POLITECNICO MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE E DELL'INFORMAZIONE

EXECUTIVE SUMMARY OF THE THESIS

A portable EEG-based Brain-Computer Interface for Imagined Speech Detection: towards an Assistive System for Restoring Communication

MASTER THESIS IN BIOMEDICAL ENGINEERING
Academic year: 2022-23

Author: Federico Cavallini (10619764)

Advisor: Professor Emilia Ambrosini

Co-advisors: Professor Marc Van Hulle (KU Leuven)
Aurélie De Borman (KU Leuven)
Bob Van Dyck (KU Leuven)

1. Introduction

Brain-Computer Interfaces (BCIs) provide an alternative method for users to interact with the world by translating brain activity into commands for external devices. BCIs can significantly benefit individuals with severe motor disabilities, enhancing their communication and independence.

Verbal communication loss due to conditions like stroke or amyotrophic lateral sclerosis (ALS) has motivated research in speech BCIs, aiming to restore communication for individuals with preserved cognitive abilities. Different speech-related modalities have been explored, including attempted speech, silent speech, inner speech and imagined speech. The latter – imagined speech (IS) – is a cognitive task where individuals mentally simulate speaking without any actual articulatory movement, akin to first-person motor imagery. IS is particularly interesting as it can be executed similarly by both healthy subjects and paralyzed patients, facilitating the transfer of information for restorative applications. Due to the lack of voluntary movement, in IS the main challenges arise from the absence of a ground truth signal. While invasive methods like ECoG [1] and microelectrode array implants [2] offer better performances for decoding IS, EEG is commonly employed due to its cost-effectiveness, safety and usability for investigation on healthy subjects. At the state-of-the-art [3]–[6], most of the studies use clinical-grade EEG systems featuring 64 electrodes montages.

In this thesis, we introduce the use of a research-grade portable EEG system only featuring 8 channels to implement a BCI for detecting IS. The primary objective is to assess the feasibility of this novel approach.

The study includes an initial comprehensive analysis aimed at disentangling IS mechanisms employing two different paradigms. Deepening into the BCI implementation details, the aim is twofold: firstly, we want to perform a fair comparison with the other studies in the state-of-the-art, enabling the evaluation of our model's offline detectability performance on six healthy subjects. Secondly, we aim at transferring the system to online settings providing neurofeedback to improve BCI performance. The model is trained to detect IS in real-time, providing visual feedback to users. The goal is to determine if the system can provide trustworthy feedback and how it helps users to adapt their imagination strategy in response to the predictive capabilities of the model.

2. Materials

2.1. Data acquisition

Brain signals are recorded with Mentalab Explore+, an 8-channel research-grade EEG recording device. It is composed of 9 electrodes: the ground electrode is affixed to hair-free forehead with a wet sticker electrode. The remaining 8 channels employ conductive polymeric electrodes with a minimal amount of conductive gel to enhance signal quality. Overall, the setup time is remarkably brief, taking just 5-10 minutes, and the device lightweight and portable design significantly improves the comfort for users.

After a comprehensive analysis of the state-of-the-art, optimal locations for the 8 channels are chosen, covering both auditory cortices (3 electrodes per hemisphere) and addressing speech areas asymmetry with two additional channels on the left hemisphere. The final montage comprises the ground at Fpz and the other 8 channels at locations TP8, C6, FT10, TP7, C6, FT9, FC5 and F5. Sample frequency was set to 500Hz.

2.2. Paradigms

In this study, the focus is to detect IS, *i.e.* whether the subjects are imaging to say a word or not. For these experiments, the vocabulary of words to be imagined is limited to the English words “LEFT” and “RIGHT”. In addition, a control resting class denoted as “NONE” is included to maintain consistent sensorial stimulation: subjects receive the same go-cue as in the imagination tasks, but no specific mental task is required. This prevents the detection of IS from being influenced by evoked reactions to visual stimuli rather than the actual processes of speech imagination. The two classes comprising the task to imagine a word will be addressed together as “IMAGINE” class, to be distinguished from the “NONE” class.

