Transfer Learning for Covert Speech Classification Using EEG Hilbert Envelope and Temporal Fine Structure

Saravanakumar Duraisamy*   Mateusz Dubiel*   Maurice Rekrut[†]   Luis A. Leiva*

*University of Luxembourg, Luxembourg

[†]German Research Center for Artificial Intelligence (DFKI), Germany

Abstract

Brain-Computer Interfaces (BCIs) can decode imagined speech from neural activity. However, these systems typically require extensive training sessions where participants imaginedly repeat words, leading to mental fatigue and difficulties identifying the onset of words, especially when imagining sequences of words. This paper addresses these challenges by transferring a classifier trained in overt speech data to covert speech classification. We used electroencephalogram (EEG) features derived from the Hilbert envelope and temporal fine structure, and used them to train a bidirectional long-short-term memory (BiLSTM) model for classification. Our method reduces the burden of extensive training and achieves state-of-the-art classification accuracy: 86.44% for overt speech and 79.82% for covert speech using the overt speech classifier.

Index Terms

Speech BCI; Electroencephalography; Hilbert Envelope; Transfer Learning

I. INTRODUCTION

Covert (or silent) speech refers to speech that is produced without any audible sound [1], [2]. Various approaches have been proposed to measure covert speech, including electrical activity of articulatory muscles or ultrasound-based lipreading [3], [4]. A more natural approach is to measure the neural correlates of speech directly from brain activity, using Brain-Computer Interfacing (BCI) technology in the so-called speech imagery or imaginary speech BCIs [5], [6], [7].

Since speech is generated in the brain before it is articulated [8], decoding it directly from brain signals bypasses the need for muscle movements. This makes the process silent and imperceptible to others. In essence, imagined speech BCIs aim to interpret words or phonemes from brain activity alone, without any audible sound or physical movement involved. Although invasive methods such as electrocorticography (ECoG) have yielded promising results [9], [10], noninvasive approaches such as electroencephalography (EEG) have also shown potential in decoding imagined speech [11], [12]. However, these results are often confined to controlled lab environments and require extensive and exhausting training sessions.

A major challenge in imagined speech BCIs is the tedious and error-prone training process, in which participants silently repeat words, leading to potential data mislabeling due to unverified output and variability in speech onset, offset, and speed. We propose a transfer learning approach to address these challenges by training a classifier on EEG data recorded during overt speech and applying it to EEG data of the same person silently repeating the words. This method not only reduces the mental and physical demands of training, but also enhances productivity by allowing simultaneous interaction and classifier training.

We used the data recorded during a previous study [13] in which participants navigated a virtual robot through a maze using both overt and imagined speech,[1] to apply a set of feature extraction methods (described later) and transferred a classifier from overt to imagined speech data, drawing inspiration from related work, which we discuss in the next section. Our research shows promise for developing BCI systems that can accurately decode covert speech using brain signals. This could enable truly usable applications in speech communication for people with disabilities or new communication tools.

II. RELATED WORK

Garcia-Salinas et al. [14] recorded EEG data from 27 participants while they silently spoke five words (SELECT, RIGHT, LEFT, DOWN, and UP). A Naive Bayes classifier trained on a subset of words and tested on all five achieved 68.93% accuracy. Lee et al. [15] investigated common spatial and temporal features between overt and imagined speech from seven participants and 12 words, achieving a 59.9% accuracy for overt speech and 16.2% for imagined speech, both above chance levels. They suggested potential for imagined speech classification based on overt speech features but did not attempt a direct transfer. Watanabe et al. [16] examined whether EEG acquired during speech perception and imagination shared a signature envelope with EEG from overt speech. Their study, involving 18 participants and three words, showed that classifiers trained on imagined

[1]Dataset available at https://doi.org/10.5281/zenodo.14645653

speech EEG envelopes could achieve 38.5% accuracy when tested on overt speech envelopes. These findings reinforce the similarity between the two paradigms.

Komeiji et al. [17] developed a Transformer-based model that decoded sentences from covert speech using ECoG signals. They outperformed a BiLSTM model in both overt and covert speech tasks. Interestingly, training the model on overt speech achieved similar performance for decoding covert speech compared to training on covert speech itself. This is a significant finding, as it bypasses the difficulty of collecting covert speech data. Surprisingly, to the best of our knowledge, no research has yet attempted a direct transfer of a classifier from overt to imagined speech. Our work therefore addresses this longstanding research gap, offering a more productive training process while maintaining high accuracy.

