# EEG-Based Silent Speech Interface and its Challenges: A Survey

Nilam Fitriah, Hasballah Zakaria, Tati Latifah Erawati Rajab*
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia

***Abstract*—People with speech disorders could have social and welfare difficulties. Therefore, the silent speech interface (SSI) is needed to help them communicate. This interface decodes the speech from a human's biosignal. The brain signals contain information from speech production to cover people with numerous speech disorders. Brain signals can be acquired non-invasively by electroencephalograph (EEG) and later transformed into the features for the input of speech pattern recognition. This review discusses the advancement of EEG-based SSI research and its current challenges. It mainly discussed the acquisition protocol, spectral-spatial-temporal characterization of EEG-based imagined speech, classification techniques with leave-one-subject or session-out cross-validation, and related real-world environmental issues. It aims to aid future imagined speech decoding research in exploring the proper methods to overcome the problems.**

***Keywords—Imagined speech; silent speech interface; electroencephalograph (EEG); speech recognition***

## I. INTRODUCTION

Communication is essential in daily human life. People would hardly communicate in noisy circumstances, in quiet environments where no sounds are allowed, in secret conversation, or when they have speech disorders.

Speech disorders could negatively affect a person's social life and welfare. WHO reported that in 2011 there was 3,6% of the world's population experienced extreme difficulty living in their community due to speech disorders [1]. Moreover, they also were hindered from getting a job, as stated by ILO in 2017 for 4,1% of Indonesian citizens [2]. Additionally, research conducted in the United States found that one in 13 adults experience speech disorders annually [3]. In this circumstance, an interface to assist communication becomes more necessary than ever.

One of the interfaces intended to help people with speech impairments to communicate is the silent speech interface (SSI), which converts the biosignal into a speech. Human speech can be categorized into overt speech (with sound), silent articulation (articulator moves but no sound), and covert speech (no sound and movement) [4]. The latter is also called silent speech or imagined speech as our focus of discussion.

The causes of speech disorders can be the absence of knowledge to speak experienced by deaf people, articulation problems, neurologic dysfunction (e.g., stroke), and paralysis (i.e., tetraplegia, muscular degenerative diseases, locked-in syndrome, or coma patients) [4]. Most causes come from brain disorders [5]. The applicable sensors mainly record brain activity, such as electroencephalograph (EEG) or electro-corticograph (ECoG). While ECoG has a higher Signal-to-Noise Ratio (SNR) than EEG, its invasive electrode placement can have a clinical risk. Moreover, ECoG only covers a specific area, while EEG has a broader coverage than ECoG. Hence, EEG is considered safer than ECoG.

This review aims to present the development of EEG-based speech imagery studies and assist researchers in finding a solution to achieve better accuracy and solve the real problem. It focused on EEG application to decode imagined speech as a pipeline consisting of signal acquisition, signal preprocessing, feature extraction, classification techniques, and the real-world application challenges that were still unnoticed. The process flow of the review is in Fig. 1, and the summary of the discussed references is in Table I.

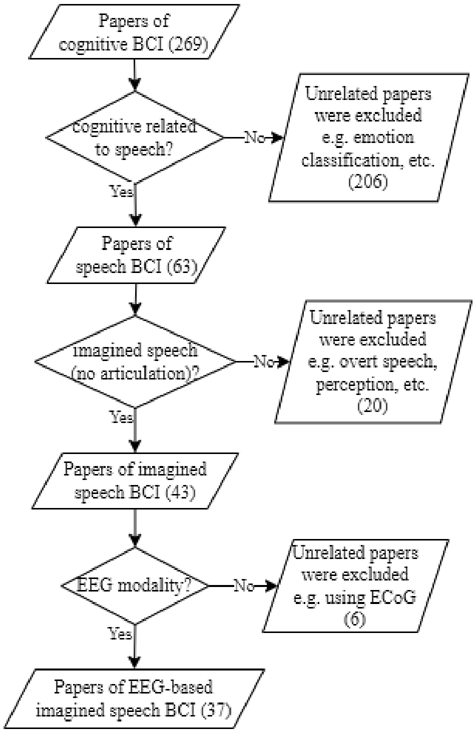
Papers of cognitive BCI (269)
cognitive related to speech?
No
Unrelated papers were excluded e.g. emotion classification, etc. (206)
Yes
Papers of speech BCI (63)
imagined speech (no articulation)?
No
Unrelated papers were excluded e.g. overt speech, perception, etc. (20)
Yes
Papers of imagined speech BCI (43)
EEG modality?
No
Unrelated papers were excluded e.g. using ECoG (6)
Yes
Papers of EEG-based imagined speech BCI (37)

Fig. 1. Review Processes.

*Corresponding Author

This work is financially supported by Direktorat Riset, Teknologi, dan Pengabdian Kepada Masyarakat, Direktorat Jenderal Pendidikan Tinggi, Riset, dan Teknologi, Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi Republik Indonesia, and Institute for Research and Community Services (LPPM-ITB) with contract 347/IT1.B07.1/SPP-LPPM/V/2022.