Two paradigms were implemented. In both of them, the “IMAGINE” classes are instructed with oriented triangles, while the “NONE” class is associated with the absence of any instruction. The core difference between the two lies in how the go-cue is presented. In the *sliding cues* paradigm (Figure 1), the go-cues are represented by a stream of crosses – surrounded by triangles or not – sliding horizontally at a constant speed. The “IMAGINE” or “NONE” task should commence when the corresponding sliding cross aligns with the central fixation cross. In the *color-changing cues* paradigm (Figure 2), subjects are required to start imagining saying the word (or not imagining anything specific) when the fixation cross turns from white to green. To optimize the duration of the acquisition, each instruction is followed by three consecutive repetitions of the go-cue. Hence, each trial is composed of three repetitions.

All the participants were recruited on voluntary basis and provided written consent. Experiments were pre-approved by the Ethical Committee of UZLeuven, Belgium.

2.3. Pilot study experiment

In the pilot study, a single subject (male; age: 23; right-handed; without neurological complaints) underwent two experimental sessions. The first

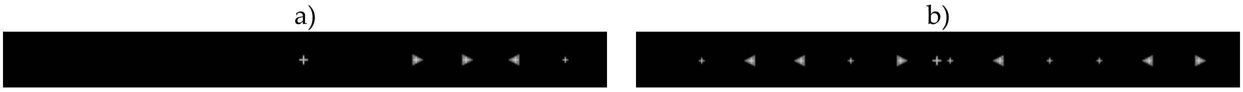

Figure 1: Sliding cues paradigm. In a) the experiment has just started and the subjects is waiting for the first cue (“RIGHT”) to slide over the fixation white central cross. In b) a “LEFT” task has just passed and the next one will be a “NONE” cue. The time between two consecutive cues is 4s.

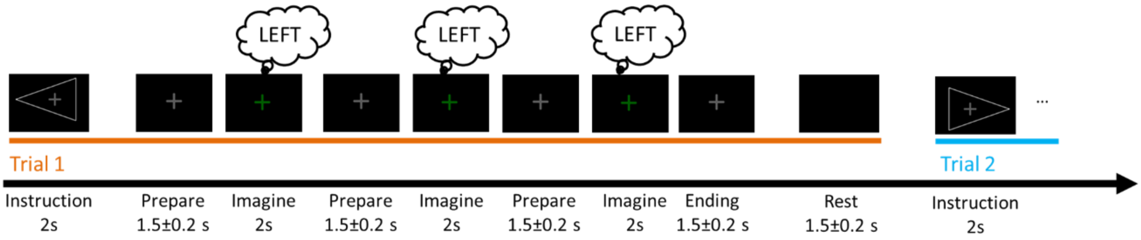

Figure 2: Color-changing cues paradigm. One trial is composed of 3 consequent repetitions of the same word. The subject should imagine saying the word when the fixation cross turns green (as depicted by the thought bubbles). All the phases have a specific duration.

employed color-changing cues paradigm. 210 repetitions for each of the three classes were recorded in 55 minutes. However, only 195 repetitions were deemed usable due to problems during the acquisition. The second session used the sliding cues paradigm. In 55 minutes, 210 repetitions per class were acquired. Some of them were excluded due to eye blink artifacts and, after a balancing process, 189 trials per class were available.

To create a balanced dataset for the detection problem, the "IMAGINE" class dataset was formed by merging half of the trials for "RIGHT" and "LEFT" through random down-sampling in a stratified manner.

2.4. Offline experiments

For the whole *corpus* offline experiment, six subjects (2 females; aged between 23 and 30; without neurological complaints) were recruited for two sessions of 50 minutes at least one week apart. Both sessions employed the color-changing cues paradigm and were divided into 12 blocks interleaved by a 30s pause. In each block six trials belonged to "NONE" class and six trials to "IMAGINE" class (3 "RIGHT" and 3 "LEFT"). The trials were randomly sorted. Each session resulted into 216 repetitions for "NONE" and "IMAGINE" classes.

2.5. Online experiments

Three out of the six subjects also participated in an online session. It was based on the color-changing cues paradigm and was composed of 12 blocks. The six initial blocks, forming the offline training set, mirrored the structure of offline sessions. Then, a predictive model was trained and used to detect IS in real-time. After each trial visual feedback about each of the three consecutive repetitions was provided as "OK" if the model correctly classified the corresponding single repetition or "X" if it did not. The model underwent initial training after the first six blocks without feedback. Subsequently, it was retrained at the conclusion of each block, cumulatively incorporating all previous trials of the session into the training set.

