D2.2 – Report on the developed imagined speech decoding approaches

BINGO

Brain Imaged-Speech Communication

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

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 12, 27/11/2024
Actual date of delivery:	Month 12, 25/11/2024
Work Package:	WP2 Neuroengineering for EEG-based Imagined Speech Decoding
Task:	T2.2 Neuro-informed imagined speech decoding algorithms
Type:	DEC
Approval Status:	Final
Version:	v1.0 (Will be updated on M21)
Number of pages:	28
Filename:	D2.2_Report_Developed_ImaginedSpeech_Decoding_Approaches_v1.docx

Executive Summary: Decoding imagined speech from electroencephalography (EEG) signals represents a significant advancement in brain-computer interface (BCI) research, offering new possibilities for assistive communication and neurorehabilitation. This report presents an in-depth investigation of three distinct decoding approaches: (i) Riemannian Geometry-based analysis for feature extraction and denoising, (ii) a hybrid deep learning framework combining EEGNet with Riemannian Geometry, and (iii) an attention-based EEG Conformer model. These methods are evaluated using publicly available datasets, incorporating neuroscientific principles from the dual-stream model of speech processing to improve decoding accuracy. The findings indicate that Riemannian Geometry methods enhance the discrimination of imagined speech patterns by leveraging spatial covariance matrices, while the EEGNet-Riemannian hybrid improves classification performance through end-to-end feature learning. The EEG Conformer, despite its promise in capturing long-range dependencies, exhibits challenges related to generalization and overfitting. Across all approaches, subject variability and the difficulty of distinguishing phonetically similar words remain significant hurdles. The code for the methodologies described in this report are available at the project's code repository

HISTORY

Version	Date	Reason	Revised by
v0.1	10/10/2024	Table of Contents	Fotis P. Kalaganis
v0.2	05/11/2024	Initial Draft	Fotis P. Kalaganis
v1.0	25/11/2024	Final	Spiros Nikolopoulos

AUTHOR LIST

Organization	Name	Contact Information
CERTH	Fotis P. Kalaganis	fkalaganis@iti.gr
CERTH	Kostas Georgiadis	kostas.georgiadis@iti.gr
CERTH	Vangelis Oikonomou	vicnmu@iti.gr
CERTH	Vasilis Kitsios	vkitsios@iti.gr
CERTH	Nikos Laskaris	laskaris@csd.auth.gr
CERTH	Spiros Nikolopoulos	nikolopo@iti.gr
CERTH	Ioannis Kompatsiaris	ikom@iti.gr

ABBREVIATIONS AND ACRONYMS

BCI	Brain Computer Interface
EEG	ElectroEncephaloGram
SCP	Spatial Covariance Matrices (interchangeable with CCV)
CCV	Cross Covariance Matrices
SPD	Symmetric Positive Definite
CNN	Convolutional Neural Network

Contents

1. INTRODUCTION

Inner speech, the silent verbalization of thoughts, plays a crucial role in cognitive processes such as memory, problem-solving, and self-regulation. Decoding inner speech from brain activity presents significant challenges but holds promise for brain-computer interfaces (BCIs), neurorehabilitation, and communication aids for individuals with speech impairments. Electroencephalography (EEG), a non-invasive neuroimaging technique, is widely used in this context due to its high temporal resolution and ease of application. Decoding inner speech from EEG signals requires advanced machine learning techniques to extract relevant features and classify neural activity patterns. Common approaches include [Lopez-Bernal, 2022]:

- **Feature extraction methods:** Time-domain (e.g., event-related potentials), frequency-domain (e.g., power spectral density), and time-frequency domain (e.g., wavelet transforms) analyses are commonly used to isolate relevant EEG features.
- **Traditional classifiers:** Machine learning models such as support vector machines (SVMs), linear discriminant analysis (LDA), and k-nearest neighbors (k-NN) are applied to distinguish inner speech-related EEG patterns from other brain activity.
- **Deep learning techniques:** Neural networks, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have demonstrated promising results in decoding complex EEG patterns. Hybrid models, such as long short-term memory (LSTM)-CNN architectures, leverage spatial and temporal dependencies for improved accuracy.
- **Transfer learning and domain adaptation:** Due to inter-subject variability in EEG signals, transfer learning techniques are employed to adapt models trained on one subject to new subjects with minimal retraining.

Despite progress, inner speech decoding remains a challenging task due to low signal-to-noise ratio, individual variability, and the subtle neural correlates of inner speech. In order to promote future research and improve decoding accuracy through multimodal approaches, we have developed and tested three different approaches that inherently consider the most prominent neuroscientific model of speech processing, namely, the two stream hypothesis [Hickok, 2022]. This model delineates the roles of two cortical pathways in speech processing:

- **Dorsal stream:** This stream is responsible for sensorimotor integration, linking auditory representations with speech production. It plays a crucial role in mapping sounds to articulatory movements and is heavily involved in phonological processing. Key brain regions include the superior temporal gyrus, premotor cortex, inferior parietal lobule, and the arcuate fasciculus, a critical white matter tract connecting auditory and motor speech areas. The dorsal stream is essential not only for overt speech but also for covert (inner) speech, enabling individuals to internally rehearse and manipulate linguistic representations.
- **Ventral stream:** This pathway is responsible for speech comprehension, facilitating the transformation of auditory input into meaningful linguistic information. It primarily involves the superior and middle temporal lobes, as well as the anterior temporal lobe, which are crucial for lexical and semantic processing. Unlike the dorsal stream, which is more concerned with the mechanics of speech, the ventral stream is engaged in understanding and interpreting spoken and inner language, allowing for internal dialogue and conceptual thinking.

In the context of inner speech, both streams are engaged in an interplay, but the dorsal stream is particularly implicated in maintaining and manipulating phonological representations without overt articulation, while the ventral stream supports the internal generation and understanding of semantic content.

To this end, in this document we present the results from three distinct approaches that may consider such an interplay between different cortical regions inherently. The first approach is based on Riemannian Geometry tools and may serve as a denoising procedure on connectivity brain data (captured by means of covariance matrices). The second approach, presented as a poster in EUSIPCO 2024 as part of BINGO's dissemination activities, corresponds to an end-to-end neural network that combines learnable temporal filters followed also by a learnable module that operates on the multiplexed functional connectivity of temporally filtered brain signals. Finally, the third approach, which is only primarily tested, constitutes a hybrid model between convolutional neural networks (CNNs) and attention mechanisms so as uncover the contextual temporal dependencies that underpin speech processing. It should be noted, that two publicly available datasets are employed so as to examine the effectiveness of these approaches. 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. DATASETS

2.1 KARA ONE

The dataset [Zhao, 2015] consists of 14 participants, with an average age of 27, who were instructed to imagine pronouncing and consequently to speak aloud 7 phonemes or syllables: (/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/) and 4 words: (pat, pot, knew, and gnaw) over the course of 30 to 40 minutes. The participants were seated in front of a computer monitor and a Microsoft Kinect camera and a research assistant placed an EEG cap on their heads. The data collected combine 3 modalities: EEG signals, face tracking and audio. A 64-channel Neuroscan Quick-cap was used, the electrode placement followed the 1020 rule and the data were sampled at 1kHz.

