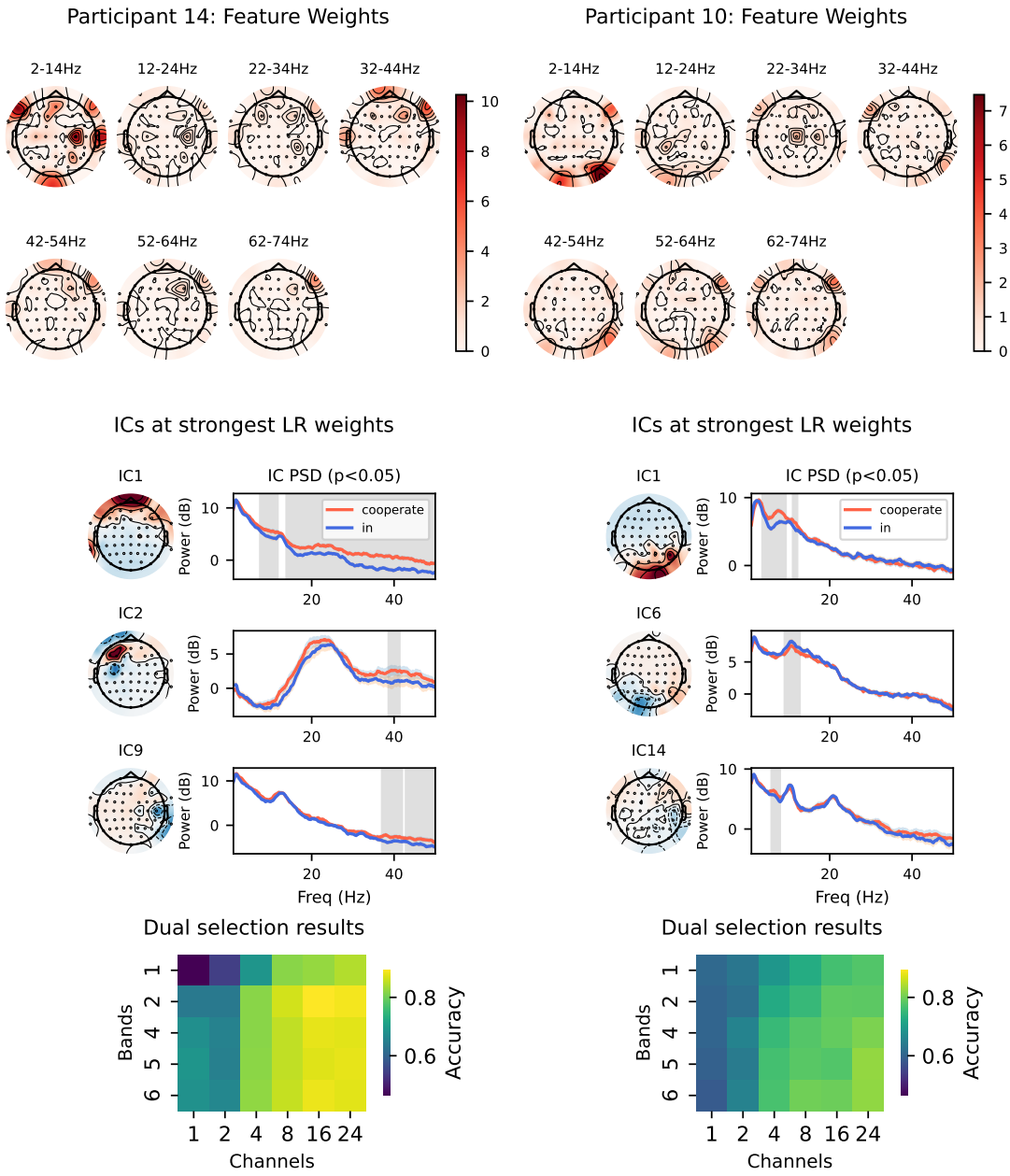

Fig. 3. **Unified spatial, spectral, and decoding analysis for Participant 14 (left) and Participant 10 (right) on long versus short word conditions.** Top: Topographic maps of L1 classifier feature weights across seven frequency bands (2–74,Hz). Middle: Scalp projections and class-wise power spectra of the three independent components projecting most strongly onto the highest-weighted channels, comparing the cooperate" and in" classes; grey bands indicate frequency ranges belonging to clusters significant at ($p < 0.01$)(cluster-based permutation test). Bottom: Heatmaps of classification accuracy from the dual-selector optimization, displaying performance as a function of restricted channel and frequency band counts.

exceeds that for the "in" condition within the significant ranges.