3. Methods

3.1. TF representations

To address high non-stationarity of the EEG signals, time-frequency (TF) analysis is essential for capturing and comprehending dynamic changes in brain signals. At first, Morlet Wavelet transform was applied to "IMAGINE" epochs to visualize the neural activation induced by IS. The Gaussian window of the wavelet is adapted to the analysed frequency: its variance σ is the time needed for 10 complete periods of the analysed frequency. Then, to cope with the absence of a ground truth signal, the most discriminant time instants and spectral components usable for classification were identified. Classification of epochs into the two classes "IMAGINE" and "NONE" is performed by a simple model on consecutive narrow time windows, using only specific frequency bands to identify the most relevant TF features. For each couple time instant-spectral component, a 5-fold cross validated accuracy score is computed. Finally, each score is used to plot an accuracy map in the TF bidimensional domain.

3.2. Classifier and metrics

The binary classification model employed for IS detection is depicted in Figure 3. Epochs featuring 8-channel EEG signal, of 2s length, are extracted around the go-cue time. They are processed in two separated branches: the first extracting 660 features through the use of a Filter-Bank Common Spatial Pattern (FB-CSP) algorithm, the second computing 16 ratios between functional frequency-band powers. Then, a supervised feature selection

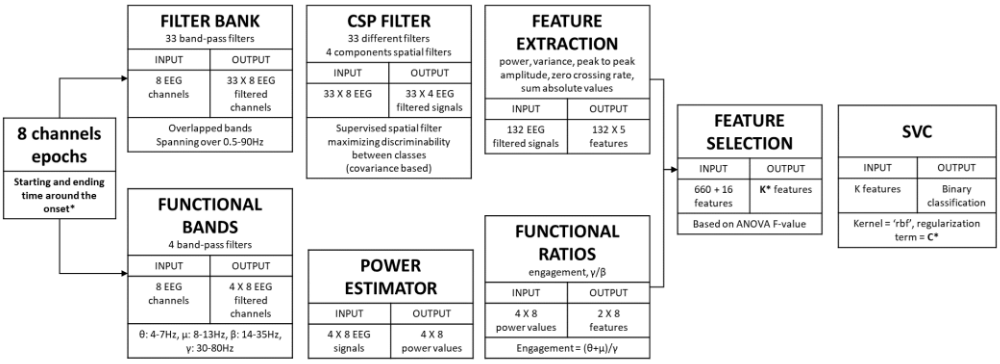

Figure 3: IS detection model. The hyperparameters to tune are shown with an asterisk.

method – based on an ANOVA F-value ranking – reduces the size of the input to a Support Vector Classifier (SVC). The hyperparameters to be tuned in this model are the starting point of the 2s epochs (T_{start}), the number of features to be selected (K) and the regularization term of the SVC (C).

To prevent data leakage, hyperparameters for the evaluation of offline detection accuracy in all 6 subjects are tuned using a grid-search procedure on the pilot study data (resulting in $T_{start} = 0.25$; $K = 200$; $C = 1$). Then, for the online model, the same grid-search procedure is applied on the best offline session of the tested subject to obtain subject-specific optimal hyperparameters.

3.3. Evaluation metrics

As often employed in BCI studies, to ensure the robustness of the performance estimates used to evaluate the system, a 10-fold cross-validation (CV) procedure is employed. The metric used to evaluate the prediction performance on each split is the classification accuracy. The final accuracy is obtained as an average of the 10 CV scores. It is essential to evaluate the achieved accuracy in comparison with the chance level for the specific problem, *i.e.* the accuracy that a random classifier would achieve. In this case, being a balanced binary classification problem, it is 50%, with 1% confidence upper boundary of 56.1% (given by the number of repetitions per class).