III. METHODOLOGY

Our goal is to apply transfer learning by using a model trained on overtly spoken EEG data and transferring it to classify covert (inner/imagined) speech EEG data. We used the dataset of Rekrut et al. [13] which acquired overt and covert speech EEG data with a game-like setup as described below.

A. Experimental design and procedure

1) Participants: Fifteen healthy, right-handed subjects (11 male, 4 female, average age 26.8 years) participated in the study. All participants were fluent in English. The study was approved by the Ethical Review Board of the Faculty of Mathematics and Computer Science at Saarland University.

2) Design: EEG data was recorded from participants while they interacted with a virtual robot using both overt (spoken aloud) and covert (imagined) speech. The participants viewed a bird's-eye view of the robot positioned at the center of the screen. The robot was controlled using five different command words: LEFT, RIGHT, UP, PICK, and PUSH. The experiment was divided into four sessions: two for overt speech and two for covert speech. Each session consisted of eight levels. In each level, each of the five command words was performed five times in random order, resulting in 200 trials per session (8 levels × 5 words × 5 repetitions). The sessions alternated between overt and covert conditions, to avoid recording data in large blocks, which could have led to classifying random brain states instead of actual cognitive processes. Fig. 1 illustrates the experimental paradigm.
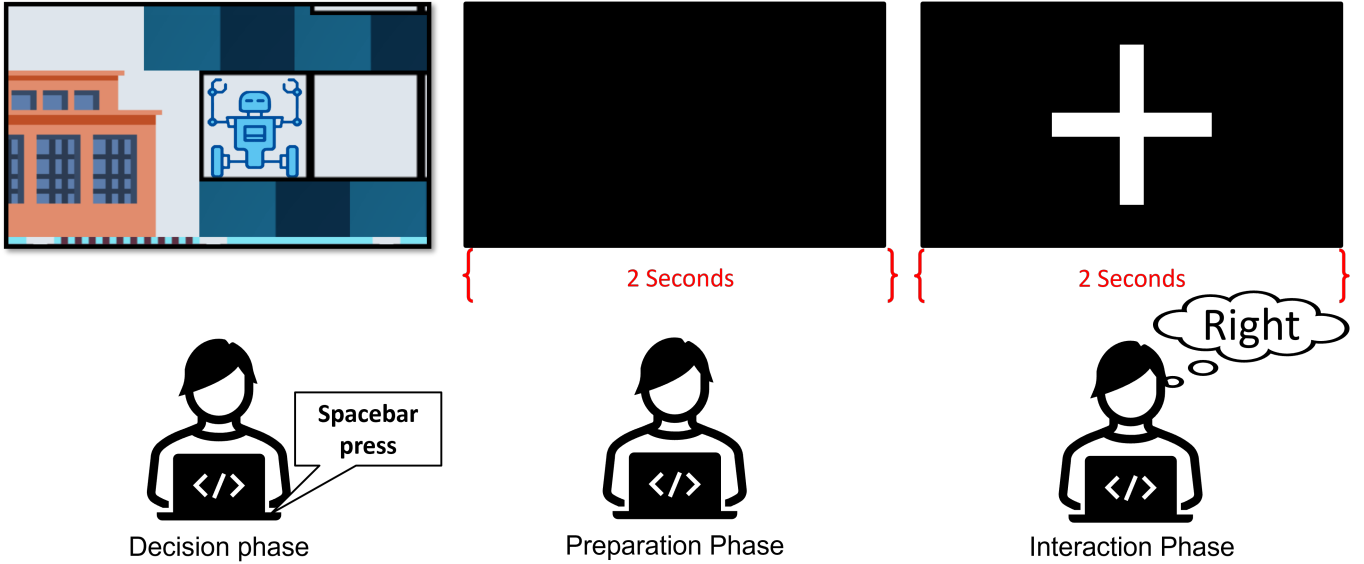

Fig. 1. Protocol for collecting imagined speech data [18]. A spacebar press triggers the interaction during overt speech, while imagined speech interactions occur in separate sessions.

