Survey on Imagined Speech BCI with EEG and Deep Learning

Qihao Xu

mc36420@umac.mo

CCBS

ICI

University of Macau

**1.Introduction**

Brain-computer interface (BCI) is a kind of technique providing a bridge between human brain and computer to convey information. There have been many studies investigating the application of BCI in many areas like healthcare, entertainment, gaming, communication, control, research and so on. As the concept of connected environments being poplar, BCI has been one of the frontiers of the research interest. The implementation of BCI nowadays usually use deep learning techniques to analyze the brain signals from devices like ECoG, MEG, EEG, fMRI, fNIRS and so on to decode brain commands to control computers, prosthetic arms and so on. Compared with other devices, EEG is safe, cheap, portable and reliable enough which make it a good choice for BCI. Deep learning methods are popular at this time for high accuracy and good generalization compared with conventional machine learning algorithms. With the classification method of deep learning combined with other step of BCI data processing: signal acquisition, pre-processing, feature extraction, a BCI can be implemented.

Imagined speech BCI is a newer kind of BCI compared with motor imagery and other BCI. It receives brain signals from brain regions related with speech and language and decoding brain signals into voice or text for convenient control of devices or making people who cannot speak communicate with others.

Though transmitting information from computer to brain still seems impossible for now, transmitting signals from human brain to computer is already practical enough. With this technique, we can give paralyzed people one more chance to speak and control devices. Also, normal people can also use imagined speech BCI to increase their efficiency of working because of the ability of convenient control and communication. We can also get deeper understanding on neuroscience as the development of brain signal analyzation.

The objective of the article is to review the important imagined papers of the past five years and summarize them to make deeper understanding of the development of imagined speech BCI and its patterns, trends, gaps and the future direction of imagined speech development. In this article, the rest of the paper is organized as follows: Section 2 reviews the important studies of the past 5 years. Section 3 discusses the limitation and future direction of imagined speech BCI. Section 4 makes a conclusion on the whole article.

1. **Literature Review**

The most important studies of the recent five years are reviewed below and sorted by time.

In the study of (Dong-Yeon Lee, 2020), they described an issue that the increase in classification performance is limited due to the small amount of data. To tackle this issue, they proposed an end-to-end framework using Siamese neural network encoder, which learns the discriminant features by considering the distance between classes. The imagined words (e.g., arriba (up), abajo (down), derecha (right), izquierda (left), adelante (forward), and atrás (backward)) of an open access database of imagined speech from (Romero et al., 2017) were classified using the raw electro-encephalography (EEG) signals. They obtained a 6-class classification accuracy of 31.40 ± 2.73% for imagined speech, which significantly outperformed other methods at that time. This was possible because the Siamese neural network, which increases the distance between dissimilar samples while decreasing the distance between similar samples, was used. In this regard, their method can learn discriminant features from a small dataset. The proposed framework would help to increase the classification performance of imagined speech for a small amount of data and implement an intuitive communication system. But generalization to other non-Spanish datasets or larger datasets of the method is not tested.

###In the paper of (Datta & Boulgouris, 2021) they propose a framework using multi-channel convolutional neural network (MC-CNN) for recognizing the grammatical class (verb or noun) of 10 covertly spoken words from electroencephalogram (EEG) signals. Their proposed network extracts features by taking into account spatial, temporal, and spectral properties of the EEG signal. Further, sets of signals acquired from different regions of the brain are processed separately within the proposed framework and are subsequently combined at the classification stage. This approach enables the network to effectively learn discriminative features from the locations of the brain where imagined speech is processed. Their network was tested using challenging experiments, including cases where the test subject did not take part in system training. In their main application scenario, where no instance of a specific noun or verb was used during training, their method achieved 85.7% recognition. Further, their proposed method was evaluated on a publicly available EEG dataset and achieved recognition rate of 93.8% in binary classification. These results demonstrate the potential of their method. Though the method has high generalization ability, it is still doing the binary classification task, which is not practical in real-life applications. Further study on the multi-class classification should be tested on the method.