TABLE I. EEG-BASED SSI STUDIES

| **Ref.** | **Cue** | **Speech** | **Freq. (Hz)** | **Feature extraction** | **Classification** | **Validation** |
|---|---|---|---|---|---|---|
| [6] | Audiovisual | “to”, “two”, “too”, “here”, “hear” | 1-20 | FFT | Min. least-squares | CV |
| [7] | Visual | 0-9; NATO phonetic alphabet (a-e); "yes", "no", etc.; sentence | 0.9-60 | STFT+LDA | HMM | CV |
| [8] | Visual | “alpha”, “bravo”, …, “echo” | 1-300 | DTCWT, LDA | HMM | CV |
| [9] | Audiovisual | “/ba/”, “/ku/” | 3-18 | Hilbert transform, spectral feature | matched-filter | CV |
| [10] | | | 3-20 | spot of interest (SOI) | LDA | CV |
| [11] | | | 4-25 | AR | kNN | CV |
| [12] | Visual | “/a/”, “/u/” | 1-45 | CSP | SVM | CV |
| [13] | | | | Statistics, geometric mean, energy sum, entropy, wavelength | LDA | CV |
| [14] | | | 8-40 | MFCC | kNN/SVM | CV |
| [15] | Audio | “/aa/”, “/ae/”, “/l/”, “/r/”, “/m/”, “/n/”, “/uu/”, “/ow/”, “/s/”, “/z/” | 4-28 | Spectrogram | LDA | CV |
| [16] | Visual | “two”, “to”, “four”, “for” | N/A | Voltage | DT | CV |
| [17] | Visual | (Korean) 3 (“sam”), 5(“oo”), 9 (“gu”), 10 (“sib”); cheek (“ppyam”), nose (“ko”), eye (“nun”), mouth (“ib”) | 1-100 | spectrogram, STFT | SVM | CV |
| [18] | Visual | (Chinese) “左” (“left”), “壹”(1) | 6-30 | CSP, DWT, AR | SVM | CV |
| [19] | Visual | /um/ | 4-20 | AR | LDA | CV |
| [20] | Audio (question) | (Arabic) “yes”, “no” | 0-48 | DWT | SVM, SOM, LDA | CV |
| [21] | Audio | “/a/”,“/i/”,“/u/”,“/e/”,“/o/” | 1-100 | Statistics | EL | CV |
| [22] | Audio (question) | (Indian & English): “yes”, “no” | 0-40 | Spectral power of FFT | ANN | CV |
| [23] | Audiovisual | “/iy/”, “/uw/”, “/piy/”, “/tiy/”, “/diy/”, “/m/”, “/n/”, “pat”, “pot”, “knew”, “gnaw” | 1-50 | Statistics | SVM | CV |
| [24] | | | | Statistics, MFCC, nonlinear features | SVM | CV |
| [25] | | | | DWT | DNN | CV |
| [26] | Visual | “/a/”, “/i/”, “/u/”, “in”, “out”, “up”, “cooperate”, “independent” | 8-70 | Riemannian manifold | RVM | CV |
| [27] | | | | channel cross-covariance (CCV) | CNN+LSTM+DAE | CV |
| [28] | | | | channel cross-correlation matrix | LSTM | CV |
| [29] | | | | DWT | DNN | CV |
| [30] | | | | Bag of Features (BoF) | RNN | LOSO-CV |
| [30] | Visual | (Spain): “arriba” (“up”), “abajo” (“down”), “izquierda” (“left”), “derecha” (“right”), “seleccionar” (“select”) | 4-25 | | | |
| [31] | | | 4-25 | DWT | RF | CV |
| [32] | | | 0-64 | Statistics, RWE | | |
| [33] | | | 40-50 | Bag of Features (BoF) | NB + TL | CV |
| [34] | Audiovisual | (Spain) “/a/”, “/e/”, “/i/”, “/o/”, “/u/”, “arriba” (“up”), “abajo” (“down”), “izquierda” (“left”), “adelante” (“forward”), “atrás” (“backward”) | 2-40 | RWE | RF | CV |
| [35] | | | | | CNN | CV |
| [36] | | | | CNN layer, FBCSP | | |
| [37] | | | | | CNN + TL | LOSO-CV |
| [38] | | | | word embedding + Siamese encoder | kNN | CV |
| [39] | Visual | “/a/”, “/e/”, “/i/”, “/o/”, “/t/” | 0.5-220 | phase per band (Hilbert transform) | SVM | CV |
| [40] | Audio | “/a/”, “/e/”, “/i/”, “/o/”, “/u/”, “yes”, “no”, “left”, “right”. | 0.1-70 | RMS, zero-crossing rate, moving window average, kurtosis, and power spectral entropy | RNN | CV |
| [41] | Audio | “Hi Bixby”, “Call Mom”, “Open Camera”, “What’s the weather”. | 0.5-70 | | | |
| [42] | Audio | “ambulance”, “clock”, “hello”, “yes”, “light”, “help me”, “pain”, “stop”, “thank you”, “toilet”, “TV”, “water”. | 0.5-40 | CSP | LDA | CV |
| [43] | Audio | “go”, “back”, “left”, “right”, and “stop” | 0.5-60 | covariance and MaxLCor | ELM | CV |
| [44] | Audio | “hello”, “help me”, “stop”, “yes”, “thank you” | 0.5-128 | DWT, MaxLCor | SVM | CV |
| [45] | Visual | ten words for every vowel: “a” (“can”, …, “tap”), “e” (“bed”, …, “vex”), etc. | 0.5-50 | coherence, PDC, DTF, transfer entropy | DBN | LOSO-CV |

ANN = artificial neural network (NN), AR = autoregression, CNN = convolutional NN, CSP = common spatial pattern, DAE = Deep Autoencoder, DNN = deep NN, DT = decision tree, DTCWT = Double-Tree Complex Wavelet Transform (WT), DTF = direct transfer function, DWT = Discrete WT, ELM = extreme learning, FBCSP = filter bank CSP, FFT = Fast Fourier Transform (FT), HMM = hidden Markov model, kNN = k-nearest neighbour, LDA = linear discriminant analysis, LSTM = long short-term memory, MaxLCor = Maximum Linear Cross-correlation Coefficient, MFCC = Mel Frequency Cepstral Coefficients, NB = naïve Bayes, PDC = partial directed coherence, RF = random forest, RMS = root mean square, RNN = recurrent NN, RVM = relevance vector machine (VM), RWE = relative wavelet energy, SOM = self-organizing map (clustering), STFT = Short Time FT, SVM = support VM, TL = transfer learning.

This paper did not compare the accuracies quantitatively between studies due to the different techniques, datasets, or computation environments. Hence, the comparison would not be apple-to-apple. Discussing techniques and challenges is more worthwhile than the accuracy comparison to find the right solutions. The remainder of this review is organized as follows. Section II discusses the data acquisition process. Section III describes the signal processing used in the reviewed studies. Section IV deals with Spectral, Spatial and Temporal analysis. Section V explains the classification and feature extraction techniques that were categorized further into non-deep learning and deep learning, including the modified validation method. This paper describes the challenges of the application of EEG-based SSI in Section VI. Finally, Section VII draws some conclusions.