3) Procedure: To allow participants to navigate the game at their own pace, we used the spacebar as the game controller (decision phase). Pressing that key caused a black screen to appear for two seconds, during which participants focused on which command they had to think (preparation phase), and when a white cross appeared in the middle of the screen for two additional seconds, as a cue, participants would produce the imagined speech (interaction phase). A two-second rest period followed, and the process was then repeated. In each level, the robot randomly presented the five action words. In the overt session, participants spoke the command aloud. In the covert session, they silently repeated the word in their mind without moving their mouth or muscles. Each command word was repeated 80 times in both the covert and overt speech sessions, resulting in 400 repetitions for each condition and a total of 800 repetitions overall.

4) Data collection: EEG data were collected in a controlled environment, with participants seated in comfortable armchairs to minimize unnecessary movements. EEG signals were recorded using a wireless 64-channel EEG system (Brain Products Live Amp 64) at a sampling rate of 500 Hz, following the 10-20 international EEG system. The robotic experimental paradigm was executed on the same Windows PC used for EEG recording, allowing for the synchronization of game events (such as spacebar presses and fixation crosses) with the EEG data. The recorded data were stored on the PC for subsequent offline analysis.

B. Data preprocessing and feature extraction

The EEG data were stored separately for each participant and individually preprocessed for each session. A 50 Hz notch filter was first applied to remove powerline interference. Subsequently, a 4th-order Butterworth Finite Impulse Response (FIR) filter with a passband of 0.5 to 80 Hz was used to filter the data. To eliminate artifacts related to eye blinks, muscles, and movement-related activities, independent component analysis (ICA) [19] was applied to the filtered EEG data. The EEG signal was then segmented into two-second epochs corresponding to the speech production phase, with a 100 ms pre-stimulus window for baseline correction. The mean signal from this 100 ms pre-stimulus window was subtracted from the corresponding epoch to minimize baseline drift and improve signal consistency. Finally, individual trials were labeled as covert or overt speech for modeling.

The EEG data during speech reflect the brain's neural response to motor articulation and cognitive processes. To extract meaningful features from EEG signals for the classification of overt and imagined speech, we employed the *Hilbert Envelope* (ENV) and *Temporal Fine Structure* (TFS) [20], [21] as key representations of EEG data. These features capture amplitude variations (using ENV, which are essential for speech-related activity) and phase information (using TFS, which is crucial for modeling neural dynamics during speech). Compared to alternative features [22] such as wavelet-based features [23], [24], common spatial patterns [5], and Riemannian geometry-based features [25], these representations provide enhanced capability for detecting subtle patterns in brain activity associated with speech tasks.

The Hilbert Envelope is the magnitude of an analytic signal $\tilde{a}(t)$, in our case EEG, and represents the slow-varying amplitude of that signal over time. It is computed as:

$$\text{ENV}(a(t)) = ||\tilde{a}(t)|| = \sqrt{a(t)^2 + \mathcal{H}\{a(t)\}^2} \tag{1}$$

where the original EEG signal $a(t)$ acts as the real part, and the Hilbert-transformed signal $\mathcal{H}\{a(t)\}$ acts as the imaginary part. $\mathcal{H}\{\cdot\}$ denotes the Hilbert transform operation.

The Hilbert transform of a signal in the time domain is defined as the convolution of $a(t)$ with $\frac{1}{\pi t}$, which can be expressed as:

$$\hat{a}(t) = \mathcal{H}\{a(t)\} = \frac{1}{\pi t} * a(t) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{a(\tau)}{t - \tau} d\tau \tag{2}$$

This envelope captures the amplitude modulation of the EEG signal, reflecting the overall energy changes that occur during speech production.

The Temporal Fine Structure, on the other hand, captures the rapid oscillations within the signal and is related to the phase information of the analytic signal. It is computed by normalizing the original EEG signal with its envelope:

$$\text{TFS}(a(t)) = \frac{a(t)}{\text{ENV}(a(t))} \tag{3}$$

The fine temporal details of the EEG signal are said to carry information about the neural dynamics underlying speech [26], [27].

The ENV and TFS features are extracted from the EEG data for each trial. Each EEG segment has an original dimension of 1000×64. After applying the ENV and TFS, each EEG segment is transformed into two feature segments of the same dimension. These segments are then concatenated horizontally, resulting in a new dimension of 1000×128.