Phonological categories in articulated speech are defined based on the place and manner of articulation. In the work of (Panachakel & G, 2021),they investigate whether the phonological categories of the prompts imagined during speech imagery led to differences in phase synchronization in various cortical regions that can be discriminated from the EEG captured during the imagination. Nasal and bilabial consonant are the two phonological categories considered due to their differences in both place and manner of articulation. Binary classification of nasals and bilabials ( /piy/, “pat” and “’pot” are considered as bilabial and “/n/”, “knew” and “gnaw” as nasal category) is conducted. Mean phase coherence (MPC) is used for measuring the phase synchronization and shallow neural network (NN) is used as the classifier. As a benchmark, they also designed another NN based on statistical parameters extracted from imagined speech EEG. The result is that a 59% accuracy for MPC(Alpha), 85% accuracy for MPC(Beta), 69% accuracy for MPC(Gamma), 72% accuracy for benchmark. The NN trained on MPC values in the beta band gives classification results superior to NN trained on alpha band MPC values, gamma band MPC values and statistical parameters extracted from the EEG. They showed that nasal and bilabial consonants lead to dissimilar activations. Hence prompts orthogonal in these phonological categories are good choices as speech imagery prompts. The study shows the potential of differences in phase synchronization in various cortical regions as a feature for classification. But phonological categories cannot be used in real life application because it is much easier that the real problem.

A new kind of sequence–to–sequence model called a transformer has been applied to electroencephalogram (EEG) systems. However, the majority of EEG–based transformer models have applied attention mechanisms to the temporal domain, while the connectivity between brain regions and the relationship between different frequencies have been neglected. In addition, many related studies on imagery–based brain–computer interface (BCI) have been limited to classifying EEG signals within one type of imagery. Therefore, it is important to develop a general model to learn various types of neural representations. In the study of (Ahn et al., 2022), they designed an experimental paradigm based on motor imagery, visual imagery, and speech imagery tasks to interpret the neural representations during mental imagery in different modalities. And the speech imagery task is binary classification of words “in” and “cooperate”. They conducted EEG source localization to investigate the brain networks. In addition, they propose the multiscale convolutional transformer for decoding mental imagery, which applies multi–head attention over the spatial, spectral, and temporal domains. The proposed network shows promising performance with 72% speech imagery accuracy with the Arizona State University dataset, as compared to the conventional deep learning models. Hence, they believe that it will contribute significantly to overcoming the limited number of classes and low classification performances in the BCI system. The study fills the gap of the newest technique Transformer which don’t focus on connectivity and frequency. But the training imagined speech dataset is too small to validate the method. The method should test on many other larger datasets.

The paper of (Kaongoen et al., 2022) presents a novel brain-computer interface (BCI) system that adopts speech imagery (SI) via electroencephalograph (EEG) centered around the user’s ears (ear-EEG) to increase its wearability and usability for daily life. The ear-EEG is acquired using a custom-made wearable BCI headphone. The system extracts features from the ear-EEG by applying common spatial pattern (CSP) filters and Riemannian tangent space projections to the covariance matrices calculated from ear-EEG. Eleven individuals participated in this work. Multi-class classification is conducted on five words “up”, “down”, “next”, “back”, “power” with the method. The proposed system was evaluated through multiple sessions of both offline and online experiments that were designed to allow participants to control an interactive simulated home appliance. In the offline experiment, all participants were able to achieve a classification accuracy significantly higher than the chance level. In the online experiments, a few participants were able to use the proposed system to freely control the home appliance with high accuracy and relatively fast command delivery speed. The best participant achieved an average true positive rate and command delivery time of 85% and 3.79 s/command, respectively. Based on the positive experimental results and user surveys, the novel ear-EEG-SI-based BCI paradigm is a promising approach for the wearable BCI system for daily life. The Ear-EEG is an innovative device which is convenient to use have the potential to be used in real-life. But the position of the electrodes are far from the brain cortex which will affect accuracy, and the accuracy on other larger datasets should be tested.