## II. Data Acquisition

The research subjects were given the cue of speech (vowels or words) shown from the monitor, heard from the earphones, or both (audiovisual cue). If the cue presentation was before the speech imagery, they had to memorize the cue, and this would separate the imagined speech task from the reading/listening task. While in the simultaneous cues, subjects performed the imagined speech task with the reading/listening task at the same time. Besides, the most active brain parts for listening and reading are different, i.e., the listening process involves the temporal lobes, and the reading process involves the occipital lobes. Thus, cue format and presentation regarding the time of imagined speech can affect different active brain areas.

The acquisition protocol of the Arizona State University dataset [26] used only visual cues, as illustrated in Fig. 2. The cue presentation was simultaneous with the imagined speech recording. The subject performed speech imagery at each "beep" sound and continued the same pattern until the visual cue disappeared (7 x *T* second). They used three short words (“in”, “out”, and “up”) and two long words (“cooperate” and “independent”). For the longer words, *T* is more than one second. This protocol was applied to 15 subjects (11 males, one left-handed, age 22-32) and the EEG signals recording used 60-channel EEG at 1000 Hz and two-channel electrooculography (EOG) to capture ocular artefacts. They applied Common Spatial Pattern (CSP) to get the most active brain areas. The results showed that brain activity almost entirely focused on the left frontal, middle and parietal sides of the brain, as the location of the motor cortex and Broca and Wernicke’s area. Even in the rest state, the brain remains highly active.

The other previous study that used visual cues presented the cues separately from the imagined speech state [23], as shown in Fig. 3, named the KARAONE dataset. The goal was to differentiate between the pronounced speech performed in the cue state and the imagined speech after the cue state. They used seven phonemes; “/iy/”, “/uw/”, “/piy/”, “/tiy/”, “/diy/”, “/m/”, “/n/”, and four words from Kent’s list [46], i.e. "pat", "pot", "knew", and "gnaw". This protocol was conducted on 12 subjects (8 males, all right-handed, age 27.4±5) and recorded with 62-channel EEG at 1024 Hz. They found that central brain areas in temporal locations had discriminative features.

A different protocol was employed by Coretto et al. [34], as illustrated in Fig. 4. They used both audio and visual cues before subjects performed imagined speech. No specific reason why they decided to use both cue types. They employed five vowels; “/a/”, “/e/”, “/i/”, “/o/”, “/u/”, and five Spanish words; “arriba” (“up”), “abajo” (“down”), “izquierda” (“left”), “derecha” (“right”), “atrás” (“backward”), and “adelante” (“forward”). This protocol was conducted on 15 subjects (8 males, one left-handed, averaged age 25) and recorded with six-channel EEG at 1024 Hz. Moreover, there was still no further study with spatial analysis to observe which brain parts are more involved in both cues.

On the other side, a Brain-Computer Interface (BCI) open dataset [47] for the speech imagery study only employed audio cues. They used five common words; “hello”, “help me”, “stop”, “thank you”, and “yes” and conducted this protocol on 15 subjects using 64-channel EEG at 256 Hz. The cue presentations were given independently from the imagined speech state, as shown in Fig. 5.

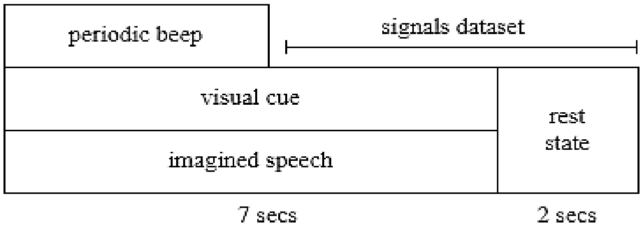

Fig. 2. Arizona State University Dataset’s Acquisition Protocol.

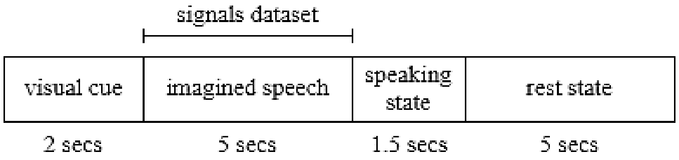

Fig. 3. KARAONE Dataset’s Acquisition Protocol.

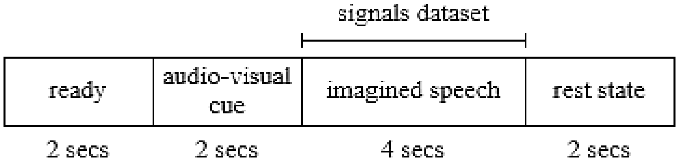

Fig. 4. Coretto Dataset’s Acquisition Protocol.

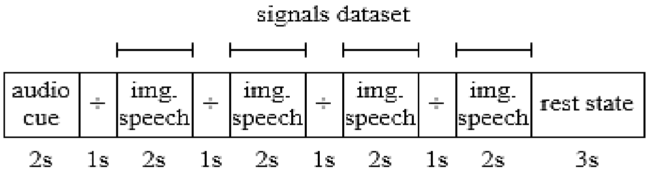

Fig. 5. BCI 2020 Track 3 Dataset’s Acquisition Protocol.

Besides the cue format and its presentation timing, the duration of each state is an important issue. In the KARAONE dataset, the length of the imagined speech state was five seconds with no repetition. Meanwhile, Arizona State University’s dataset [26] and Coretto’s [34] applied repetition during one imagined speech state using the beep sound. The Arizona State University dataset had seven repetitions in one imagined speech state. Coretto et al. applied repetition only for words, but not vowels, i.e., the subject must complete the task within four seconds of the vowel imagined speech state. BCI dataset [47] also used repetitions, but rather than using the beep sound, it applied a fixation cross shown from the monitor. The purpose of repetition block use was to maintain focus consistency [8]. Moreover, the longer the duration of the imagined speech, the easier for the subject to get sleepy because they performed the imagined speech state in silence.

While performing imagined speech, subjects were asked to refrain from moving articulation, swallowing, moving their eyes, or blinking. These restrictions aim to reduce the muscular and ocular artefacts as the significant artefacts in the EEG signal. These artefacts are generated by the signal from muscle activity surrounding the head, particularly the region near the articulator and eye activity. The range of muscular and ocular signals has intersected with the range of EEG signals [48].