C. Dataset splits and model description

We train subject-specific models using LSTM, Gated Recurrent Unit (GRU), BiLSTM, and Bidirectional GRU (BiGRU) architectures to classify EEG data. Each subject provided 400 EEG segments for covert speech and 400 segments for overt speech. We rely on transfer learning, where models are initially trained on overt speech EEG data. The BiLSTM model, which achieved the best performance, is then used for transfer learning. The LSTM layers of the BiLSTM model are frozen, and the fully connected layer is fine-tuned using 15%, 20%, 25% and 30% of the EEG features of covert speech to create a covert speech classification model.

The BiLSTM model begins with a sequence input layer, accepting an input size of $(B, 1000, 128)$, where B represents the batch size (we used $B = 32$), 1000 is the number of time steps, and 128 is the number of features. The model is trained with Adam optimizer using a learning rate $\eta = 0.0001$ and decay rates $\beta_1 = 0.9, \beta_2 = 0.999$. The loss function is categorical entropy. Table I shows the details of the model architecture.

TABLE I
MODEL CONFIGURATION OF BiLSTM MODEL.

Layer Type	Learnable hyperparams	Remarks
Input	None	Shape: (B, 1000, 128)
BiLSTM 1	Input weights: 4×128×512 Recurrent weights: 4×512×512 Bias: 4×512	Hidden units: 512 State fn: tanh Gate fn: sigmoid Dropout rate: 0.3
BiLSTM 2	Input weights: 4×1024×256 Recurrent weights: 4×256×256 Bias: 4×256	Hidden units: 256 State fn: tanh Gate fn: sigmoid Dropout rate: 0.2
Fully Conn.	Weights: 256×5	Size: 5 Bias: 5×1
Output	None	Softmax. Size: 5

IV. RESULTS AND DISCUSSION

A. Covert-Overt EEG decoding

Previous work [16], [28], [29] revealed spatio-temporal dynamics and waveform synchronization between covert and overt speech EEG data. This foundation guided our feature extraction, using ENV and TFS to analyze such a waveform synchronization. Fig. 2 shows the average Hilbert envelope and temporal fine structure for covert and overt speech commands of participant 10 (command: RIGHT). The ENV and TFS of the EEG data for overt and covert speech show structural similarities. The max. correlation coefficients between the EEG envelopes of covert and overt speech for commands are 0.5234 for UP, 0.6412 for LEFT, 0.6607 for RIGHT, 0.4791 for PICK, and 0.5624 for PUSH, indicating varying degrees of similarity in the envelopes of amplitude of the EEG signals.
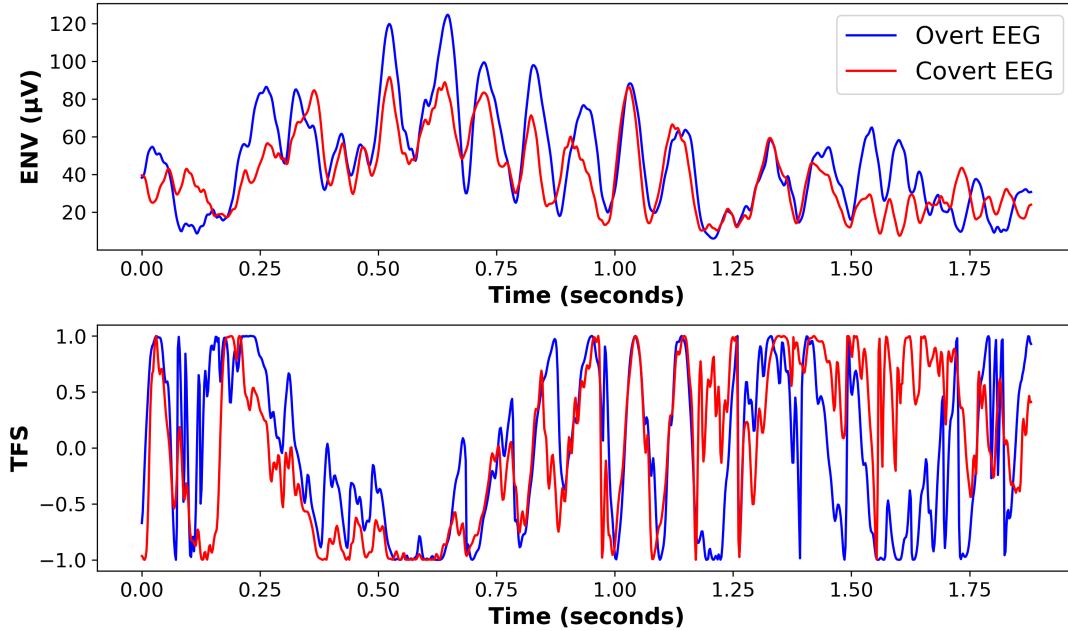

Fig. 2. ENV and TFS of the speech command RIGHT.