Invasive devices have recently led to major milestones in this regard: deep-learning algorithms trained on intracranial recordings can now start to decode elementary linguistic features such as letters, words and audio-spectrograms. However, extending this approach to natural speech and non-invasive brain recordings remains a major challenge. In the study of (Défossez et al., 2023) they introduce a model trained with contrastive learning to decode self-supervised representations of perceived speech from the non-invasive recordings of a large cohort of healthy individuals. To evaluate this approach, they curate and integrate four public datasets, encompassing 175 volunteers recorded with magnetoencephalography or electro-encephalography while participants listened to short stories and isolated sentences. The results show that their model can identify, from 3 seconds of electroencephalography signals, accuracy of 25.7%±2.9 on Brennan dataset and 17.7%±0.6 on Broderick dataset. The comparison of their model with a variety of baselines highlights the importance of a contrastive objective, pretrained representations of speech and a common convolutional architecture simultaneously trained across multiple participants. Finally, the analysis of the decoder’s predictions suggests that they primarily depend on lexical and contextual semantic representations. Overall, this effective decoding of perceived speech from non-invasive recordings delineates a promising path to decode language from brain activity, without putting patients at risk of brain surgery. The method is innovative considering directly transform EEG signals to sound waves. The dataset is recorded when participants listening to speech rather than imagining on their own will, this method should be tested on imagined speech with the participants’ own will to make the BCI applicable. And the accuracy is still too low for real-life applications.

In a recent study of auditory evoked potential (AEP) based brain–computer interface (BCI), it was shown that, with an encoder–decoder framework, it is possible to translate human neural activity to speech (T-CAS). Current encoder–decoder-based methods achieve T-CAS often with a two-step approach where the information is passed between the encoder and decoder with a shared vector of reduced dimension, which, however, may result in information loss. In the paper of (Guo et al., 2023), they propose an end-to-end model to translate human neural activity to speech (ET-CAS) by introducing a dual–dual generative adversarial network (Dual-DualGAN) for cross-domain mapping between electroencephalogram (EEG) and speech signals. In this model, they bridge the EEG and speech signals by introducing transition signals which are obtained by cascading the corresponding EEG and speech signals in a certain proportion. We then learn the mappings between the speech/EEG signals and the transition signals. The resulted accuracy is 78.53%. We also develop a new EEG dataset where the attention of the participants is detected before the EEG signals are recorded to ensure that the participants have good attention in listening to speech utterances. The proposed method can translate word-length and sentence-length sequences of neural activity to speech. Experimental results show that the proposed method significantly outperforms state-of-the-art methods on both words and sentences of auditory stimulus. The consideration of the participants’ attention is an innovative aspect, which can increase the quality of the recorded EEG data, this will of great help for future BCI study. But the problem is that the collected EEG imagined speech data is not conducted by positive imagining of the participants, which may not be useful in real-life application.

In the study from (Jeong et al., 2023), the human mind directly decoded the neural languages based on speech imagery using the proposed deep neurolinguistic learning. A multi-class classification of 8 words (“ I”, “partner”, “move”,“have”, “drink”, “box”, “cup” and “phone”) is conducted. It reached an average success rate of each task represented approximately 72.36% of the performance. Through real-time experiments, they evaluated whether BCI-based cooperative tasks between multiple users could be accomplished using a variety of neural languages. They successfully demonstrated a BCI system that allows a variety of scenarios, such as essential activity, collaborative play, and emotional interaction. This outcome presents a novel BCI frontier that can interact at the level of human-like intelligence in real time and extends the boundaries of the communication paradigm. Generation of neural language as a sentence-level form is quite innovative but realizing prosthetic arm control should be more related with motor imagery rather than speech imagery.