Each trial consisted of 4 states. First, there is a 5-second rest state where the participants were instructed to relax. Next, in the stimulus state, the prompt text appeared on the screen and its corresponding audio played from the speakers. A 5-second imagined speech state follows where the participants imagined pronouncing the prompt without moving their articulators and finally, they spoke the prompt aloud. In this work, we only employed the EEG segments corresponding to imagined speech. The data from 11 out of 14 participants were utilized to maintain uniformity in the number of trials. For each participant, 132 trials were conducted, 12 for each prompt.

2.2 2020 BCI COMPETITION DATASET

In this case [Jeong, 2022], 15 participants, aged between 20-30 years, were instructed to imagine pronouncing five words/phrases, namely: (“hello,” “help me,” “stop,” “thank you,” and “yes”). During the experiment, the subjects were seated in a comfortable chair in front of a 24-inch LCD monitor screen and were asked to solely focus on the given task without moving their articulators nor making any sound. For the recording, 64 EEG electrodes following a 10-20 international configuration were used.

An auditory cue of a randomly chosen prompt is introduced to the participants for 2 s, followed by the visual cue of a cross mark on the screen that lasted between 0.8-1.2 s. The subjects imagined pronouncing the given prompt as soon as the cross mark disappears. During this phase, the participants received no stimulus at all in order to avoid any unwanted brain activity. The cross-mark presentation on the screen and the consequent imagined speech phase (2s) were repeated 4 times in a row for each random cue. Before proceeding to the next word/phrase, the subjects were given 3s to relax and clear their minds.

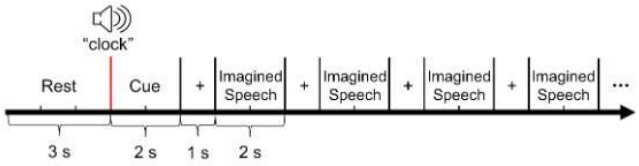

Fig. 2.1. Timeline of the experimental procedure. Image source: https://osf.io/ymvjz.

In total, 400 trials were conducted for each subject, 80 trials per prompt, out of which 300 are dedicated for training, 50 for validation and the remaining 50 for testing.

2. RIEMANNIAN GEOMETRY

2.1 MOTIVATION

Riemannian geometry has emerged as a powerful mathematical framework for EEG-based decoding, including inner speech recognition. Unlike traditional machine learning approaches that rely on feature extraction, Riemannian geometry leverages the structure of spatial covariance matrices (SCP; also referred to as Cross Covariance matrices, CCV), treating EEG data as points on a geometric manifold. This method enhances classification performance by capturing the intrinsic geometry and connectivity of brain signals. By modeling EEG covariance matrices in a Riemannian framework, these approaches reduce sensitivity to noise and enhance discrimination between inner speech states, making them promising candidates for future advancements in BCI applications. In this section we present a methodological approach that can be interpreted as a robust estimation for covariance matrices under the constraint of a common mixing matrix and uncorrelated activity in the source space.

2.2 METHODOLOGY

ELEMENTS OF RIEMANNIAN GEOMETRY

Let $\mathbf{X}_i \in \mathrm{R}^{E\times T}, i = 1,\ldots,n$ be a multichannel EEG segment, where E denotes the number of electrodes, T the number of time samples and n the number of available segments (or trials). Each segment (assuming zero mean signals) can also be described by the corresponding spatial covariance matrix $C_i = \frac{1}{T-1} X_i X_i^{\top} \in R^{E\times E}$, where $(\cdot)^{\top}$ denotes the transpose operator. By definition and under a sufficiently large T value to guarantee a full rank covariance matrix, spatial covariance matrices are Symmetric Positive Definite (SPD) that lie on a Riemannian manifold instead of a vector space (e.g. scalar multiplication does not hold on the SPD manifold). In the field of differential geometry, a Riemannian manifold is a real, smooth manifold endowed with an inner product on the tangent space of each point that varies smoothly from point to point.

When treating EEG data, the manifold of SPD matrices denoted by $Sym^{+}_{E} = \{\mathbf{C} \in \mathrm{R}^{E\times E} : \mathbf{x}^{\top}\mathbf{C}\mathbf{x} > 0$, for all non-zero $\mathbf{x} \in \mathrm{R}^{E}\}$, is typically studied when it is equipped with the AIRM [Conney, 2020],

$$\langle \mathbf{A}, \mathbf{B} \rangle_{\mathbf{P}} \triangleq Trace(\mathbf{P}^{-1}\mathbf{A}\mathbf{P}^{-1}\mathbf{B}) \quad (1)$$

for $\mathbf{P} \in Sym^{+}_{E}$ and $A, B \in T_E^{+}(P)$, where $T_E^{+}(P)$ denotes the tangent space of Sym^{+}_{E} at $\mathbf{P}$. Then, the following geodesic distance is induced

$$\delta(C_i, C_j) = \| logm\left(C_i^{-1/2} C_j C_i^{-1/2}\right) \|_F = \sqrt{\sum_{q=1}^{E} log^2 \lambda_q} \quad (2)$$

where $logm(\cdot)$ denotes the matrix logarithm operator and λ_q the eigenvalues of the matrix $\mathbf{C_i}^{-1/2}\mathbf{C_j}\mathbf{C_i}^{-1/2}$ or equivalently of the matrix $\mathbf{C}_i^{-1}\mathbf{C}_j$. We note that these two matrices are similar (i.e., hold the same eigenvalues) and that the indices i and j can be permuted. Among the other useful properties that are discussed thoroughly in [Pennec, 2006], δ is congruent invariant for non-singular matrices $\mathbf{W}$, i.e. $\delta(\mathbf{W}\mathbf{C}_i\mathbf{W}^{\top}, \mathbf{W}\mathbf{C}_j\mathbf{W}^{\top}) = \delta(\mathbf{C}_i, \mathbf{C}_j)$. This is an important property in EEG signal processing since it provides equivalence between the sensor and the source space [Congedo, 2017]. According to the prevailing EEG model, the recorded activity is well approximated by a linear mixture of source signals. Hence, $\mathbf{X}_i = \mathbf{M}\mathbf{S}_i$ with $\mathbf{M}$ denoting the mixing matrix and $\mathbf{S}_i$ the source signals. Then, by substituting the observed signal with

the equivalent mixing of sources, one may obtain the following covariance matrix, $C_i = \frac{1}{T-1} M S_i S_i^\top M_i^\top$. Therefore, the mixing procedure in the time domain results in a congruent transformation in the corresponding covariance matrices. It becomes obvious that since δ is invariant to such transformations, the two spaces are considered equivalent. In a strict mathematical sense this is partially true (e.g., for certain forms of **M**) and this topic is thoroughly discussed in [Congedo, 2017]. Hereafter, the terms "AIRM-induced geodesic distance" or simply "geodesic distance" will be used interchangeably and will refer to Equation 2.