B. Classification results

The ENV and TFS features are combined as input to the models to learn the underlying EEG patterns during different speech tasks. Four different subject-specific classification models were trained on the overt speech EEG data using a five-fold cross-validation approach. Among these, the BiLSTM model (average accuracy of 86.44%) outperformed the others, making it the preferred model for transfer learning to covert EEG classification. The BiLSTM's superior performance is attributed to its ability to process EEG data bidirectionally, effectively capturing both forward and backward temporal dependencies while

leveraging the non-stationary and dynamic statistical characteristics of EEG signals to decode speech-related neural activity. Table II summarizes the results.

TABLE II
CLASSIFICATION ACCURACY RESULTS, IN PERCENTAGE. FOR EACH PARTICIPANT, THE BEST RESULT IS HIGHLIGHTED IN BOLD FACE.

| User | Overt | | | | Covert |
	LSTM	GRU	BiLSTM	BiGRU	BiLSTM
P1	70.37	63.78	**75.15**	64.80	58.92
P2	66.34	63.21	**71.80**	62.39	58.33
P3	**87.00**	83.50	86.30	84.50	86.21
P4	88.68	85.77	**90.65**	86.28	85.90
P5	**85.00**	74.96	84.33	73.90	85.72
P6	91.00	90.32	**93.07**	91.17	86.20
P7	94.11	93.43	**96.53**	92.47	89.74
P8	84.40	80.50	**88.46**	81.69	79.17
P9	**86.00**	80.74	85.22	79.80	73.36
P10	**95.32**	95.00	95.00	94.37	90.81
P11	90.07	91.84	**92.38**	90.90	86.55
P12	86.45	84.00	**90.49**	85.53	83.68
P13	71.28	70.03	**76.82**	71.00	67.34
P14	79.95	83.40	**84.00**	82.30	85.60
P15	75.71	77.50	**80.86**	78.60	68.59
Avg. Accuracy (%)	83.99	81.46	**86.44**	81.51	79.82

For covert speech classification, the trained BiLSTM model was used with the LSTM layer weights frozen. The fully connected layers were re-trained using 15%, 20%, 25%, and 30% of the covert EEG data. To compare the classification accuracy of the transferred covert and standard overt speech models, 20% of the unknown covert EEG data were used to test the accuracy of the transferred covert model, as the standard overt classifier was trained and tested with a 80:20 split. Classification results are illustrated in Fig. 3, showing that with 25% and 30% of the covert training data, the classifier performance did not deviate significantly ($p < .05$, paired t-test with Bonferroni correction). Furthermore, the performance of the transferred model becomes more consistent and less variable when exposed to more training examples.

We also trained the same models shown in Table II from scratch, specifically for covert speech EEG classification. Table III summarizes the results, demonstrating that transfer learning significantly improves performance. For example, while the BiLSTM model was also the best performer when trained from scratch, it was nevertheless 20 points below the BiLSTM trained with transfer learning.

TABLE III
COVERT SPEECH EEG CLASSIFICATION RESULTS (ACCURACY $\pm$ STDEV) WITHOUT TRANSFER LEARNING.

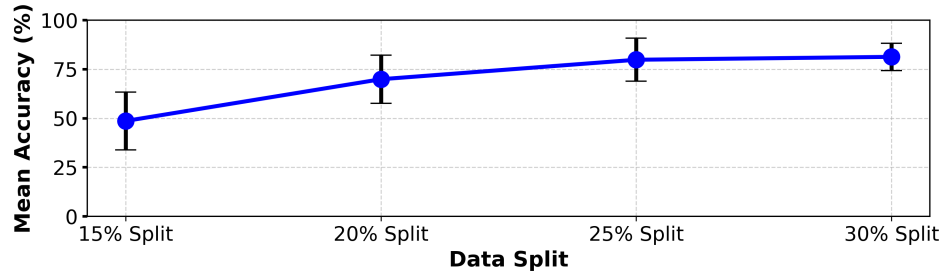|	LSTM	GRU	BiLSTM	BiGRU
Avg. Accuracy (%)	65.70 $\pm$ 3.39	65.0 $\pm$ 3.68	66.94 $\pm$ 4.65	65.83 $\pm$ 4.72

Fig. 3. BiLSTM model performance with varying data splits of covert EEG data. Error bars denote SDs.