Decoding imagined speech had limited success, mainly because neural signals are weak and more variable than overt speech, hence challenging to decode by machine-learning (ML)-based algorithms. In recent years, deep learning (DL) with convolutional neural networks (CNNs) has transformed computer vision and can perform pattern recognition better than the traditional ML-based algorithms. The objective of the article of (Kamble, Ghare, et al., 2023a) is to design a smoothed pseudo-Wigner–Ville distribution (SPWVD) and CNN-based automatic imagined speech recognition (AISR) system to recognize imagined words. This article uses a publicly available 64-channel EEG dataset, collected from 15 healthy subjects for three categories: long words, short words, and vowels. The EEG signals were transformed into time–frequency representation (TFR) using SPWVD, which are used as an input to CNN such that the EEG dataset was identified and classified into binary and multiclass categories. In addition, the CNN model was optimized using a recently developed Keras-tuner library to achieve optimal performance. The performance of the SPWVD-CNN-driven AISR system is evaluated using seven performance evaluation metrics: accuracy (ACC), recall (REC), precision (PREC), Mathew’s correlation coefficient (MCC), Cohen’s kappa (κ), F1-score, and area under the curve (AUC). It is found that the proposed system achieved the maximum classification ACCs of 94.82%, 94.26%, 94.68%, and 84.50% for long words, short–long words, short words, and vowels, respectively. The accuracy of the method is very high but binary classification is far from real-life applications. Study on multi-class classification of long words should be done on multi-class classification with the method.

Extracting meaningful information from the raw EEG signal is a challenging task due to the nonstationary nature of EEG signals. Decomposing a signal into several sub-bands (SBs) using rational dilation wavelet transform (RADWT) requires selecting predefined factual parameters, which is an arduous task. The main objective of the study from (Kamble, Ghare, et al., 2023b) is to propose an adaptive RADWT method capable of decomposing EEG signals by adaptively selecting the tuning parameters and classifying the EEG signals into distinct categories. The optimum tuning parameters of RADWT are obtained using particle swarm optimization (PSO) and used to decompose the EEG signals into several SBs. Several statistical features are elicited from each SB and used to input six different machine learning (ML) algorithms. This work uses a 64-channel EEG dataset recorded from 15 healthy people for three categories: long words, short words, and vowels. The performance of the proposed AISR system is evaluated using seven performance evaluation metrics: accuracy, recall, precision, Cohen’s kappa, F1-score, and area under the curve. The proposed system achieved the average classification accuracies of 87.26% ± 1.12%, 89.23% ± 0.95%, 95.5% ± 0.68%, and 92.16% ± 0.83% for long words, short–long words, short words, and vowels, respectively. When compared with the existing state-of-the-art, the proposed non-parametric decomposition approach and the Bagging algorithm achieved a 3%–5% improvement. The performance of the proposed method is validated using an open-access dataset. The accuracy of the method is also very high but binary classification is far from real-life applications. Study on multi-class classification of long words should be done on multi-class classification with the method.

###The article of (Kamble, Ghare, Kumar, et al., 2023) investigates the feasibility of spectral characteristics of the electroencephalogram (EEG) signals involved in imagined speech recognition. Eleven subjects were recruited to perform the speech imagination task. This article analyses the spectral features for binary and multiclass classification of imagined words in six different frequency bands (FBs). The 1-D EEG signals were converted into time–frequency representation (TFR) plots using smoothed pseudo-Wigner–Ville distribution (SPWVD) and classified using a convolutional neural network (CNN). In addition, the analysis was performed for subject-dependent, subject-independent, and leave-one-subject-out (LOSO) approaches along with the all-data approach. The proposed method achieved promising results in the gamma band with a binary classification accuracy of 82.04% ± 2.45%, 81.66% ± 4.93%, 78.97% ± 3.12%, and 81.04% ± 3.08% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively, and a multiclass classification accuracy of 51.44% ± 3.55%, 50.20% ± 1.35%, 49.93% ± 1.72%, and 50.42% ± 2.18% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively. Finally, the multiclass scalability in decoding the imagined words is investigated by increasing the number of classes from 2 to 15. The study’s findings demonstrate that the EEG-based imagined speech recognition using spectral analysis has the potential to be an effective tool for speech recognition in practical BCI applications. The design of the paradigm is great which applicable for the paralyzed patients who can be main consumer of BCI. But the binary classification is not necessary when there is multi-class classification considering real-life demand.