AJD-BASED COVARIANCE ESTIMATION

The mixing matrix, denoted as **M**, is determined by the position and orientation of dipoles in the brain, the physical characteristics of the head, and the placement of electrodes on the scalp. It is therefore reasonable to assume that **M** remains constant for a certain period, such as during a single recording session. Assuming that sources are independent and the associated activity (i.e., source signals) are uncorrelated, the spatial covariance matrices of the sources are diagonal.

The process of estimating the mixing matrix, denoted as **M**, from the observed sensor signals is an ill-posed problem known as Blind Source Separation (BSS) [Müller, 2004]. Two approaches are commonly used to tackle the BSS problem: The first approach is Independent Component Analysis (ICA), which aims to transform the data so that the components become as independent as possible [Hyvärinen, 1999]. An alternative approach involves using the diagonality of certain characteristic matrices derived from the data to approximate $\mathbf{M}^{-1}$ through the concept of AJD [Ziehe, 2005]. This involves finding an orthonormal change of basis denoted as **U**, which makes the set of symmetric square matrices as diagonal as possible. This, second approach, intuitively uncovers the 'average eigenspace' of matrices that are approximately jointly diagonalizable [Cardoso, 1996].

Following the notation of previous section, we denote by $\mathbf{C}_i$ covariance matrix that corresponds to the EEG trial, $\mathbf{X}_i$. Let **U** be the orthonormal matrix calculated by AJD over the set of $\mathbf{C}_i$ with $i = 1,\ldots,n$ that estimates the mixing matrix **M**. Then, each $\mathbf{C}_i$ can be transformed to a dominantly diagonal matrix through $\mathbf{U}^\top\mathbf{C}_i\mathbf{U}$. As such, we can reconstruct (i.e., re-estimate) all the spatial covariance matrices under the constraint of a common eigenspace by using the formula $\widetilde{C}_i = \mathbf{U}\,diag(\mathbf{U}^\top\mathbf{C}_i\mathbf{U})\,\mathbf{U}^\top$. Here, the $diag(\cdot)$ operator, which discards the nondiagonal elements of a matrix and obtains a strictly diagonal matrix, is applied upon an almost diagonal matrix and hence achieves a good re-estimation of the original covariance matrix.

This estimation approach forces all the spatial covariance matrices to admit a common mixing matrix and, hence, acts as a denoising procedure that abides to well-established neuroscientific theories. In addition, the estimated covariance matrices are guaranteed to hold the SPD property which allows the employment of Riemannian geometry. A more detailed description about the advantages and the mathematical properties of this covariance estimation can be found in [Kalaganis, 2022].

2.3 RESULTS

PRELIMINARY STUDY: SPECTROTEMPORAL ANALYSIS AND SENSOR SELECTION

Taking into account the high subject variability encountered in EEG data, a preliminary analysis for each subject was performed that aimed to identify the exact brain areas (i.e. sensors), timing (i.e. trial segments) and spectral components (i.e. frequency ranges) that the phenomenon of imagined speech takes place, with the scope of decoding the underlying phenomenon in the best possible way. In this context, a wavelet filter bank approach that disentangles the input signal into multiple frequency components without losing the signal's temporal characteristics is employed. It is noted that wavelets are characterized by time locality, allowing an efficient capture of transient behavior in a signal, which is of essence in the case of imagined speech decoding. Working on the training set for each subject

independently, we applied the continuous wavelet transform (FBCWT, based on morse wavelet function and Matlab filter bank implementation) within the [1-100]Hz frequency range and derive the associated scalogram for each trial separately. Following the aforementioned procedure, all single-trial scalograms were averaged, regardless their label, to derive a spectrotemporal profile of activation for every sensor. Finally, using the baseline period, the mean and std of each scale was estimated and used to derive a threshold value (mean+3std) that in turn was employed to reveal the significant event-related spectral perturbations. The process is completed with detection of the sensors, segments and frequencies of interest based on the thresholding process.

Fig.2 illustrates the averaged FBCWT patterns for an indicative set of sensors for an exemplar subject (i.e. subject S1), after the thresholding process is completed. It is important to note here, that for clarity purposes only a selected number of sensors is presented. The visual inspection of the figure provides answers regarding the three research questions posed in this subsection. Starting from identification of the brain areas that the imagined speech phenomenon takes place, it is evident that the most informative sensors are located over the Broca's area (e.g., FT7, FT9 and T7), a trend that aligns well with what is reported in relevant bibliography regarding the brain areas activated during the task of imagined speech [Si, 2021]. On the contrary, the activation levels on sensors located over areas that are not associated with the mental speech task, like the middle area (e.g., sensors Pz, CPz and Fz), is significantly lower. Moving to the temporal domain, it is obvious that a reaction period of approximately 500ms is required before the mental imagery process is initiated by the participant, which is typical, while varying among individuals, when cue-based triggers mark the initiation of a task. Consequently, this process, upon appropriate modifications, can be employed as an onset detection procedure, which is of paramount importance in self-paced and online BCI paradigms. In the spectral domain, and specifically for the sensors characterized by high activity (such as FT7, FT9 and T7), three frequency ranges of interest can be identified: (i) Low ([5-20]Hz), (ii) Medium ([40-55]Hz) and, (iii) High ([≥70]Hz), with the High frequency range being empirically identified, based on the validation set, as the one with the highest discriminative power. Finally, we should note that while the trends observed for subject S1 are similar for the other subjects, the exact optimal sensors, segments and frequencies, as expected differ among them, showcasing the necessity and importance of this preliminary study.

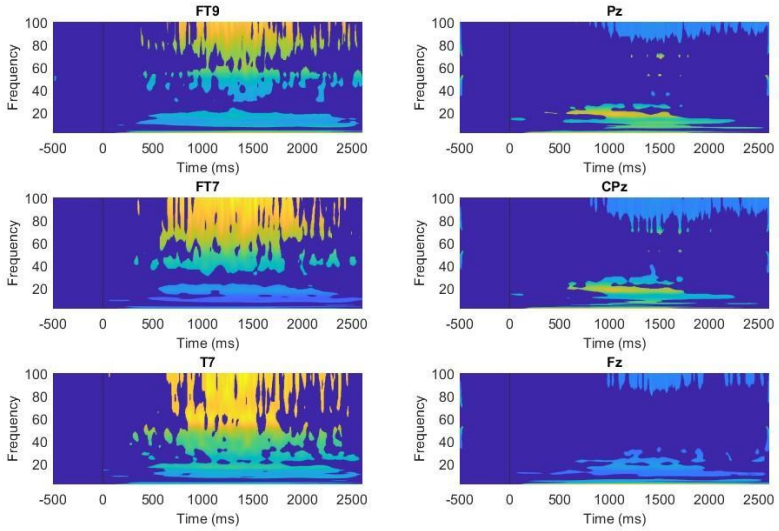

Fig.3.1. Spectrotemporal analysis for the sensors characterized by the highest (left panel) and lowest (right panel) activation levels. The stimulus onset is indicated by the black vertical identified at t=0s and corresponds to cross disappearance as depicted in Fig. 2.1.