The previous studies used different types of speech, such as vowels [26], phonemes [23], syllables [10], words [6]–[8], [16]–[18], [23], [26], [31], [32], sentences [7], and binary questions [20], [22]. The use of speech parts (vowels, phonemes, and syllables) mainly was to observe the brain when planning to produce the sound of words. Meanwhile, using words or binary questions to observe the brain when planning to respond/send a message earlier than sound production. Thus, the former is more syntactic, whereas the latter is more semantic. The distribution of the speech used in this related research is shown in Fig. 6.

Using syllables or vowels as the cue was for their discriminative sounds or articulator movements. To maintain the number of nasals, plosives, and vowels, Zhao and Rudzicz [23] used phonemes and short words with similar sounds. DaSalla et al. [12] only used the vowels “/a/” and “/u/” because of their articulation differences; “/a/” with an opened mouth regulated by digastricus muscles and “/u/” with rounded lips controlled by orbicularis oris muscles. On the other side, Nguyen et al. [26] utilized vowels to discover the relationships between vowel voices and model accuracies. They reported that different voices had an impact on the performances. Then, the use of the syllables “/ba/” and “/ku/” [9], [10] was mainly to classify the rhythm.

Most SSI studies employed the word cues for several reasons. It represents natural communication, such as asking, responding, or making a statement [18], [32] and semantically differentiates homonyms [6], [16], [17], e.g., "two" and "to". Additionally, employing words with different lengths could help the model to differentiate words (such as “in” and “cooperate”) [26] or repetition [8] affected the model and use words as commands for moving the cursor on the monitor (“up”/”down”) [31].

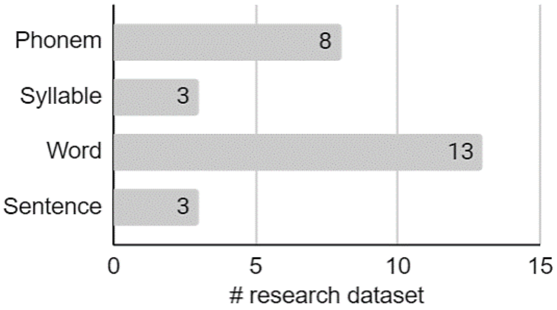

Fig. 6. Used Speech Cues.

Wester [7] also applied several words grouped based on their usability; digits (0-9), alpha corpus, phone-related words, MP3 player commands, Graduate Record Examination (GRE) Corpus, and Lecture Corpus. The alpha corpus consists of “alpha”, “beta”, “theta”, “delta”, and “echo”. The phone-related words covered some phone communication words, such as "yes", "no", "accept", "deny", and "wait". The phrases for MP3 player commands were "start", "back", "next", "louder", and "turn down". GRE Corpus represent the rarely used words like "brittle" or "profundity". Lecture Corpus to examine the sequence of words as a sentence such as "good afternoon, ladies and gentlemen, welcome to the interact centre. My name is …, thank you for your attention, any questions."

The cognitive aspect is vital in the question cue because subjects must first comprehend the question. Then, different questions with the same answer could raise other cognition. For example, the answer to "Are you a scholar?" and "Are you a human?" are assumed to be "yes", but the cognition is different. Furthermore, the brain activity to respond with a "yes" is different from saying a "yes" with no intention. It will need the adaptability of the model to overcome the difference.

The multilanguage issue represents the other chance for SSI study. Many languages have been applied, e.g. English [6], [7], [16], [23], [26], Chinese [18], Hindi [22], Arabic [20], Korean [17], and Spanish [31]. Suppose the model can classify words from one language and their equivalent in another. In that case, this sparks the chance to build a complete set of EEG-based imagined speech datasets, regardless of the source languages of speech.

## III. Signal Preprocessing

Even though some restrictions, e.g. to move, were applied in the recording protocol, artefacts and noises are still unavoidable in EEG signals acquisition because of the low frequency and voltage, which are easily interfered with muscular/ocular artefacts and other noises. It still becomes a challenging issue for EEG studies in data cleaning without losing significant information/features for later analysis or pattern recognition. The noise removal must be performed before the downsampling step to prevent the downsampled values from being falsely interpreted as noise. Previous studies used Independent Component Analysis (ICA) [49], [50] or artefact detectors based on the joint use of spatial-temporal features (ADJUST) [51].

If EEG acquisition used a high sampling rate, e.g., 1000Hz, most studies applied the downsampling process to lower the computational complexity. According to the Nyquist theorem and the brain signal frequency range of 0.5-100 Hz [48], 256 Hz covers more than twice the commonly observed maximum frequency. From the 37 reviewed studies, only seven studies used it ([11], [12], [17], [21], [26], [34], [42]). There will be thread off for the downsampling; the smaller the sample size, the lower the needed computation resources, but the more important features lost. Thus, using the original sample size could help observe the discriminative speech recognition features while also considering the available resources.

## IV. Spectral, Spatial, and Temporal Analysis

Choosing only certain frequency bands, such as the alpha and beta bands, could decrease the number of features [20] because alpha and beta bands contain discriminative information. Statistic calculation (e.g. maximum value, average, standard deviation, kurtosis, and others) in beta, delta, and theta gave higher accuracy than the other bands in imagined vowel classification [13] with 81.25-98.75% accuracy for classifying the combined task, e.g., features from imagined speech state of “/a/” was combined with rest state, “/a/” and “/u/”, and so on. However, the higher gamma band is not discriminative for speech imagery [11], [18], [31], [39] since muscular artefacts produce high gamma activity [52]. Additionally, from the reviewed papers, only a few studies (8 of 37) used it, as shown inFig. 7. Two of them reported that the gamma did not provide discriminative characteristics to decode speech [39] except for a speech with articulation [21].