There are many methods to analyze speech imagery signals, but those based on deep neural networks achieve the best results. However, more research is necessary to understand the properties and features that describe imagined phonemes and words. In the paper of (Macias-Macias et al., 2023), they analyze the statistical properties of speech imagery EEG signals from the KaraOne dataset to design a method that classifies imagined phonemes and words. With this analysis, they propose a Capsule Neural Network that categorizes speech imagery patterns into bilabial, nasal, consonant-vocal, and vowels /iy/ and /uw/. The method is called Capsules for Speech Imagery Analysis (CapsK-SI). The input of CapsK-SI is a set of statistical features of EEG speech imagery signals. The architecture of the Capsule Neural Network is composed of a convolution layer, a primary capsule layer, and a class capsule layer. The average accuracy reached is 90.88%±7 for bilabial, 90.15%±8 for nasal, 94.02%±6 for consonant– vowel, 89.70%±8 for word-phoneme, 94.33%± for /iy/ vowel and, 94.21%±3 for /uw/ vowel detection. Finally, with the activity vectors of the CapsK-SI capsules, they generated brain maps to represent brain activity in the production of bilabial, nasal, and consonant-vocal signals. Classifying phonemes rather than words is more practical than classifying words because recording the EEG signal of every word seems impossible. But classify phonemes into many phoneme classes is not so that useful in real life. The accuracy of multi-class classification of recognizing each phoneme should be tested on the method.

Translating imagined speech from human brain activity into voice is a challenging and absorbing research issue that can provide new means of human communication via brain signals. Efforts to reconstruct speech from brain activity have shown their potential using invasive measures of spoken speech data but have faced challenges in reconstructing imagined speech. In the paper of (Young-Eun Lee, 2023), they propose NeuroTalk, which converts non-invasive brain signals of imagined speech into the user’s own voice. Their model was trained with spoken speech EEG which was generalized to adapt to the domain of imagined speech, thus allowing natural correspondence between the imagined speech and the voice as a ground truth. In their framework, an automatic speech recognition decoder contributed to decomposing the phonemes of the generated speech, demonstrating the potential of voice reconstruction from unseen words. Though the paradigm is great which fits real-life demand. Using spoken speech may be worse than pure imagination for the influence of signal of speak.

Advances in deep learning have shown great promise towards the application of performing high-accuracy Electroencephalography (EEG) signal classification in a variety of tasks. However, many EEG-based datasets are often plagued by the issue of high inter-subject signal variability. Robust deep learning models are notoriously difficult to train under such scenarios, often leading to subpar or widely varying performance across subjects under the leave-one-subject-out paradigm. Recently, the model agnostic meta-learning framework was introduced to increase the model’s ability to generalize towards new tasks. While the original framework focused on task-based meta-learning, the research from (Ng & Guan, 2024) aims to show that the meta-learning methodology can be modified towards subject-based signal classification while maintaining the same task objectives and achieve state-of-the-art performance. Namely, they propose the novel implementation of a few/zero-shot subject-independent meta-learning framework towards multi-class inner speech and binary class motor imagery classification. Compared to current subject-adaptive methods which utilize large number of labels from the target, the proposed framework shows its effectiveness in training zero-calibration and few-shot models for subject-independent EEG classification. The proposed few/zero-shot subject-independent meta-learning mechanism performs well on both small and large datasets and achieves robust, generalized performance across subjects. The results obtained shows a significant improvement over the current state-of-the-art, with the binary class motor imagery achieving 88.70% and the accuracy of multi-class inner speech achieving an average of 31.15%. Using meta-learning is a good attempt, this may accelerate the development of BCI study. But the accuracy of imagined speech on this method is too low. Further research should be done on the usage of meta-learning in BCI.