CLASSIFICATION RESULTS

Fig.3 presents the overall accuracy of the proposed decoder (Fig.3.2A, Fig.3.2B) and also the accuracy scores obtained for each subject independently (Fig.3.2C). In particular, the employed decoder is based on the estimation of covariance matrix (as described in section 3.2) from EEG signals in the frequency range above 70Hz while employing a Riemannian k-NN classifier where distance is calculated according to Eq. 2. By exploiting the validation set for each subject independently, k = 3 was identified as the most suitable value in terms of accuracy. We note that the provided test set is employed only for the purposes of obtaining and reporting the classification performance in this section.

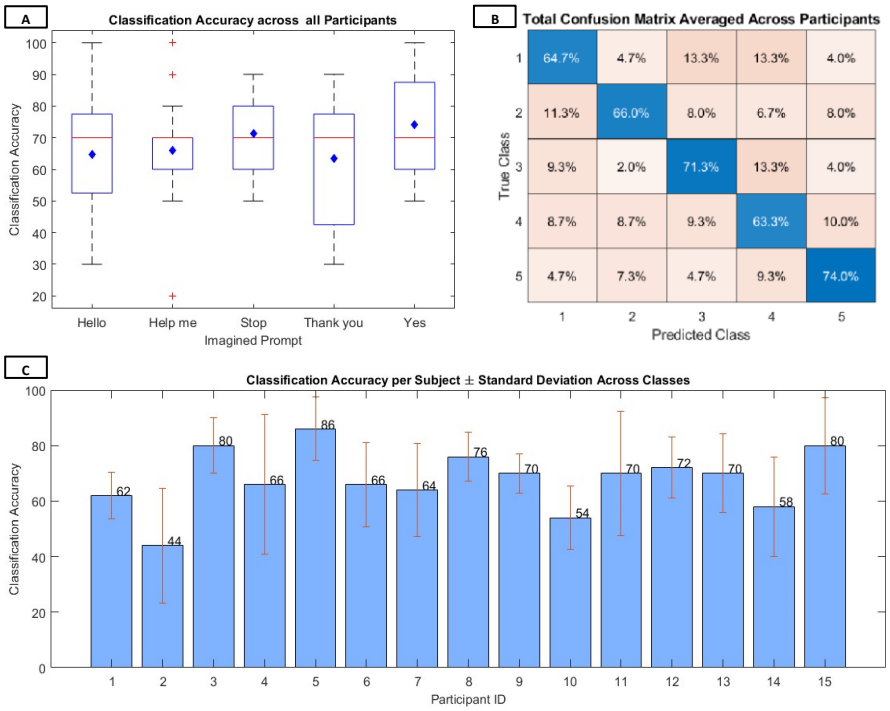

Fig.3.2. The global and subject-wise performance of the proposed decoder. (A) The overall classification accuracy compartmentalized for each imagined prompt, (B) The total confusion matrix, and (C) The average classification accuracy per subject.

It is evident that despite the high subject variability, the majority of the subjects perform well when the imagined prompt is monosyllabic (see Fig.3.2A). This trend may imply that a different approach focusing on syllables rather than words may be required to better decipher the phenomenon of imagined speech. In the same direction, disentangling the two prompts starting with the same syllable "He" (i.e. "Hello" and "Help me") seems highly challenging, given the high false positive values. Returning to the subject variability issue, while the accuracy for the majority of the subjects revolve around 70%, there are subjects with accuracy lower or barely exceeding 50% (i.e. S2, S10, S14), while there are also cases characterized by near-optimal performance (e.g. S3, S5). Considering the nature of the task (i.e. mental task) that in some cases may not be completely straightforward, it is not unlikely that some participants may require a familiarization period prior to the engagement with such tasks, as in the case of the motor imagery paradigm [Georgiadis, 2019].

Despite the aforementioned, the proposed decoding scheme provides classification scores that significantly exceed the random level for this five-class problem that comes at 20%. Finally, it must be noted that the achieved performance surpasses all but one the competitive approaches regarding the

selected dataset [Jeong, 2022]. Additionally, the employed AJD-based covariance estimator surpasses the classical covariance estimator, under the same classification setting, by 3.1% while exhibiting the same trends in class-specific classification results.

3. EEGNET COMBINED WITH RIEMANNIAN GEOMETRY

3.1 MOTIVATION

Preliminary attempts of decoding imagined speech from EEG signals comprised extraction of statistical features (such as mean, variance, skewness and kurtosis) [Min, 2016; Zhao, 2015] or wavelet transform coefficients and classification [Jahangiri, 2018; Jahangiri, 2019a; Jahangiri, 2019b; Panachakel, 2019; Sereshkeh, 2019; Torres-García, 2012] with conventional machine learning algorithms such as Support Vector Machines (SVM), Deep Belief Networks (DBN) and Extreme Learning Machines (ELM). In another approach, instead of working with raw EEG data, researchers used Cross Covariance matrices (CCV) encoding statistical correlation between EEG channels [Saha, 2019; Saha, 2013; Kalaganis, 2023]. A basic characteristic of CCV matrices, namely being symmetric positive definite, allows for alternative processing directions utilizing basic manifold properties that originate from Riemannian geometry. Recent approaches following this pathway have achieved remarkable performance on similar BCI applications [Kalaganis, 2019; Kalaganis, 2022; Georgiadis, 2022; Georgiadis, 2023]. These results motivated the integration of Riemannian geometry with deep learning techniques, one of the most prominent example of which is the SPDNet [Huang, 2017]. More recent studies focused on implementing different deep learning methods, mostly Convolutional Neural Network (CNN) architectures, that demonstrated very promising results, at least when examined on other BCI paradigms. Some of the widely used architectures are the shallow ConvNet, the deep ConvNet [S.R., 2017] and the EEGNet [Lawhern, 2018].

Our approach utilizes some basic concepts, stemming from both CNNs and CCVs, in an effort to combine the best of the two worlds towards building a novel, end-to-end, deep learning architecture. Specifically, the first part of the introduced architecture consists of a convolutional layer of temporal filters as implemented in EEGNet. The output feature maps (i.e., EEG signals filtered in the temporal domain) correspond to selected frequency bands where the most significant brain activation occured. The second part involves converting the resulting maps into CCV matrices. Each matrix is then subjected to multiple linear transformations such that the output matrices also lie in Riemannian manifolds (potentially of varying dimensions) in accordance with SPDNet architecture. Ultimately, the LogEuclidean metric is computed for each matrix and the information is transferred to a fully connected layer for the purpose of classification. As stated earlier, the combination of the above mentioned parts constitute an end-to-end trainable network. In essence, the proposed architecture calculates CCV matrices that can capture the brain connectivity structure that underpins the imagined speech paradigm at various frequency bands. The motivation for employing this particular architecture for the task at hand is related to the “dual stream model” according to which several brain regions are involved and interconnected during speech formulation and understanding [Cooney, 2018]. We note that the methodology presented in this has been presented in EUSIPCO 2024 as part of the BINGO’s outcomes. The code for this methodology will be uploaded as part of project’s toolbox.