The last row provides the dual-selection accuracy grids, quantifying classification performance at minimal spatial and spectral resolutions. For Participant 10, the model achieves a decoding accuracy of 0.609 using only a single channel and a single frequency band. For Participant 14, while the single-channel/single-band configuration yields an accuracy of 0.464, performance reaches 0.627 when the resolution is increased to two channels and one frequency band.

The pattern observed in these two participants extends to the rest of the group. Figure 4 shows the summed L1 weights across all frequency bands for each of the six participants in the short-versus-long word condition, ordered by decoding accuracy. In four of the six, the weight mass is concentrated at peripheral electrodes. The two exceptions, S01 and S05, show more diffuse distributions extending over central sites, and are also the two lowest-performing participants.

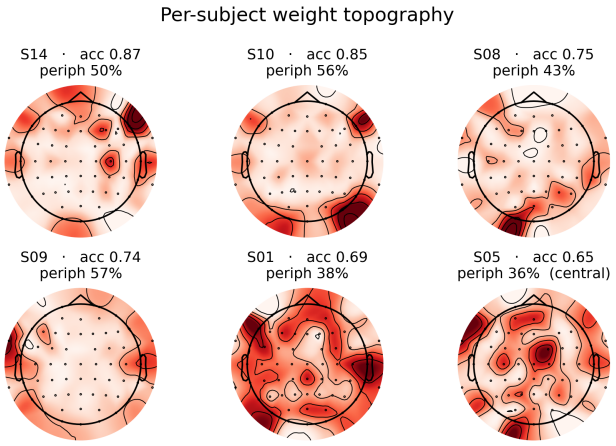

Fig. 4. **Per-subject topography of classifier feature weights on the long versus short word condition for Nguyen dataset.** Topographic maps of the absolute L1 weights, summed across all frequency bands, for each of the six participants in the short-versus-long word group. Participants are ordered by decoding accuracy under Group CV, reported above each map together with the proportion of total weight mass falling on peripheral electrodes. Participants 14 and 10, examined in detail in Figure 3, show the most focal peripheral concentrations, whereas Participants 01 and 05 show more diffuse distributions extending over central sites.

IV. Discussion

Our results demonstrate that improper cross-validation strategies are a primary source of artificial inflation in decoding performance. In the BCI Competition dataset, performance remained invariant between Stratified and Group CV, confirming that the dataset's randomized publication obscured the temporal trial order. The drastic drop to below-chance accuracy for one specific participant under the grouped condition highlights how trial sequencing can distorted the outcomes. The necessity of proper evaluation is most evident in the Nguyen dataset: when overlapping segments from the same trial were not grouped (Stratified CV), accuracies spiked to near 90%. Applying a rigorous Group CV effectively corrected this inflation while proving that true decoding capabilities still remained above chance. The post-task sanity check provides definitive proof of this inter-trial leakage. The BCIComp pipeline maintained high accuracy during non-imagery window, indicating that the classifier was exploiting temporal artifacts rather than SI intent. In stark contrast, performance on the Nguyen dataset appropriately collapsed to chance levels during the sanity check, confirming that its high accuracies are driven by task-related SI. Furthermore, restricting the classifier to the $< 40\,\text{Hz}$ condition demonstrated that the temporal leakage driving the BCIComp results is heavily reliant on higher frequency ranges. Conversely, the authentic features isolated in the Nguyen dataset are predominantly located within the standard, physiologically $< 40\,\text{Hz}$ bands.

The pronounced disparity in decoding performance across the linguistic conditions in the Nguyen dataset is highly revealing. The anomalously high accuracy observed exclusively in the Short-Long Words condition suggests that the classifier is exploiting systematic, task-related artifacts introduced by the contrast in word length, rather than intrinsic neural representations of inner speech. Because imagining a long word inherently takes more time than a short word, this temporal difference creates a structural bias, increasing the probability of duration-dependent physiological confounds occurring systematically in one class over the other.

The topographic distribution of the most discriminative features for the two highest-performing subjects, Participants 14 and 10, indicates that the classifier's success does not rest on speech-imagery-specific activity. In both cases the highest weights are localized to the extreme periphery of the montage—right frontal-temporal channels for Participant 14 and posterior-occipital channels for Participant 10—rather than to the frontal and temporal language regions that inner speech would be expected to recruit [15]. The independent components projecting onto these channels, however, point to two distinct explanations.