Table1: Imagined speech review summary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Reference | Task | Main Methods | Results | Strengths | Limitations |
| (Dong-Yeon Lee, 2020) | Multi-class classification of 6 Spanish words | End-to-end framework using  Siamese neural network encoder | 6-class classification accuracy of an average of 31.40 ± 2.73% | The method performs well in small dataset | Generalization on other, non-Spanish or larger datasets not tested |
| (Datta & Boulgouris, 2021) | Binary classification on nouns and verbs from 5 nouns and 5 verbs | Multi-channel convolutional neural network (MC-CNN) | 85.7% accuracy and 93.8% accuracy on KaraOne dataset | High generalization | Binary classification of grammatical class is not applicable in real-life |
| (Panachakel & G, 2021) | Binary classification of nasals and bilabials ( /piy/, “pat” and “’pot” are considered as bilabial and “/n/”, “knew” and “gnaw” as nasal category) | Mean phase coherence (MPC) is used  for measuring the phase synchronization and shallow neural  network (NN) is used as the classifier | 59% accuracy for MPC(Alpha), 85% accuracy for MPC(Beta), 69% accuracy for MPC(Gamma), 72% accuracy for benchmark. | Shows the potential of differences in phase synchronization in various cortical regions as a feature for classification | Phonological categories cannot be used in real life application |
| (Ahn et al., 2022) | Binary classification of words “in” and “cooperate” | Multiscale convolutional transformer | 72% speech imagery accuracy with Arizona  State University dataset | The study fills the gap of the newest technique Transformer which don’t focus on connectivity and frequency. | The training imagined speech dataset is too small to validate the method. The method should test on many other larger datasets. |
| (Kaongoen et al., 2022) | Multi-class classification of five words “up”, “down”, “next”, “back”, “power” | Common spatial pattern (CSP) filters and Riemannian tangent  space projections to the covariance matrices | The average TPRs (True Positive Rate) across all tasks of all participants for first, second and final sessions  were 0.44, 0.45, and 0.59, respectively | Innovative and convenient device which have the potential to be used in real-life | The position of electrodes will affect accuracy and limited dataset size |
| (Défossez et al., 2023) | Zero-shot classification CLIP loss of 3s of speech signal and 3s of EEG signal listening to stories or sentences | Contrastive learning | 25.7%±2.9 on Brennan dataset and 17.7%±0.6 on Broderick dataset | Innovative method | Low accuracy, the paradigm is not positively imagining speech |
| (Guo et al., 2023) | Transition from EEG signal listening to speech file to speech | Dual–dual  generative adversarial network (Dual-DualGAN) | Accuracy rate: 78.53  PCC: 0.838  MCD: 3.793 | Consideration of attention | The imagination process is not positive conducted by the participants |
| (Jeong et al., 2023) | Multi-class classification of 8 words (“ I”, “partner”, “move”,“have”, “drink”, “box”, “cup” and  “phone”) | Real-Time Deep Neurolinguistic Learning | Average success rate of each task represented approximately 72.36% of the performance | Innovative sentence level generation | The control of prosthetic arm should be more related with motor imagery rather than speech imagery |
| (Kamble, Ghare, et al., 2023a) | Binary classification for long words(“Independent” “Cooperate”) and short words(“IN”, “OUT” and “UP”), multi-class classification for short words and vowels(“a” “i” and “u”) | Smoothed pseudo-Wigner–Ville distribution (SPWVD) and CNN | Accuracy of 94.82%, 94.26%,  94.68%, and 84.50% for long words, short–long words, short  words, and vowels, respectively | High accuracy | Multi-class classification of long words not tested |
| (Kamble, Ghare, et al., 2023b) | Binary classification for long words(“Independent” “Cooperate”) and short-long (short and long) words, multi-class classification for short words (“IN”, “OUT” and “UP”) and vowels(“a” “i” and “u”) | Adaptive rational dilation wavelet transform (RADWT) | Average  classification accuracies of 87.26% ± 1.12%, 89.23% ± 0.95%,  95.5% ± 0.68%, and 92.16% ± 0.83% for long words, short–long  words, short words, and vowels, respectively | High accuracy | Multi-class classification of long words not tested |
| (Kamble, Ghare, Kumar, et al., 2023) | Multi-class classification of 15 words (“HELP,” “LIGHT,” “PAIN,” “STOP,” “YES,” “NO,” “RIGHT,”  “LEFT,” “THANK YOU,” “BACKWARD,” “DOWN,” “TOILET,” “TELEVISION,” “WATER,” and “MEDICINE.”) and binary classification of each kind of combination of two of the 15 words | Spectral analysis | Gamma band with a binary classification accuracy of 82.04% ±2.45%, 81.66% ± 4.93%, 78.97% ± 3.12%, and 81.04% ±3.08% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively, and a multiclass classification accuracy of 51.44% ± 3.55%, 50.20% ± 1.35%, 49.93% ±1.72%, and 50.42% ± 2.18% in all-data, subject-dependent, subject-independent, and LOSO approaches, respectively | Good paradigm | Binary classification not necessary |
| (Macias-Macias et al., 2023) | Multi-class classification of phonemes and words into bilabial, nasal, consonant-vocal, and vowels/iy/ and/uw/. (/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, and /n/, /pat/, /pot/, /knew/, /gnaw/) | Capsule Neural Network | Average accuracy reached is 90.88%±7 for bilabial, 90.15%±8 for nasal, 94.02%±6 for consonant–vowel, 89.70%±8 for word-phoneme, 94.33%± for/iy/ vowel and, 94.21%±3 for/uw/ vowel | Using phonemes is more practical than words | Classifying phoneme class is not so that useful in real-life |
| (Young-Eun Lee, 2023) | Multi-class classification of pronounced and imagined 12 words (Ambulance, clock, hello, help me, light, pain, stop,  thank you, toilet, TV, water, and yes) | NeuroTalk | root mean square error (RMSE), spokenEEG:0.17, imaginedEEG:0.18, Unseen spokenEEG:0.19, Unseen imaginedEEG:0.19 | Good paradigm | Spoken speech EEG signal may cause noise |
| (Ng & Guan, 2024) | Multi-class classification of four words (‘up’ ,‘down’, ‘left’ and ‘right’) | Subject-independent meta-learning framework | Accuracy of multi-class inner speech  achieving an average of 31.15% | Innovative meta-learning | Low accuracy |