3.2 METHODOLOGY

PREPROCESSING

In the Kara One dataset, the raw EEG signals are filtered using a 3rd order high-pass Butterworth filter with cut-off frequency 1Hz. A wavelet ICA algorithm is then employed for the purpose of artifact removal and denoising. The EEG segments corresponding to the imagined speech stage are ultimately fed to the classification algorithms.

No preprocessing steps are followed in the second dataset since the 2020 BCI competition provided the signals in the form of labelled EEG trial segments.

NETWORK ARCHITECTURE

The proposed architecture is an end-to-end trainable network that combines the convolutional temporal filters as implemented in EEGNet with the linear Symmetric Positive Definite (SPD) matrix transformations utilized in SPDNet architecture as shown in Fig.2.

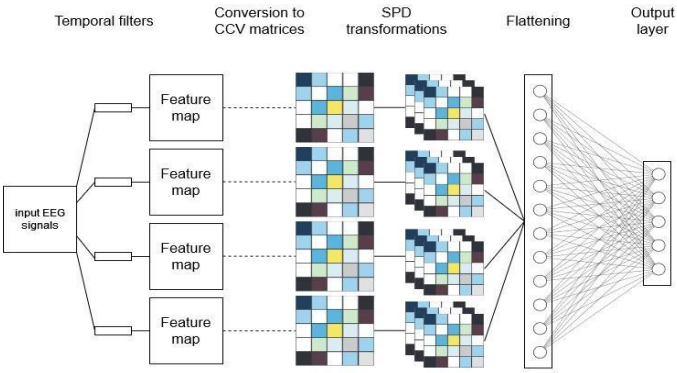

Fig. 4.1. EEGNet-SPDNet architecture.

Below we provide the key components of the network, tailored to imagined speech decoding:

- The temporal filters are convolutional 2D filters with shape $(1, \frac{f_s}{2})$, where f_s is the sample frequency. The convolutional layer is followed by a batch normalization layer, a non-linearity ReLU, a mean pooling layer of size (1,4) and a dropout layer with probability 0.5. The filters output feature maps containing the EEG signal in different band-pass frequencies leaving the channel dimension unaffected.
- The different versions (feature maps) of the EEG signal are then converted to the corresponding covariance matrices.
- Each matrix is then subjected to multiple bilinear transformations that map SPD matrices to other SPD matrices of different dimension. If the input matrix is denoted as: $\mathbf{X} \in R^{N \times N}$, the output matrix as: $\mathbf{Y} \in R^{M \times M}$ and the transformation matrix as: $\mathbf{W} \in R^{M \times N}$ then the mapping is as follows:

$$\mathbf{Y} = \mathbf{W}\mathbf{X}\mathbf{W}^{T}. \qquad (1)$$

The trainable parameters in this part are the elements of the transformation matrix. The output matrix *Y* is symmetric positive definite if the transformation matrix *W* is full rank on the rows. Thus, an exclusive optimizing procedure take place here such that this essential matrix property is preserved. The transformation is followed by a non-linearity layer the function of which is the

rectification of the eigenvalues. Given the diagonalization of **Y** as **X** = **UΣU**^{T} and the output matrix denoted as:
Z, the rectification is:

$$\mathbf{Z} = \mathbf{U}max(\varepsilon\mathbf{I},\mathbf{\Sigma})\mathbf{U}^T \qquad (2)$$

where $\varepsilon > 0$ is the rectification threshold. The $max(\varepsilon\mathbf{I},\mathbf{\Sigma})$ is a diagonal matrix where each diagonal value (eigenvalue) of **Σ** is replaced by: $max(\varepsilon,e_i)$. Essentially, this process prevents eigenvalues from approaching zero. Finally, the matrices are mapped to Euclidean space via the implementation of Log-Euclidean metric and consequently to a linear output layer through flattening.

3.3 RESULTS

A. KARA ONE

Fig.4.2 (top) presents the overall accuracy of the proposed architecture when tested with Kara One dataset as well as the accuracy scores obtained for each subject. Both training and validation are conducted on a personalized level, fitting a different network for each subject. It is apparent from the confusion matrix (Fig.4.2 - Top Right) that the network was able to adequately distinguish phoneme from word prompts but hardly disentangled phonetically similar words (pat/pot and knew/gnaw). The best accuracy is measured in one letter phonemes (/m/, /n/). The novel method implemented in this work exhibits superior performance compared to the EEGNet that barely exceeded random level (9.1%). When it comes to intra-subject multi-classification of words and phonemes, our approach outperforms handcrafted features (i.e. MFCC) combined with SVMs [Cooney, 2018] by 4%. It achieved similar overall accuracy with inter-subject training and validation attempts that employed CNN architectures as well [Rusnac, 2020]. The overall performance of our approach as shown in Table I surpassed all but one competitive approach found in the literature regarding this particular classification task [Panachakel, 2019].