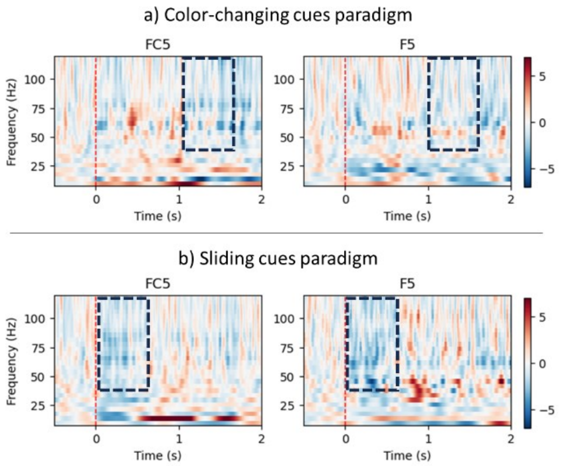

Figure 4: TF wavelet representation of "IMAGINE" class for channels F5 and FC5. In a) trials of the color-changing cues paradigm session are presented; in b) sliding cues paradigm. Time 0s refers to the moment when the go-cue is given. The power values (always positive) obtained by the Wavelet transform, are normalized with respect to a baseline (-0.5s-0s) period using a z-score scaler. Blue intensity encodes for negative values (hence power reduction), red intensity for positive values (hence power increase). Dashed boxes encircle the identified ERD (reduction of the signal power) happening in γ band at different timing in the two paradigms.

For the evaluation of offline sessions, two splitting strategies – *i.e.* the way each epoch is assigned to the train or test set to create CV folds – are implemented and compared:

- *random split*: all the repetitions of each task are divided in a pseudo-random way between the two sets, allowing for example the presence of two consecutive repetitions in the train split and the third in the test.
- *trial-wise split*: the three consecutive repetitions following the same instruction are kept together either in the train or in the test set. It is used to avoid the model to base the prediction about the test set on the shared temporal features which are not related to IS.

The influence of the employed splitting strategy on the obtained accuracy is analysed via the Wilcoxon signed-rank test for paired samples, by comparing the accuracies achieved with the two strategies on each session.

4. Results

4.1. Pilot study

In Figure 4 the TF representation of "IMAGINE" epochs are visualized and the two paradigms are compared. A 0.75s negative inflection of the power levels associated to γ band (30-70Hz) indicates an Event Related Desynchronization (ERD) happening when the subject imagines saying a word. However, this ERD is shifted by 1s when comparing the two paradigms: while it starts about synchronously with the go-cue in the sliding cues paradigm, it starts 1s after the go-cue for color-changing paradigm. Figure 5 depicts the same 1s shift between the two paradigms in terms of discriminability. The most differentiable regions of the TF 2D map lie within the γ range (30-70Hz) and β range (14-35Hz) for both the paradigms. However, they are delayed when color-changing cues are employed. Quantitatively, a 1s mismatch

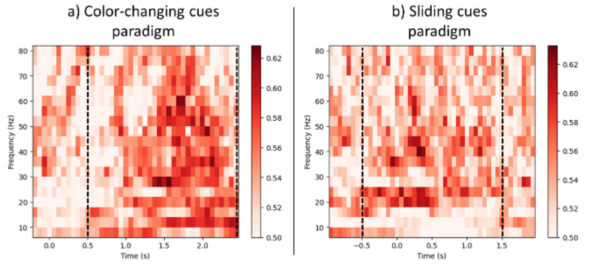

Figure 5: TF classification maps for the two paradigms. Time 0s refers to the moment when the go-cue is given. Red intensity encodes the accuracy level achieved by using epochs cutted around the specific time instant and considering the relative frequency.

is also found in the optimal time windows identified for each paradigm through CV. The optimal time windows are reported with dashed lines in the figure.

4.2. IS offline detection

The Wilcoxon signed-rank test found a statistically significant difference (22 paired samples, p<0.001) in the classification accuracy achieved by the model when it is evaluated with a CV procedure employing the random splitting strategy (69.6±11.6%) or using the trial-wise strategy (63.8±13.2%). So, when the paradigm employs multiple consecutive repetitions related to the same instruction, using a random splitting strategy inflates the reported accuracy.

The trial-wise splitting strategy was employed to evaluate the accuracy in different sessions. In Table 1 for each subject it is reported the accuracy achieved in their best session (it was the second for all except subject 5, but not significative difference was found). All the subjects except subject 3 surpassed the chance level (50%) and the 1% confidence upper boundary (56.1%).

Sub 1	Sub 2	Sub 3	Sub 4	Sub 5	Sub 6	**Avg**
67.2%	74.0%	47.2%	62.1%	60.6%	86.6%	**66.3%**

Table 1: IS detection accuracy. Best session for each subject.