When observing which frequency bands give the highest accuracy, there are some considerations about physiological activities associated with the specific bands. The high-frequency bands, e.g., beta or gamma, are dominated by muscular artefacts. Even though these waves correspond to the concentration or active attention [53], thus, they could have information for imagined speech recognition. Meanwhile, the low-frequency bands do not correlate with concentration [54]. The alpha waves correspond to relaxed awareness, the theta waves appear when the consciousness moves to drowsiness, and the delta waves are further away from concentration since they relate to deep sleep. Furthermore, low-frequency bands often get interferences with ocular artefacts or lead movements [55]. Since cognitive task, e.g. imagined speech, requires concentration, the role of low-frequency bands that yield higher accuracy than the high ones needs more examination.

EEG spatial analysis can give a better insight into which brain area has essential information for imagined speech recognition; the electrodes can be selected further. When a conversation happens, the most active regions are the auditory cortex, motor cortex, Broca’s area, and Wernicke area [56]. Broca’s area is in the inferior frontal gyrus, while Wernicke’s area is in the superior temporal gyrus, with the arcuate fasciculus connecting both of them to build auditory-motor interaction [57], as illustrated inFig. 8. Some SSI research has validated this brain region [7], [18], [22], [26]. Furthermore, in early speech production, the auditory potential existed in the superior temporal gyrus (STG) [58], located in the temporal area in both hemispheres.

Most of the reviewed studies reported high accuracy achievement by using the features extracted from frontal lobes (location of the Broca’s area) [9], [10], [17], [26], [42] and temporal lobes (location of the Wernicke’s area) [17], [21], [39], [42], followed by parietal lobes [9], [17], [26]. The others found that occipital lobes contributed to achieving the highest accuracy, which is obvious since they employed visual cues in their protocols [10], [39].

Suppose the EEG signals were treated as an event-related potential (ERP) during imagined speech production. The signals can be aligned to the onset of imagined speech production and then averaged to focus between the preceding and following onset time [59]. Previous studies have examined it and found that the potential in the left hemisphere arises one-two seconds before speech production [36] and appears a moment before cued speech is produced [60]. One study reported that speech-related ERP reached the peak at 350ms after the cue [12], and the highest significant level (from the paired t-test) existed between 400-600ms after the cue in the frontal area [17]. Moreover, from the five alphabets (“/a/”, “/e/”, “/i/”, “/o/”, and “/t/”) classification result, data in 100-600ms after the cue gave the highest accuracy of 46.61% (chance level of 20%) [39].

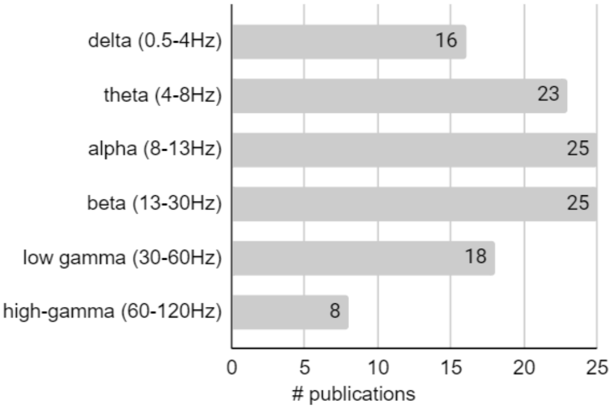

Fig. 7. Observed Frequency Bands in EEG-Based SSI.

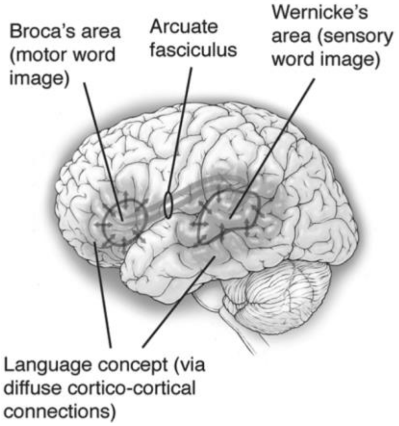
Broca’s area (motor word image)
Arcuate fasciculus
Wernicke’s area (sensory word image)
Language concept (via diffuse cortico-cortical connections)

Fig. 8. Language Organization in the Left Brain Hemisphere [57].

## V. Classification and Feature Extraction Techniques

### A. Non-Deep Learning

Most features were the result of the transformation to the frequency domain, as shown inFig. 9, e.g., Fourier, Wavelet, or Hilbert-Huang Transform. Common Spatial Patterns (CSP) and Principal Component Analysis (PCA) are other extraction methods. Filter bank CSP (FBCSP) [61] was the state-of-the-art for feature selection [35] or classification [36], [37], [62].

Since EEG signals are also time-series data, some studies applied auto-regression (AR) [11], [18], [19] or Mel Frequency Cepstral Coefficient (MFCC) [24], [41]. MFCC was more discriminative than the AR coefficient for vowel recognition in DaSalla's dataset [12] by yielding an accuracy of 75% [14] and performed better than statistics and nonlinear features in the KARAONE dataset [23]. MFCC gained an accuracy of 19.69% (chance level of 9.09% for 11 classes), while the accuracy of statistics features was 15.91%, and the accuracy of nonlinear features was 14.67% [24]. Several studies also treated EEG signals as a sequence of words by applying the Bag of Features (BoF) [30], [33]. In text pattern recognition, BoF was often used to represent a word existence using the feature values calculated from its previous words.

The other feature types are connectivity features, which relate to the brain's neural pathways when subjects perform a specific task, such as imagined speech production. They are structural [63], functional [64], and effective connectivity [65]. Structural connectivity refers to the tracts of white matter that physically interconnect brain regions. Functional connectivity refers to the statistical dependence (i.e. correlation) of time-series data between a pair of brain regions influenced by structural connectivity. Meanwhile, effective connectivity refers to a causal model representing the interactions between connected neurons.