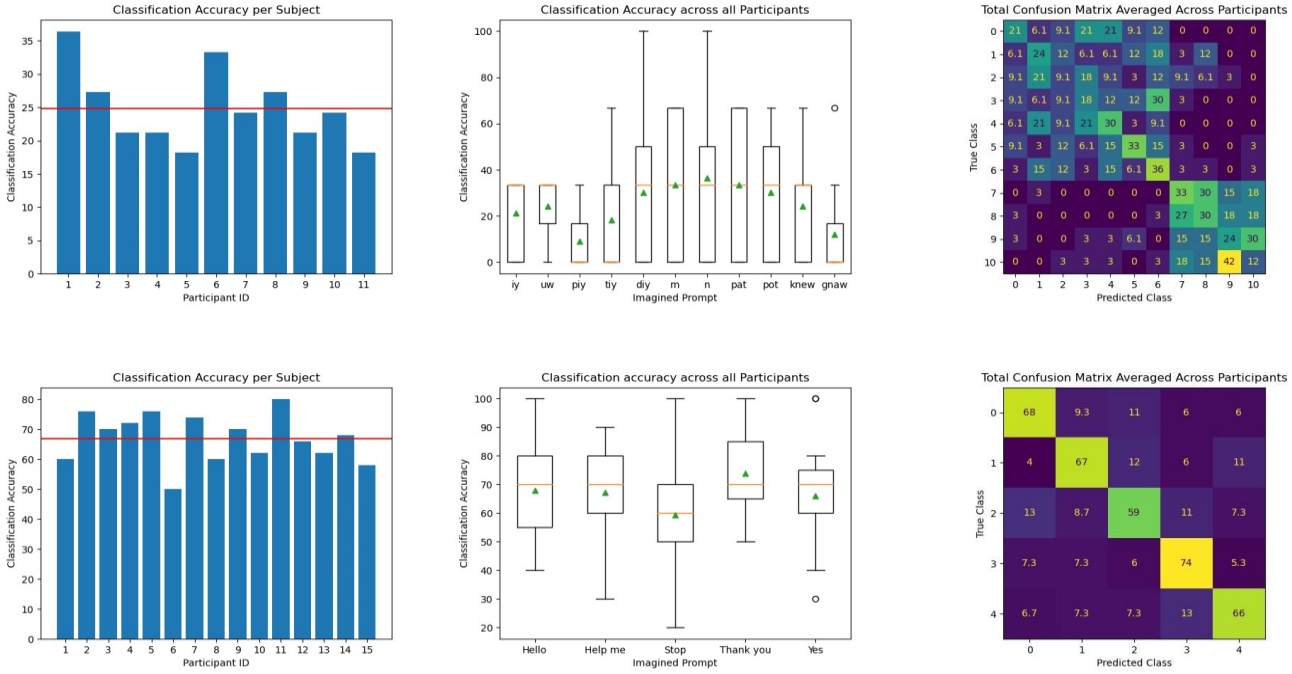

Fig. 4.2. The overall and subject-wise performance on Kara One dataset (top), 2020 BCI Competition dataset (bottom). Left: Classification accuracy per subject and overall (red line). Middle: Classification accuracy across all participants for each imagined prompt. Right: Total confusion matrix

Table 4.1. Mean accuracy for each classification method test on two distinct datasets

	Mean Accuracy (%)	
	Kara One	**2020 BCI Comp.**
MFCC - SVM	17.84	37.86
EEGNet	14.05	50.26
EEGNet + SPDNet	24.79	66.93

B. 2020 BCI COMPETITION

Training and validation took place separately for each subject in this case as well. A major observation from the overall accuracies for each subject (Fig.4.2 - Bottom Left) is the subject variability: there are subjects with accuracy as high as 80% (S11) while others that do not exceed 60% (S6, S15). The complexity of the specific task (imagined speech) could be one factor contributing to this. The novel architecture employed significantly outperformed EEGNet in this case as well (Table 4.1). As mentioned earlier, the attained performance here is on par with the top results of the competition, namely similar to the competitor approach with the second highest average accuracy [J.J., 2022].

4. EEG CONFORMER

4.1 MOTIVATION

The EEG Conformer architecture has recently gained attention as an advanced deep learning model for decoding inner speech from EEG signals. The Conformer model, originally designed for speech processing, combines convolutional neural networks (CNNs) and transformers, enabling both local feature extraction and long-range temporal dependencies.

Motivating its use in inner speech decoding, the Conformer architecture is particularly suited for EEG data due to:

- **Temporal Context Modeling:** Inner speech exhibits complex temporal dependencies, and transformers within the Conformer architecture excel at capturing these long-range patterns.
- **Local Feature Extraction:** Convolutional layers effectively extract local EEG features, enhancing the model's ability to recognize fine-grained neural activity patterns.
- **Self-Attention Mechanism:** This allows the model to focus on the most relevant neural representations while filtering out noise, a crucial factor in EEG-based decoding.
- **Robust Generalization:** The combination of CNNs and transformers helps mitigate inter-subject variability, improving model adaptability across different individuals.

By leveraging these capabilities, EEG Conformer models may provide a state-of-the-art solution for inner speech decoding, potentially improving the accuracy and reliability of BCIs for communication and assistive technologies. It is noted that the code for this methodology will be uploaded as part of project's toolbox

4.2 METHODOLOGY

The overall framework is depicted in Fig. 4.1. 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.

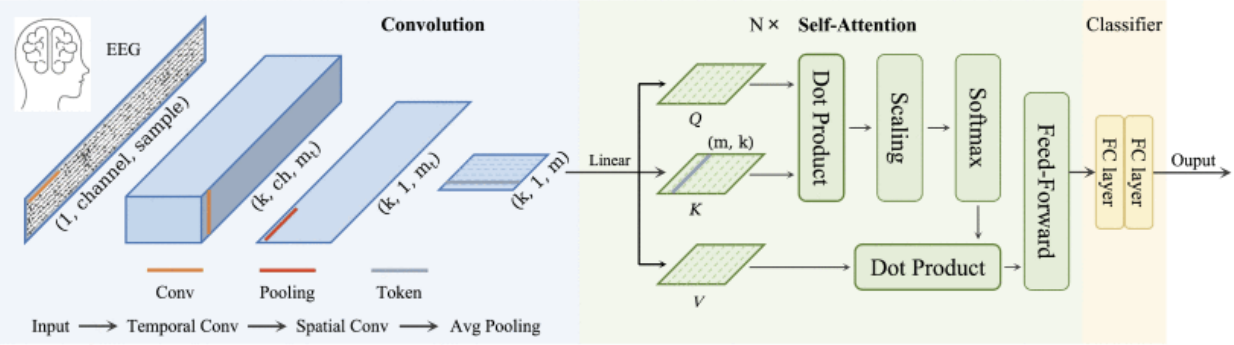

Fig. 4.1. The framework of Convolutional Transformer (Conformer), including a convolution module, a self-attention module, and a classifier module. Image source: https://ieeexplore.ieee.org/document/9991178

PREPROCESSING

THE RAW EEG TRIALS ARE OF SIZE *CH* × *SP* , WHERE *CH* REPRESENTS ELECTRODE CHANNELS AND *SP* DENOTES TIME SAMPLES. WE ONLY USE A FEW STEPS TO PRE-PROCESS THE RAW EEG DATA. FIRST, BAND-PASS FILTERING IS EMPLOYED TO FILTER OUT EXTRANEOUS HIGH AND LOW-FREQUENCY NOISE. HERE, WE USE A 6-ORDER CHEBYSHEV FILTER TO PRESERVE TASK-RELEVANT RHYTHMS. THEN, A Z-SCORE STANDARDIZATION IS PERFORMED TO REDUCE THE FLUCTUATION AND NONSTATIONARITY AS

$$x_o = \frac{x_i - \mu}{\sqrt{\sigma^2}}$$

where x_i and x_o denote band-pass filtered data and the output of standardization, respectively. μ and σ^2 represent the mean and variance, calculated with the training data and used directly for the test data.

NETWORK ARCHTECTURE

Convolution Module: Inspired by [Schirrmeister, 2017] and [Lawhern, 2018], the convolution module is designed by decomposing the two-dimensional convolution operation into two sequential one-dimensional layers: a temporal convolution and a spatial convolution. The first layer applies k kernels of size (1,25) with a stride of (1,1), focusing on capturing temporal patterns. The second layer maintains k kernels of size (ch,1) with a stride of (1,1), where ch represents the number of EEG electrode channels. This layer functions as a spatial filter, learning inter-channel interactions in EEG signals. To enhance training efficiency and mitigate overfitting, batch normalization is employed. Exponential Linear Units (ELUs) are used as the activation function to introduce nonlinearity, following [Lawhern, 2018]. The third layer consists of an average pooling operation along the time axis, utilizing a kernel size of (1,75) and a stride of (1,15), which smooths temporal features, reduces overfitting, and decreases computational demands. Finally, the feature maps from the convolution module are rearranged by compressing the electrode channel dimension and swapping the convolution channel dimension with the time dimension. This transformation ensures that all feature channels corresponding to each time step are treated as tokens and fed into the next module.