**3. Discussion**

The review above discloses important papers of BCI with EEG and Deep Learning in the past five years. As we can see, most of the papers above use CNN, GAN, decoder and other popular deep learning methods. Some of them also focus on specific features related with EEG or neuroscience, but few of the methods are specifically designed for BCI with EEG or even for imagined speech. With more specification, we can design methods better than these normal and non-specific deep learning methods, which can better fit the demand of BCI. The future study of BCI should focus more on special features of brain which may give inspiration to the development of BCI.

Also, most the datasets used now contain only few words or phonemes. There haven’t been any dataset contains all phonemes or even words (which is too huge to collect, so recording phoneme data will be more practical). Only with enough data, then we can train generalizable enough models. The data for now is far from enough, we should also do more work on data collection which will help in future BCI study. The future collection of data should be mainly phonemes because its limited and acceptable number rather than words or even sentences that will be expensive to collect because of their huge amount. With the combination of phonemes, we can get infinite number of pronunciations of words then we can reach practical BCI system generation voice.

Nowadays, the use of BCI in many areas like healthcare, entertainment, gaming, communication, control, research and so on have been done many studies. All the mentioned areas need highly accurately recognizing many kinds of signals. Despite of the demand of online response time for BCI considering most of the studies above are offline, many of the important papers above still use binary classification and the methods using multi-class classification are with low accuracy which cannot satisfy the demand of the usages above. The road to BCI that can be used in real life circumstances is still a long way to go.

**4. Conclusion**

In conclusion, this article shows the condition of the imagined speech BCI in the past 5 years. We found that though many up-to-date deep learning methods are used in these studies, the study of imagined speech BCI is still in primary stage for shortage of data and high-accuracy BCI multi-classification algorithms. And using phonemes rather than words is a more practical choice for paradigm. We should develop more BCI-specific algorithms, focusing on more brain features, to make imagined speech BCI applicable in real-life and a more connected and convenient world.