Few studies of imagined speech recognition have considered applying functional or effective connectivity features. Qureshi et al. [43] used functional connectivity features fed into an extreme learning machine (ELM) to classify imagined speech of five words. These features were covariance and maximum linear cross-correlation coefficient (MaxLCor), with the same calculation for phase-only time-series data. MaxLCor is one of the spatial connectivity features to measure functional connectivity and extract EEG characteristics by calculating the normalized product of two time-series signals and then measuring their similarities [66]. They reported that covariance features yielded the highest accuracy of 87.90% on binary classification. Pawar et al. [44] also used MaxLCor combined with DWT features to classify imagined speech words [47] and achieved an accuracy of 40.64±2.45% (chance level 20%). Furthermore, Chengaiyan et al. [45] identified vowels and consonants by applying brain connectivity features on each frequency band; coherence [67] as functional connectivity and partial directed coherence (PDC) [68], direct transfer function (DTF) [69], and transfer entropy [70] as effective connectivity. They fed the features into deep learning methods, a Recurrent Neural Network (RNN) and a Deep Belief Network (DBN), where RNN gave lower accuracy of 72% than DBN with an accuracy of 80%.

Many different features were classified with Support Vector Machine (SVM) ([17], [18], [20], [39]) or Linear Discriminant Analysis (LDA) ([7], [13], [20]) as shown in Fig. 1, since both are good at separating discriminative values. Meanwhile, the Decision Tree (DT) [16] and its ensemble variant called Random Forest (RF) [31], [32], [34] were used due to their capability to distinguish between the classes and process a large number of features. The RF feed with Wavelet features outperformed SVM in the same dataset [31], [32].

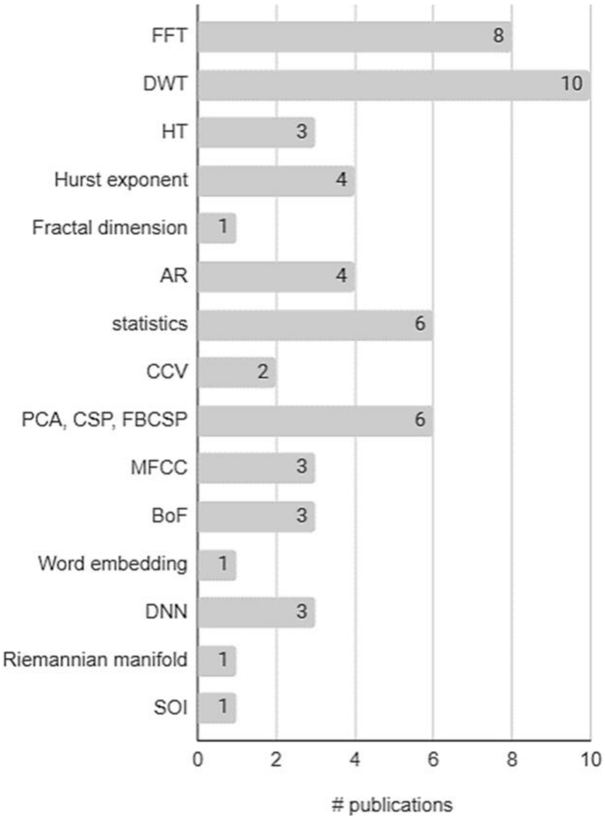

Fig. 9. Features in EEG-Based SSI.

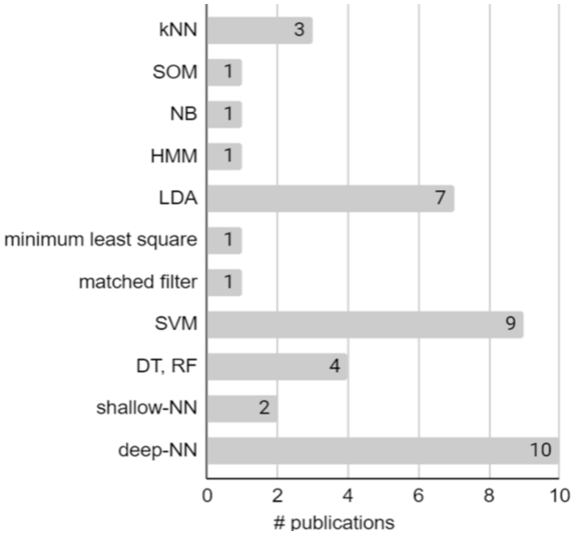

Fig. 10. Classification Methods in EEG-Based SSI.

Another classification technique formerly used in imagined speech classification was k-Nearest Neighbor (kNN). When combined with MFCC features, it gained higher accuracy than SVM and the Hidden Markov Model (HMM) to identify the vowel of "/a/" and "/u/" from the DaSalla's dataset [12] with kNN's averaged accuracy of 86.89% compared with 75.83% and 70.56% respectively [14]. It also still outperformed SVM in classifying "/a/", "/e/", "/i/", "/o/", and "/u/" in the additional datasets of the same study [14]. Since kNN depends on the centroids for the *k* classes derived during training and the test dataset was classified based on the majority class of its neighbours, new samples may require k-NN retraining.

The shallow Artificial Neural Network (ANN), in the form of ELM, was also applied due to its architecture of layers and nodes that can achieve good generalization. With only using statistical values of the signals to classify vowels, ELM's accuracy was higher than SVM or LDA, in which ELM's accuracy was 87.07% compared with 51.07% for SVM and 81.98% for LDA [21]. Furthermore, with average power features to classify "yes" and "no", ANN also performed better than SVM and RF, with 92.18% compared with 83.07% and 79.95%, respectively [22].

The ANN's capacity for generalization motivated further research using deep learning (DL). While non-DL techniques depend on the features input, DL uses its layers to learn the data characteristics directly.

### *B. Deep Learning*

Several attempts employed deep learning (DL) to extract features, e.g., Convolutional Neural Network (CNN) to extract spatial features and Recurrent Neural Network (RNN) for temporal characteristics [27]. These feature vectors were concatenated in the form of a channel covariance matrix as the input for Deep Autoencoder (DAE) classifier. Siamese Network increased the distance of different labelled samples and vice versa, which gave higher accuracy (31.40 ± 2.73%) [38] than the baselines [33], [34], [37], [62] by using the same Coretto's dataset [34].

Although DL has become the common feature and representation learning method, it is also well-known for being data-hungry. A small-size dataset might generate a final model that overfits. Therefore, only a few previous studies have used DL compared with non-DL, e.g., SVM or LDA. Furthermore, it is impossible to cover the whole language corpus, as subjects would feel uncomfortable for more than 30 minutes of EEG recording.