Self-Attention Module: We assume that incorporating context-dependent representations into low-level temporal-spatial features can enhance EEG decoding, given the coherence of neural activity. To achieve this, self-attention is utilized in this module to capture global temporal dependencies in EEG features, addressing the constrained receptive field of the convolution module. The tokens processed in the previous stage are linearly transformed into three identical components: query (Q), key (K), and value (V). The correlation between tokens is computed using a dot product between Q and K. A scaling factor is

introduced to prevent gradient vanishing, ensuring stable training. The resulting values are then processed through a Softmax function to generate an attention score matrix. Finally, this attention score is applied to V using a dot product to weight the features accordingly [Vaswani, 2017]. This process can be formulated as

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{k}}\right)V$$

where *k* denotes the length of a token. Besides, two fully-connected feed-forward layers are connected behind to enhance the fitting ability. The input and output sizes of this process remain the same. The entire attention computation is repeated *N* times in the self-attention module.

We also employ the multi-head strategy to further improve representation diversity. The tokens are equally divided into *h* segments and fed into the self-attention module separately, and the results are concatenated as the module output [Vaswani, 2017]. The process can be expressed as

$$MHA(Q, K, V) = [head_0; \cdots; head_{h-1}],$$
$$head_l = Attention(Q_l, K_l, V_l)$$

where MHA stands for multi-head attention, $Q_l, K_l, V_l \in R^{m \times k/h}$ denote the query, key, and value obtained by linear transformation of divided token in the *l* -th head, respectively.

Classification Module: Finally, we adopt two fully-connected layers as the classifier module, which outputs an *M* -dimensional vector after *Softmax* function. Cross-entropy is used as the loss function of the whole framework as

$$L = -\frac{1}{N_b}\sum_{i=1}^{N_b}\sum_{c=1}^{M} y\ log\ log\ (\hat{y})$$

where *M* represents the number of EEG categories, y and $\hat{y}$ are the ground truth and predicted label, respectively. N_b denotes the number of trials in a batch.

4.3 RESULTS (PRELIMINARY)

In this subsection we provide preliminary results of employing the EEG Conformer architecture on the 2020 BCI competition dataset in a subject agnostic manner. The provided image below, Fig.4.2, illustrates the training and validation loss curves (under two different initialization settings) of an EEG Conformer model for decoding imagined speech over 20 epochs. The training loss (solid line) decreases consistently, demonstrating that the model is effectively learning from the training data. However, the validation loss (dotted line) does not exhibit a similar phenomenon and remains significantly higher than the training loss. This indicates potential overfitting, where the model is learning patterns specific to the training set but failing to generalize to unseen data. The divergence between the training and validation loss is a common issue in deep learning models, particularly when working with complex EEG signals, which inherently have high variability and noise. Possible solutions to improve generalization include regularization techniques such as dropout, weight decay, or data augmentation, as well as providing more diverse training data. Additionally, early stopping based on validation loss could help prevent overfitting. The high validation loss suggests that the model might not yet be reliable for real-world applications without further tuning and improvements.

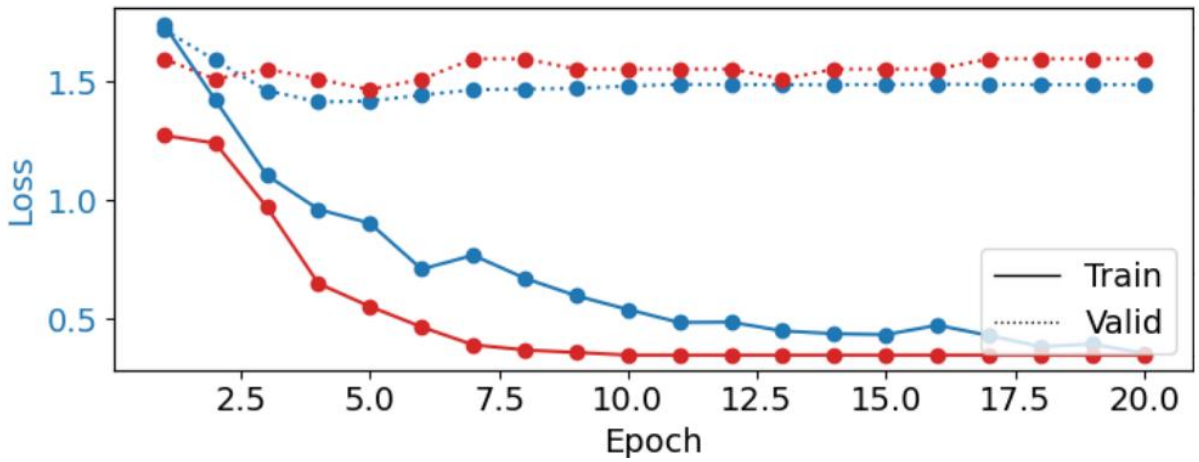

Fig. 4.2. The training process of an EEG conformer, under two different initialization settings, for the 2020 BCI competition dataset.

5. CONCLUSIONS AND FUTURE WORK

This report presents a comprehensive evaluation of imagined speech decoding approaches, leveraging advanced machine learning techniques and neuroscientific models. The study explored three distinct methodologies: (i) Riemannian Geometry-based decoding, (ii) EEGNet combined with Riemannian Geometry, and (iii) EEG Conformer. Each approach incorporates different perspectives on neural signal processing and classification, addressing the challenges posed by the inherent variability and low signal-to-noise ratio of EEG data.

The findings suggest that Riemannian Geometry-based methods provide robust feature extraction and denoising capabilities, enhancing classification performance in EEG-based imagined speech decoding. The integration of EEGNet with Riemannian Geometry demonstrated significant improvements over conventional neural networks, capturing spatiotemporal relationships while maintaining computational efficiency. Lastly, the preliminary results of the EEG Conformer model indicate the potential of attention-based architectures in capturing long-range dependencies, though challenges related to overfitting and generalization remain.

Despite promising results, inter-subject variability continues to pose a major challenge. The observed differences in classification accuracy among participants highlight the necessity of adaptive learning models that can accommodate individual neural patterns. Furthermore, the ability to differentiate phonetically similar words remains an open problem, suggesting the need for more sophisticated feature extraction techniques and alternative classification strategies.

Building upon the current findings, future research will focus on the following areas:

1. **Enhancing Generalization and Adaptability**: Developing subject-independent models through transfer learning, domain adaptation, and meta-learning approaches to improve generalization across individuals.
2. **Advanced Neural Architectures**: Refining deep learning models by incorporating transformer-based architectures with improved attention mechanisms, regularization techniques, and self-supervised learning paradigms.
3. **Improved Data Collection Strategies**: Conducting large-scale experiments with more diverse participants and incorporating real-world settings to ensure the validity of the proposed methods.
4. **Expanding the Lexicon**: Extending the vocabulary of decoded words and syllables to enable more complex and naturalistic communication through BCIs.