**References**

Ahn, H. J., Lee, D. H., Jeong, J. H., & Lee, S. W. (2022). Multiscale Convolutional Transformer for EEG Classification of Mental Imagery in Different Modalities. *IEEE Trans Neural Syst Rehabil Eng*, *PP*. <https://doi.org/10.1109/TNSRE.2022.3229330>

Datta, S., & Boulgouris, N. (2021). Recognition of grammatical class of imagined words from EEG signals using convolutional neural network. *Neurocomputing*, *465*, 301-309. <https://doi.org/10.1016/j.neucom.2021.08.035>

Défossez, A., Caucheteux, C., Rapin, J., Kabeli, O., & King, J.-R. (2023). Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, *5*(10), 1097-1107. <https://doi.org/10.1038/s42256-023-00714-5>

Dong-Yeon Lee, M. L., Seong-Whan Lee. (2020). Classification of Imagined Speech Using Siamese Neural Network. *IEEE International Conference on Systems, Man, and Cybernetics(SMC)*.

Guo, Y. A., Liu, T., Zhang, X. F., Wang, A. H., & Wang, W. W. (2023). End-to-end translation of human neural activity to speech with a dual-dual generative adversarial network. *Knowledge-Based Systems*, *277*. <https://doi.org/ARTN> 110837

10.1016/j.knosys.2023.110837

Jeong, J. H., Cho, J. H., Lee, B. H., & Lee, S. W. (2023). Real-Time Deep Neurolinguistic Learning Enhances Noninvasive Neural Language Decoding for Brain-Machine Interaction. *IEEE Trans Cybern*, *53*(12), 7469-7482. <https://doi.org/10.1109/TCYB.2022.3211694>

Kamble, A., Ghare, P. H., & Kumar, V. (2023a). Deep-Learning-Based BCI for Automatic Imagined Speech Recognition Using SPWVD. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-10. <https://doi.org/Artn> 4001110

10.1109/Tim.2022.3216673

Kamble, A., Ghare, P. H., & Kumar, V. (2023b). Optimized Rational Dilation Wavelet Transform for Automatic Imagined Speech Recognition. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-10. <https://doi.org/Artn> 4002210

10.1109/Tim.2023.3241973

Kamble, A., Ghare, P. H., Kumar, V., Kothari, A., & Keskar, A. G. (2023). Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition. *IEEE Transactions on Instrumentation and Measurement*, *72*, 1-9. <https://doi.org/Artn> 4009409

10.1109/Tim.2023.3300473

Kaongoen, N., Choi, J., & Jo, S. (2022). A novel online BCI system using speech imagery and ear-EEG for home appliances control. *Comput Methods Programs Biomed*, *224*, 107022. <https://doi.org/10.1016/j.cmpb.2022.107022>

Macias-Macias, J. M., Ramirez-Quintana, J. A., Chacon-Murguia, M. I., Torres-Garcia, A. A., & Corral-Martinez, L. F. (2023). Interpretation of a deep analysis of speech imagery features extracted by a capsule neural network. *Comput Biol Med*, *159*, 106909. <https://doi.org/10.1016/j.compbiomed.2023.106909>

Ng, H. W., & Guan, C. (2024). Subject-independent meta-learning framework towards optimal training of EEG-based classifiers. *Neural Netw*, *172*, 106108. <https://doi.org/10.1016/j.neunet.2024.106108>

Panachakel, J. T., & G, R. A. (2021). Classification of Phonological Categories in Imagined Speech using Phase Synchronization Measure. *Annu Int Conf IEEE Eng Med Biol Soc*, *2021*, 2226-2229. <https://doi.org/10.1109/EMBC46164.2021.9630699>

Romero, E., Lepore, N., Brieva, J., Larrabide, I., Pressel Coretto, G. A., Gareis, I. E., & Rufiner, H. L. (2017). *Open access database of EEG signals recorded during imagined speech* 12th International Symposium on Medical Information Processing and Analysis,

Young-Eun Lee, S.-H. L., Sang-Ho Kim, Seong-Whan Lee. (2023). *Towards Voice Reconstruction from EEG during Imagined Speech* The Thirty-Seventh AAAI Conference on Artificial Intelligence (AAAI-23),