As shown in Table IV, limited research has been conducted on the transfer of BCI models of overt speech to covert speech. Our proposed implementation outperforms previous methods. This was possible thanks to the waveform synchronization between overt and covert speech. Our work addresses the limitations of covert speech recording, including the challenges of data collection and reliability (e.g., mislabeling, variability in speech onset, offset, and speed).

TABLE IV
COMPARISON AGAINST PREVIOUS WORK.

Ref.	No. commands	Data	Classification acc.
[15]	12	EEG	16.20%
[17]	8	ECoG	46.60%
[30]	5	EEG	35.68%
[13]	5	EEG	61.78%
[18]	5	EEG	69.10%
Ours	5	EEG	**79.82%**

C. Limitations and Future Work

Since EEG responses are highly subjective, the internal decoding of overt and covert word thoughts may vary between individuals. This variability poses a challenge in creating "universal" models. Future work could focus on analyzing waveform synchronization between participants to address intersubject variability and explore speaker independence.

V. CONCLUSION

Our work emphasizes the importance of two key EEG features: Hilbert envelope and temporal fine structure, in decoding both overt and covert speech neural patterns. Our work also underscores the importance of transfer learning for accurately classifying covert speech EEG data. We find that our BiLSTM model outperforms state-of-the-art techniques, representing significant advancements for training imagined speech BCIs.

ACKNOWLEDGMENTS

REFERENCES

[1] J. Freitas, A. Teixeira, M. S. Dias, S. Silva, et al., *An introduction to silent speech interfaces*, Springer, 2017.

[2] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[3] A. Kapur, S. Kapur, and P. Maes, "Alterego: A personalized wearable silent speech interface," in *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, 2018, pp. 43–53.

[4] N. Kimura, M. Kono, and J. Rekimoto, "Sottovoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–11.

[5] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery eeg for bci," *Biomedical signal processing and control*, vol. 8, no. 6, pp. 901–908, 2013.

[6] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. M. Doñas, J. L. Pérez-Córdoba, and A. M. Gomez, "Silent speech interfaces for speech restoration: A review," *IEEE access*, vol. 8, pp. 177995–178021, 2020.

[7] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech communication*, vol. 52, no. 4, pp. 367–379, 2010.

[8] J. S. Brumberg, D. J. Krusienski, S. Chakrabarti, A. Gunduz, P. Brunner, A. L. Ritaccio, and G. Schalk, "Spatio-temporal progression of cortical activity related to continuous overt and covert speech production in a reading task," *PloS one*, vol. 11, no. 11, pp. e0166872, 2016.

[9] T. Proix, J. Delgado Saa, A. Christen, S. Martin, B. N. Pasley, R. T. Knight, X. Tian, D. Poeppel, W. K. Doyle, O. Devinsky, et al., "Imagined speech can be decoded from low-and cross-frequency intracranial eeg features," *Nature communications*, vol. 13, no. 1, pp. 48, 2022.

[10] S. Luo, Q. Rabbani, and N. E. Crone, "Brain-computer interface: applications to speech decoding and synthesis to augment communication," *Neurotherapeutics*, vol. 19, no. 1, pp. 263–273, 2023.

[11] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A state-of-the-art review of eeg-based imagined speech decoding," *Frontiers in human neuroscience*, vol. 16, pp. 867281, 2022.

[12] A. S. S. Reddy and R. B. Pachori, "Multivariate dynamic mode decomposition for automatic imagined speech recognition using multichannel eeg signals," *IEEE Sensors Letters*, 2024.

[13] M. Rekrut, A. M. Selim, and A. Krüger, "Improving silent speech bci training procedures through transfer from overt to silent speech," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, pp. 2650–2656.