4.3. IS online detection

Subjects 1, 4 and 6 participated in online experiments. Table 2 reports the accuracy of the online predictions which were provided as visual feedback to subject in real-time during the last six blocks of the session. For subject 1, the initial block starts at chance level and exhibits a gradual improvement, eventually reaching peak accuracies of 83.3% and 77.8% in blocks 10 and 12. For subject 6, in the first four online blocks the real-time feedback was not properly working; instead in the last two blocks, the real-time feedback was given correctly 88.9% of the times. Finally, for subject 4, the model never performed much better than chance level, reaching a peak at the last block of 63.9%. For subjects 1 and 6, the model learnt how to provide real-time feedback significantly better than random and with usable and trustable outcomes for the user. They both display a common pattern: the feedback provided in the initial blocks appears to be meaningless, but later, after a different number of blocks, it begins to deliver correct real-time predictions.

Sub	Block 7	Block 8	Block 9	Block 10	Block 11	Block 12
1	50%	55.5%	69.4%	***83.3%**	61.1%	***77.8%**
4	52.8%	47.2%	55.6%	61.1%	55.6%	63.9%
6	50%	61.0%	50%	50%	***88.9%**	***88.9%**

Table 2: Online detection accuracy per block. In each blocks the support is 18 epochs per class. 1% confidence upper boundary for the random classifier is 70%: the blocks with performances significantly better than random are indicated with an asterisk.

5. Discussion

5.1. Time shift in IS with different paradigms

Aligned with other EEG studies, we observed an ERD in the γ band related to IS task. Also from the analysis of TF classification maps, the TF couples related to that γ band ERD were most effective in discriminating "IMAGINE" from "NONE" trials. The consistent 1s time shift, evident in both TF representations and in time window optimization, shows that the chosen paradigm influences the timing of the IS task. This time shift was expected: in the sliding cues paradigm, participants could anticipate the task's initiation by observing the cue gradually approaching the fixation cross, leading to precise timing (optimal window starts before the go-cue time). Conversely, in the color-changing cues paradigm, the subject faced challenges in instantly performing the task due to the reaction time required to perceive the cue changing color, resulting in delayed imagination processes. Despite this delay, subjects exhibited consistency in the color-changing cues paradigm too. Indeed, the similar intensity of observed phenomena suggests that a proper tuning of the time window might lead to equivalent results. For the whole *corpus* experiments, the color-changing paradigm was preferred for its similarity to studies present in literature allowing an unbiased comparison with state-of-the-art benchmarks.

5.2. Relevance of the splitting strategy

In the evaluation of model performances through k-fold CV, the statistical test has shown that the splitting strategy employed to create the splits influences the model accuracy. A wise choice is essential to prevent the model from achieving inflated performances based on the non-stationarity of EEG signals. Specifically, the three repetitions associated with the same instruction share certain features unrelated to the imagination process but arising from their temporal proximity. Hence, for the purpose of IS detection these time-related features are unsuitable for online classification. Therefore, their impact should be minimized in offline analysis to ensure a valid assessment of the model's ability to detect IS. A trial-wise splitting procedure coupled with a random sorting of instruction within blocks – as employed in this study – removes the influence of these proximity features on model’s predictions. Conversely, employing a random splitting procedure has been demonstrated to inflate model accuracy by exploiting this effect. It is reasonable to presume that this inflationary effect may escalate with an increasing number of consecutive repetitions.

Similar conclusions were drawn in [4] where the impact of the EEG non-stationarity was also investigated and demonstrated. They employed a dataset with epochs related to word imagination or silence tasks. Data were acquired from distinct time intervals. Their model achieved a detection accuracy ranging from 97% when inter- and intra-class time distances are different (indicating trials of different classes coming from distinct recording sections) to 58% when they were the same (indicating trials of different classes coming from the same time interval) hence removing the temporal proximity effect on classification.