### *C. Model Evaluation*

Although many reported models and their features claimed their highest accuracy, different datasets and experiment environments caused the accuracies to be incomparable since those factors can affect the chosen discriminative features and the model training process. Furthermore, the cognitive variance across subjects [71] needs more consideration since it caused the recognition model's accuracy to be more accurate when it was trained and tested in one specific subject's data but lower accuracy when it was tested to recognize the other's. This problem is called the inter-subject case. Hence, the model's accuracy in most brain signal studies was evaluated by each person, i.e., the accuracy was calculated for each subject before the whole-averaged accuracy was reported. From the reviewed collected references, the validation was categorized into three versions.

The first type of validation is the usual cross-validation (CV) in machine learning treated subject-wise. The subjects' datasets were gathered into one massive dataset. Then it was split into training, validation, and testing parts with the configurable percentages for each part, e.g., training and validation took 60% and 20%, respectively, for k-fold cross-validation. The remaining 20% for each subject was kept unlabeled for the model to later predict in the testing session. In this type of CV, the training of the model uses the features from all subjects that could make the model achieve very high accuracy because it also learned a part of data whose the same variance as the testing dataset. Still, it became weak when facing the subjects' cognitive variance as an inter-subject problem exists. This case is essential to be solved but still unnoticed by many previous studies.

The second type is leave-one-subject-out cross-validation (LOSO-CV). It is similar to leave-one-out cross-validation (LOOCV) by using one subject's dataset as a validation dataset. This method aimed to measure the robustness of the trained model related to the inter-subject issue. It can prevent the model from peeking at the test dataset variance and overfitting. This method aimed to measure the robustness of the trained model related to the inter-subject issue. It also can be extended to be the leave-N-subject-out CV. Only a few gathered studies, listed in Table I, employed LOSO-CV.

The third type is leave-one-session-out cross-validation (LOSeO-CV) to face the intra-subject problem, i.e., the model's accuracy degrades when recognizing a new recording dataset of the same subjects whose datasets were used for training the model before. This validation type is the modification of LOSO-CV with a different perspective. Even though some studies were aware of the intra-subject problem [7], [15], they did not apply LOSeO-CV since their goal was to get higher accuracy than the baseline with the current data distribution only.

It is essential to note that the comparison of the higher accuracy achieved from the general CV with the lower accuracy gained from LOSO-CV was irrelevant. It is because LOSO-CV aimed to prepare the model to become adaptive to different subjects' data distribution due to the cognitive variance. Besides, the general CV only considers the current data distribution and potentially peeks distribution information from the same subject in the testing dataset. Thus, the model tends to be overfitting.

On the recognition of Arizona State University's dataset [26] in Table II, the researchers could train their model using the general CV on the deep learning model, and they achieved higher accuracy than the baseline (49%), with the highest accuracy being 96,79 ± 4.19% [28] for vowel recognition only. Another study also used deep learning with the general CV that boosted the accuracy of long-word recognition to 81.65%, higher than the baseline of 66.20%. Although, for short-long word discrimination, a deep learning implementation [29] still did not achieve higher accuracy than the baseline (77.60% of

80.10%). Meanwhile, another research [30] used LOSO-CV while recognizing only the long words. It reported a lower accuracy of 62,99 ± 4.78% than the reported accuracy in Table II. There was still no further observation for short-word classification with higher accuracy than the baseline. Moreover, deep learning also successfully yielded higher accuracy on the KARAONE dataset [23] for multi-class classification (i.e. not a binary classification), as shown in Table III, with an accuracy of 57.15% [25] higher than the baseline (33.3%) [24]. These KARAONE dataset classification studies used the general CV for validation.

Before the development of Coretto's dataset, Torres-Garcia et al. constructed an EEG-based imagined speech dataset [31] with five similar words to Coretto's dataset; Torres-Garcia used "seleccionar" ("select") rather than "adelante" ("forward") and "atrás" ("backward"). The non-deep-learning model achieved the highest accuracy (70.33%) for this dataset, with a general CV for its validation, as shown in Table IV. One deep learning implementation [30] with a transfer learning approach still gained slightly lower accuracy (65.65%) validated by LOSO-CV.

Similar to Arizona State University's dataset, the studies on Coretto's dataset also achieved the highest accuracy by implementing a deep learning model (30% for vowel classification [35] and 62.37% for word classification [36]). It was validated by a general CV, as shown in Table V. Further research validated the deep learning model with LOSO-CV to classify the vowels. It successfully yielded higher accuracy (32.75%) [37] than the baseline (30%) [35] in recognizing vowels. Thus, the model became quite robust since it could still accurately recognize the unseen subject's dataset.

The other reviewed studies used a general CV for validation; DaSalla's EEG-based imagined speech dataset consists of "a" and "u" speech [12], and Dzmura's consists of "ba" and "ku" [9]. There was still no deep learning exploration for DaSalla's dataset. Although, the researchers can achieve higher accuracy than the baseline for binary classification, as shown in Table VI, by using different feature extraction algorithms. Meanwhile, the baseline study of Dzmura's dataset still had the highest accuracy (74.25%) with the spectral feature and matched-filter classification.

From observing several EEG-based imagined speech datasets, some deep-learning studies yielded higher accuracy than the baseline studies validated by general CV, e.g. in Arizona State University's, Torres-Garcia's, Coretto's, and KARAONE datasets. Although, the deep learning models gained lower accuracy with LOSO-CV, e.g. in Arizona State University's dataset, Torres-Garcia's dataset, and Coretto's dataset. Some studies applied the transfer learning approach to build a more robust model with LOSO-CV validation; one successfully gained higher accuracy [37], but the other still got slightly lower accuracy [30]. Nevertheless, the transfer learning approach could have the capability to train the robust model. Additionally, the non-deep-learning models could gain higher accuracy when the informative features fed into them, as in DaSalla's and Dzmura's datasets. The accuracy of the trained models from the same datasets validated with the general CV in Table II-VII can become the benchmark for further studies.