By addressing these challenges and research directions, our future studies will be aimed towards enhancing the feasibility and accuracy of imagined speech decoding, paving the way for practical applications in neurorehabilitation, assistive communication, and human-computer interaction.

REFERENCES

Cardoso, J. F., & Souloumiac, A. (1996). Jacobi angles for simultaneous diagonalization. *SIAM Journal on Matrix Analysis and Applications, 17*(1), 161–164.

Congedo, M., Barachant, A., & Bhatia, R. (2017). Riemannian geometry for EEG-based brain-computer interfaces: A primer and a review. *Brain-Computer Interfaces, 4*(3), 155–174.

Cooney, C., Folli, R., & Coyle, D. (2018). Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG. *2018 29th Irish Signals and Systems Conference (ISSC)*, 1–7.

Cooney, C., Korik, A., Folli, R., & Coyle, D. (2020). Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech EEG. *Sensors, 20*(16), 4629.

Georgiadis, K., Kalaganis, F. P., Oikonomou, V. P., Nikolopoulos, S., Laskaris, N. A., & Kompatsiaris, I. (2022). RNeumark: A Riemannian EEG analysis framework for neuromarketing. *Brain Informatics, 9*(1), 22.

Georgiadis, K., Kalaganis, F. P., Oikonomou, V. P., Nikolopoulos, S., Laskaris, N. A., & Kompatsiaris, I. (2023). Harnessing the potential of EEG in neuromarketing with deep learning and Riemannian geometry. *International Conference on Brain Informatics*, 21–32. Springer.

Georgiadis, K., Laskaris, N., Nikolopoulos, S., & Kompatsiaris, I. (2019). Connectivity steered graph Fourier transform for motor imagery BCI decoding. *Journal of Neural Engineering, 16*(5), 056021.

Hickok, G. (2022). The dual stream model of speech and language processing. *Handbook of clinical neurology*, *185*, 57-69.

Huang, Z., & Van Gool, L. (2017). A Riemannian network for SPD matrix learning. *Proceedings of the AAAI Conference on Artificial Intelligence, 31*(1).

Hyvärinen, A. (1999). Survey on independent component analysis.

J. J. et al. (2022). 2020 International brain-computer interface competition: A review. *Frontiers in Human Neuroscience, 16*.

Jahangiri, A., Chau, J. M., Achanccaray, D. R., & Sepulveda, F. (2018). Covert speech vs. motor imagery: A comparative study of class separability in identical environments. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE.

Jahangiri, S. F. (2019a). The relative contribution of high gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in EEG data. *Journal of Medical Systems, 43*(2), 20.

Jahangiri, S. F., & Achanccaray, D. (2019b). A novel EEG-based four-class linguistic BCI. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 3050–3053. IEEE.

Jeong, J. H., Cho, J. H., Lee, Y. E., Lee, S. H., Shin, G. H., Kweon, Y. S., Millán, J. d. R., Müller, K. R., & Lee, S. W. (2022). 2020 International brain–computer interface competition: A review. *Frontiers in Human Neuroscience, 16*, 898300.

Kalaganis, F. P., Georgiadis, K., Oikonomou, V. P., Nikolopoulos, S., Laskaris, N. A., & Kompatsiaris, I. (2023). Exploiting approximate joint diagonalization for covariance estimation in imagined speech decoding. *International Conference on Brain Informatics*, 409–419. Springer.

Kalaganis, F. P., Laskaris, N. A., Chatzilari, E., Nikolopoulos, S., & Kompatsiaris, I. (2019). A Riemannian geometry approach to reduced and discriminative covariance estimation in brain-computer interfaces. *IEEE Transactions on Biomedical Engineering, 67*(1), 245–255.

Kalaganis, F. P., Laskaris, N. A., Oikonomou, V. P., Nikopolopoulos, S., & Kompatsiaris, I. (2022). Revisiting Riemannian geometry-based EEG decoding through approximate joint diagonalization. *Journal of Neural Engineering, 19*(6), 066030.

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).

Lopez-Bernal, D., Balderas, D., Ponce, P., & Molina, A. (2022). A state-of-the-art review of EEG-based imagined speech decoding. *Frontiers in human neuroscience*, *16*, 867281.

Min, B., Kim, J., Park, H.-J., & Lee, B. (2016). Vowel imagery decoding toward silent speech BCI using extreme learning machine with electroencephalogram. *Biomedical Research International*, *2016*, 1–11.

Müller, K. R., Vigario, R., Meinecke, F., & Ziehe, A. (2004). Blind source separation techniques for decomposing event-related brain signals. *International Journal of Bifurcation and Chaos, 14*(2), 773–791.

Panachakel, A. G. R. J. T., & Ananthapadmanabha, T. V. (2019). Decoding imagined speech using wavelet features and deep neural networks. *16th India Council International Conference*, 1–4. IEEE.

Pennec, X., Fillard, P., & Ayache, N. (2006). A Riemannian framework for tensor computing. *International Journal of Computer Vision, 66*(1), 41–66.

Rusnac, A.-L., & Grigore, O. (2020). Generalized brain-computer interface system for EEG imaginary speech recognition. *2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC)*, 184–188.

S. R. et al. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization. *Human Brain Mapping, 38*(11), 5391–5420.

Saha, F. S., & A.-M., M. (2019). Deep learning the EEG manifold for phonological categorization from active thoughts. *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2762–2766.

Saha, P., & Fels, S. (2020). Hierarchical deep feature learning for decoding imagined speech from EEG. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(1). IEEE.

Schirrmeister, R. T., Springenberg, J. T., Fiederer, L. D. J., Glasstetter, M., Eggensperger, K., Tangermann, M., ... & Ball, T. (2017). Deep learning with convolutional neural networks for EEG decoding and visualization: Convolutional neural networks in EEG analysis. *Human Brain Mapping, 38*(11), 5391–5420.

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.

Si, X., Li, S., Xiang, S., Yu, J., & Ming, D. (2021). Imagined speech increases the hemodynamic response and functional connectivity of the dorsal motor cortex. *Journal of Neural Engineering, 18*(5), 056048.

Torres-García, A. A., Reyes-Garcia, C. A., & Villaseñor-Pineda, L. (2012). Toward a silent speech interface based on unspoken speech. *BIOSIGNALS 2012 - Proceedings of the International Conference on Bio-Inspired Systems and Signal Processing*, 370–373.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 1–11.

Zhao, S., & Rudzicz, F. (2015). Classifying phonological categories in imagined and articulated speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 992–996.

Ziehe, A. (2005). Blind source separation based on joint diagonalization of matrices with applications in biomedical signal processing (Doctoral dissertation, Universität Potsdam).