For Participant 14, the components are consistent with non-neural contamination. IC1 and IC9 are localized to frontal and right-temporal edge sites and show class differences confined to broad band ranges, while IC2 exhibits a pronounced peak near 40 Hz. Sustained high-frequency power at peripheral electrodes is a well-established signature of muscular tension [14], and the elevation for 'cooperate' relative to 'in' is consistent with tension accumulating over the longer interval required to imagine a polysyllabic word. Notably, these differences are largely absent from the frequency ranges in which cortical speech-related activity is typically reported.

For Participant 10 the picture differs, and the evidence points to a task-correlated confound of plausibly cortical origin rather than to artifact. The selected components are localized to posterior sites, and their class differences are confined to low frequencies and the alpha band (10–20 Hz), with no significant separation at higher frequencies. This profile could involve visual and attentional processing rather than of muscular contamination. It is also directly anticipated by the paradigm: the visual cue specifying the word remained on screen throughout the analysed interval, so the two classes differ not only in the content to be imagined but in the physical stimulus continuously presented to the participant. A sustained difference in posterior alpha is therefore an expected consequence of the design.

Finally, the dual-selection optimization results add further evidence for non-neural source of success. The analysis demonstrates that the classifier can still achieve robust decoding accuracies using a minimal feature space of just one channel and a single frequency band. Inner speech generation relies on the coordination of a distributed cortical network, encompassing frontal and temporal language areas [19]. The fact that the model can successfully separate the classes based on a highly localized, single-feature spatial-spectral pair strongly indicates that the driving signal is a concentrated, task-correlated artifact.

V. Conclusion

In this study, we have evidenced important sources of score inflation that can lead to the overestimation of SI feasibility. While the issue of temporal correlation related to block designs is known in BCI research [20], our analysis highlights how dataset publication formats dictate evaluation integrity. In the case of the Nguyen dataset, although the original timeline of blocks was scrambled, the trials within each block remained sequential. This allowed us to perform Group CV, revealing that the standard CV procedure used by the original authors [4] resulted in significantly inflated accuracies. Conversely, the BCI Competition dataset was published with a completely randomized trial order, making the reconstruction of blocks—and thus a true Group CV—impossible in practice. Consequently, published decoding results on the BCIComp dataset are likely driven by trial temporal leakage that cannot be mitigated.

Our results "unpack" the success of the LR+TS pipeline on two relevant SI datasets, demonstrating its efficiency in aggressively isolating discriminative features from covariance matrices. However, our analysis exposes a critical vulnerability in SI experimental designs: contrasting imagery tasks of varying temporal/stimulus durations can introduce systematic physiological artifacts. As evidenced by our spatial and spectral mapping, these non-neural confounds provide a distinct signal that data-driven algorithms can readily exploit. Given the foundational influence of the Nguyen dataset over speech imagery as a feasible paradigm, our findings prompt a necessary reevaluation of prior high-performance claims. Ultimately, this work underscores the need to critically reassess the feasibility of speech imagery in EEG.

References

[1] A Tates et al. "Speech imagery brain–computer interfaces: a systematic literature review". In: *Journal of Neural Engineering* 22.3 (June 2025), p. 031003. DOI: 10.1088/1741-2552/ade28e. URL: http://dx.doi.org/10.1088/1741-2552/ade28e.

[2] A Tates et al. "Decoding Speech Imagery or Just Noise?: A Symptom of the Replicability Crisis". In: *Journal of Neural Engineering* (). DOI: 10.13140/RG.2.2.26172.55684. URL: http://dx.doi.org/10.13140/RG.2.2.26172.55684.

[3] Sylvain Chevallier et al. "The largest EEG-based BCI reproducibility study for open science: the MOABB benchmark". In: (Apr. 2024). working paper or preprint. URL: https://universite-paris-saclay.hal.science/hal-04537061.

[4] Chuong H Nguyen, George K Karavas, and Panagiotis Artemiadis. "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features". In: *Journal of Neural Engineering* 15.1 (Nov. 2017), p. 016002. DOI: 10.1088/1741-2552/aa8235. URL: http://dx.doi.org/10.1088/1741-2552/aa8235.

[5] Netiwit Kaongoen, Jaehoon Choi, and Sungho Jo. "Speech-imagery-based brain–computer interface system using ear-EEG". In: *Journal of Neural Engineering* 18.1 (Feb. 2021), p. 016023. DOI: 10.1088/1741-2552/abd10e. URL: http://dx.doi.org/10.1088/1741-2552/abd10e.