[14] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, "Transfer learning in imagined speech eeg-based bcis," *Biomedical Signal Processing and Control*, vol. 50, pp. 151–157, 2019.

[15] S.-H. Lee, M. Lee, and S.-W. Lee, "Eeg representations of spatial and temporal features in imagined speech and overt speech," in *Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5*. Springer, 2020, pp. 387–400.

[16] H. Watanabe, H. Tanaka, S. Sakti, and S. Nakamura, "Synchronization between overt speech envelope and eeg oscillations during imagined speech," *Neuroscience research*, vol. 153, pp. 48–55, 2020.

[17] S. Komeiji, T. Mitsuhashi, Y. Iimura, H. Suzuki, H. Sugano, K. Shinoda, and T. Tanaka, "Feasibility of decoding covert speech in ecog with a transformer trained on overt speech," *Scientific Reports*, vol. 14, no. 1, pp. 11491, 2024.

[18] A. Mohamed Selim, M. Rekrut, M. Barz, and D. Sonntag, "Speech imagery bci training using game with a purpose," in *Proceedings of the 2024 International Conference on Advanced Visual Interfaces*, 2024, pp. 1–5.

[19] J. Iriarte, E. Urrestarazu, M. Valencia, M. Alegre, A. Malanda, C. Viteri, and J. Artieda, "Independent component analysis as a tool to eliminate artifacts in eeg: a quantitative study," *Journal of clinical neurophysiology*, vol. 20, no. 4, pp. 249–257, 2003.

[20] D. Hilbert, "Fundamentals of a general theory of linear integral equations," in *Integral equations and equations with infinitely many unknowns*, pp. 6–169. Springer, 1989.

[21] L. He and V. Dellwo, "A praat-based algorithm to extract the amplitude envelope and temporal fine structure using the hilbert transform," ISCA, 2016.

[22] M. Saeidi, W. Karwowski, F. V. Farahani, K. Fiok, R. Taiar, P. A. Hancock, and A. Al-Juaid, "Neural decoding of eeg signals with machine learning: a systematic review," *Brain Sciences*, vol. 11, no. 11, pp. 1525, 2021.

[23] S. Biswas and R. Sinha, "Wavelet filterbank-based eeg rhythm-specific spatial features for covert speech classification," *IET Signal Processing*, vol. 16, no. 1, pp. 92–105, 2022.

[24] T. Hernández-Del-Toro, C. A. Reyes-García, and L. Villaseñor-Pineda, "Toward asynchronous eeg-based bci: Detecting imagined words segments in continuous eeg signals," *Biomedical Signal Processing and Control*, vol. 65, pp. 102351, 2021.

[25] M. A. Bakhshali, M. Khademi, A. Ebrahimi-Moghadam, and S. Moghimi, "Eeg signal classification of imagined speech based on riemannian distance of correntropy spectral density," *Biomedical Signal Processing and Control*, vol. 59, pp. 101899, 2020.

[26] I. J. Moon and S. H. Hong, "What is temporal fine structure and why is it important?," *Korean journal of audiology*, vol. 18, no. 1, pp. 1, 2014.

[27] G. Ni, Z. Xu, Y. Bai, Q. Zheng, R. Zhao, Y. Wu, and D. Ming, "Eeg-based assessment of temporal fine structure and envelope effect in mandarin syllable and tone perception," *Cerebral Cortex*, vol. 33, no. 23, pp. 11287–11299, 2023.

[28] W. Zhang, M. Jiang, K. A. C. Teo, R. Bhuvanakantham, L. Fong, W. K. J. Sim, Z. Guo, C. H. V. Foo, R. H. J. Chua, P. Padmanabhan, et al., "Revealing the spatiotemporal brain dynamics of covert speech compared with overt speech: A simultaneous eeg-fmri study," *NeuroImage*, vol. 293, pp. 120629, 2024.

[29] G. M. Oppenheim and G. S. Dell, "Motor movement matters: The flexible abstractness of inner speech," *Memory & cognition*, vol. 38, no. 8, pp. 1147–1160, 2010.

[30] C. Cooney, R. Folli, and D. Coyle, "Optimizing layers improves cnn generalization and transfer learning for imagined speech decoding from eeg," in *2019 IEEE international conference on systems, man and cybernetics (SMC)*. IEEE, 2019, pp. 1311–1316.