5.3. Offline IS detection: a fair comparison with literature

These considerations about the exploitation of time proximity more than IS related features force a critical analysis of the state-of-the-art performance report. For example, they challenge the 80% detection accuracy reported in [3], where four repetitions were associated to each instruction and a random split strategy was employed to create the 10 folds used for CV. In our analysis, 69.6% average detection accuracy was achieved with random splitting strategy, but we employed one repetition less and only 8 EEG channels. Hence in comparison with [3] we achieved competitive perfomances, but being based on random splitting strategy we cannot ensure how much these models relate to IS or to temporal proximity features for classification. When considering [4] (where 64 EEG channels are employed too), to get rid of the temporal features effect, it should be taken into consideration the modality where intra- and inter-class time differences are the same which led to a detection accuracy of 58%. This modality can be compared to our trial-wise splitting procedure, where we achieved much higher accuracy, 66.3%.

All participants except one surpassed chance level, and its confidence upper boundary. Considering the achieved performances in a fair comparison with the literature for IS detection, a considerable advancement was proposed in this study by the use of a portable, 8-channel EEG device.

5.4. Online IS detection feasibility

Processing data in real-time and providing user with neurofeedback is crucial for practical BCI applications. Closing the BCI loop through neurofeedback facilitates mutual learning of the user and the system, potentially enhancing BCI performance over time. For two of the three subjects online feedback was finally properly provided surpassing random prediction level with a similar improving trend. In average, we achieved lower performance compared to the two other studies present in literature which implemented an online IS BCI with 64-channel EEG systems [5], [6] attaining respectively 76% and 75% mean online accuracy. However, the observed trends suggest that sustained higher accuracy levels could be achieved with additional blocks. Peak detection accuracies of 83.3% and 89.9% in the final blocks show the potential of Mentalab Explore+ for future implementations of IS detection BCI with online feedback, enabling the two adaptive controllers – the model and the user – to improve together in a synergic learning process.

6. Conclusion

In this thesis we have highlighted the importance of identifying the best time window for the employed experimental protocol to cope with the lack of a ground truth signal in the IS paradigm. Mentalab Explore+, an 8-channel research-grade EEG portable amplifier, was successfully used to implement an offline BCI pipeline for detecting IS in healthy subjects, achieving 66.3% accuracy over six participant (with peaks of 74% and 87% in best sessions). Finally, the system was applied in a real-time framework, revealing its potential by reaching online detection accuracies up to 89% in the final stages of the sessions.

Future developments should focus on enlarging the pool of participants in online sessions and extending the problem to multi-class IS decoding, through the essential employment of neurofeedback to enable mutual learning of both the system and the user.

References

[1] D. A. Moses *et al.*, "Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria," *New England Journal of Medicine*, vol. 385, no. 3, pp. 217–227, Jul. 2021, doi: 10.1056/nejmoa2027540.

[2] F. R. Willett *et al.*, "A high-performance speech neuroprosthesis," *Nature*, vol. 620, no. 7976, pp. 1031–1036, Aug. 2023, doi: 10.1038/s41586-023-06377-x.

[3] S. H. Lee, M. Lee, and S. W. Lee, "Neural Decoding of Imagined Speech and Visual Imagery as Intuitive Paradigms for BCI Communication," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 12, pp. 2647–2659, Dec. 2020, doi: 10.1109/TNSRE.2020.3040289.

[4] M. R. Asghari Bejestani, G. R. Mohammad Khani, V. R. Nafisi, and F. Darakeh, "EEG-Based Multiword Imagined Speech Classification for Persian Words," *Biomed Res Int*, vol. 2022, 2022, doi: 10.1155/2022/8333084.

[5] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG Classification of Covert Speech for Brain-Computer Interfacing," *Int J Neural Syst*, vol. 27, no. 8, Dec. 2017, doi: 10.1142/S0129065717500332.

[6] J. Moon and T. Chau, "Online Ternary Classification of Covert Speech by Leveraging the Passive Perception of Speech," *Int J Neural Syst*, vol. 33, no. 9, Sep. 2023, doi: 10.1142/S012906572350048X.

Acknowledgements

I sincerely thank Professor Van Hulle for giving me the opportunity to develop my thesis project under his supervision in the Laboratory for Neuro- and Psychophysiology of the KU Leuven Medical School. I express my gratitude to Aurélie and Bob for their fundamental daily supervision and guidance throughout the project. A special thanks goes to my advisor, Professor Ambrosini, for her availability in supporting me and providing valuable advice in remote mode during the whole research conducted at the foreign location.