[6] A Tates et al. "Consolidating the Speech Imagery Paradigm: Evidence that Rhythmic Protocols Drive Superior Decoding Accuracy". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* (). DOI: 10.13140/RG.2.2.21394.11203. URL: http://dx.doi.org/10.13140/RG.2.2.21394.11203.

[7] Eric James McDermott et al. "Artifacts in EEG-Based BCI Therapies: Friend or Foe?" In: *Sensors* 22.1 (Dec. 2021), p. 96. DOI: 10.3390/s22010096. URL: http://dx.doi.org/10.3390/s22010096.

[8] BCI Committee. "2020 International BCI Competition". In: (). DOI: 10.17605/OSF.IO/PQ7VB. URL: https://osf.io/pq7vb/overview.

[9] Michael Tangermann et al. "Review of the BCI Competition IV". In: *Frontiers in Neuroscience* 6 (2012). DOI: 10.3389/fnins.2012.00055. URL: http://dx.doi.org/10.3389/fnins.2012.00055.

[10] Maurice Rekrut et al. "How low can you go: evaluating electrode reduction methods for EEG-based speech imagery BCIs". In: *Frontiers in Neuroergonomics* 6 (July 2025). DOI: 10.3389/fnrgo.2025.1578586. URL: http://dx.doi.org/10.3389/fnrgo.2025.1578586.

[11] Saravanakumar Duraisamy, Maurice Rekrut, and Luis A. Leiva. "Functional Connectivity and Hilbert-Based Features for Covert Speech EEG Variability Analysis and Classification". In: *Interspeech 2025*. interspeech$_2$025. ISCA, Aug. 2025, pp. 2915–2919. DOI: 10.21437/interspeech.2025-1986. URL: http://dx.doi.org/10.21437/Interspeech.2025-1986.

[12] Yasser F. Alharbi and Yousef A. Alotaibi. "Decoding Imagined Speech from EEG Data: A Hybrid Deep Learning Approach to Capturing Spatial and Temporal Features". In: *Life* 14.11 (Nov. 2024), p. 1501. DOI: 10.3390/life14111501. URL: http://dx.doi.org/10.3390/life14111501.

[13] Anne Porbadnigk et al. "EEG-based Speech Recognition - Impact of Temporal Effects". In: *BIOSIGNALS 2009 - Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, Porto, Portugal, January 14-17, 2009*. Ed. by Pedro Encarnação and António P. Veloso. INSTICC Press, 2009, pp. 376–381.

[14] I.I Goncharova et al. "EMG contamination of EEG: spectral and topographical characteristics". In: *Clinical Neurophysiology* 114.9 (Sept. 2003), pp. 1580–1593. DOI: 10.1016/s1388-2457(03)00093-2. URL: http://dx.doi.org/10.1016/S1388-2457(03)00093-2.

[15] Stephanie J. Forkel and Peter Hagoort. "Redefining language networks: connectivity beyond localised regions". In: *Brain Structure and Function* 229.9 (Nov. 2024), pp. 2073–2078. DOI: 10.1007/s00429-024-02859-4. URL: http://dx.doi.org/10.1007/s00429-024-02859-4.

[16] Eric Larson et al. *MNE-Python*. 2025. DOI: 10.5281/ZENODO.15928841. URL: https://zenodo.org/doi/10.5281/zenodo.15928841.

[17] Olivier Ledoit and Michael Wolf. "Improved estimation of the covariance matrix of stock returns with an application to portfolio selection". In: *Journal of Empirical Finance* 10.5 (Dec. 2003), pp. 603–621. DOI: 10.1016/s0927-5398(03)00007-0. URL: http://dx.doi.org/10.1016/S0927-5398(03)00007-0.

[18] Alexandre Barachant et al. *pyRiemann/pyRiemann: v0.3*. 2022. DOI: 10.5281/ZENODO.7547583. URL: https://zenodo.org/record/7547583.

[19] Xing Tian. "Mental imagery of speech and movement implicates the dynamics of internal forward models". In: *Frontiers in Psychology* 1 (2010). DOI: 10.3389/fpsyg.2010.00166. URL: http://dx.doi.org/10.3389/fpsyg.2010.00166.

[20] Ren Li et al. "The Perils and Pitfalls of Block Design for EEG Classification Experiments". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43.1 (2021), pp. 316–333. DOI: 10.1109/TPAMI.2020.2973153.