TABLE II. ACCURACY WITH GENERAL CROSS-VALIDATION FOR ARIZONA STATE UNIVERSITY'S DATASET

| Ref. | Accuracy (%) | | | |
|---|---|---|---|---|
| | Vowel | Short word | Long word | Short vs long word |
| [26] | 49.00 | **50.10** | 66.20 | **80.10** |
| [27] | - | - | **81.65** | - |
| [28] | **96.79** | - | - | - |
| [29] | - | - | - | 77.60 |

TABLE III. ACCURACY WITH GENERAL CROSS-VALIDATION FOR KARAONE DATASET

| Ref. | Binary-class Accuracy (%) | | | | | Multi-class Accuracy (%) |
|---|---|---|---|---|---|---|
| | Bila-bial | Nasal | C/V | /uw/ | /iy/ | |
| [23] | 56.64 | 63.50 | 18.08 | 79.16 | 59.6 | - |
| [24] | - | - | - | - | - | 33.3% |
| [25] | - | - | - | - | - | **57.15%** |

TABLE IV. ACCURACY WITH GENERAL CROSS-VALIDATION FOR TORRES-GARCIA'S DATASET

| Ref. | Accuracy (%) |
|---|---|
| [31] | 41.21 |
| [32] | **70.33** |
| [33] | 61.02 |

TABLE V. ACCURACY WITH GENERAL CROSS-VALIDATION FOR CORETTO'S DATASET

| Ref. | Accuracy (%) | |
|---|---|---|
| | Vowel | Word |
| [34] | 22.72 | 19.60 |
| [35] | **30.00** | 24.97 |
| [36] | - | **62.37** |
| [38] | - | 31.40 |

TABLE VI. ACCURACY WITH GENERAL CROSS-VALIDATION FOR DASALLA'S DATASET

| Ref. | Binary-class Accuracy (%) | | |
|---|---|---|---|
| | /a/-rest | /u/-rest | /a/-/u/ |
| [12] | 72.33 | 78 | 62.67 |
| [14] | 75.00 | **93.83** | **91.83** |
| [13] | **75.83** | 77.5 | 72.5 |

TABLE VII. ACCURACY WITH GENERAL CROSS-VALIDATION FOR D'ZMURA'S DATASET

| Ref. | Accuracy (%) |
|---|---|
| [9] | **74.25** |
| [10] | 58.05 |
| [11] | 68.83 |

## VI. CURRENT CHALLENGES

### A. Laboratory Environment

In the imagined speech decoding research, EEG signals were recorded in a conducive laboratory environment with a proper procedure to minimize the noise and artefacts. When the interface is intended to be practical, e.g., for patients at the hospital or as an in-house assistive tool, it will face a noisier environment and inevitable artefacts.

The previously discussed artefact removal approaches still require validation since they only focused on increasing the accuracy without reexamining the effectiveness of artefact removal. The acquisition should have relaxed restrictions, such as allowing subjects to swallow or blink during imagined speech production, to validate the result of artefact removal.

### *B. Related Channels*

Much spatial information could be observed using many electrodes, but it will be less convenient, less impractical, and have bigger feature dimensionality. This issue also happens in motor imagery BCI. Several approaches to overcome it are channel selection, spatial filter, and feature selection [32]. Feature selection aimed to select the most discriminative features, spatial filter to extract characteristics from channels employed, and channel selection to choose several channels with similar/better accuracy as the whole channels.

### *C. Time-Lock*

In EEG signals, there is temporal information related to the onset time and spatial features related to the brain area. Spatial and temporal information of overt speech and imagined speech correlated [42]; the spatial pattern is not significantly different, but the temporal one is. It is due to the difficulty in determining the time-lock of the imagined speech, compared with the overt speech whose time-lock is easily detected from the voice. Besides, the time-lock can be different in the different sessions.

### *D. Intra-Subject and Inter-Subject Problem*

One main problem of EEG-based imagined speech studies is the limited speech data. Should the recording cover the whole vocabulary in a language, the subjects will be exhausted, and it will need a very long time. Besides, there are different patterns produced even if a person imagines the same speech, i.e., the inter-subject issue [15], [41], or if he imagines the same speech at a different time, i.e., the intra-subject problem [42]. Thus, the model must be adaptive to recognize new data from new sessions/subjects without training from scratch.

One approach to solving the adaptation issue is transfer learning (TL). Traditional machine learning assumes that the distribution of the present learned data and the future data are the same. In contrast, TL assumes their domain is different, or the future labelling task may differ. Some studies reported that TL did not decrease the accuracy [36], [37]. However, as the accuracy was still poor (35-60%), it needs further observation.

### *E. Connectivity*

The brain works as a neural pathway, and its existing connectivity can contain informative features for cognitive tasks, including speech imagery. Functional connectivity (correlation between brain areas) and effective connectivity (the causal model of brain areas' interaction) become potential characteristics to help identify imagined speech.

### *F. Subject Limitation*

Current SSI studies were limited to healthy subjects. The subjects' brains must be in good condition to record signals. When the study includes brain-impaired participants, the problem-related brain area might affect the data acquired and its recognition accuracy, which needs more profound observation. Additionally, observing subjects with health issues, such as the locked-in syndrome (LIS) patient, was also challenging. The moment the LIS subject began producing the speech could not be identified precisely, although the subject has been instructed to confirm the speech production attempt [72]. Other health disorders, e.g., Aphasia, Apraxia, Dysarthria, laryngectomy (i.e. removal of the larynx by operation procedure), and tracheostomy, also have specific conditions. For example, the brain activity in speech production for a new laryngectomy patient may differ from a one-year patient.

### *G. Online Learning*

There would be a requirement for online learning where the model training is simultaneous with EEG recording. It could exploit the users' feedback to retrain the model and recognize the pattern more accurately. Moreover, the subjects could also be trained to modify their brain waves to adapt to BCI [73]. Currently, most studies used the offline training (, i.e., outside the recording session). Although few studies performed online learning [7], [20], the performance was low and inconsistent.

## VII. Conclusions

This review discussed the pipeline of EEG-based SSI, which consists of signal acquisition, signal preprocessing, feature extraction, and classification, to see problems that often arise in each step. The acquisition process needs a proper design of subject inclusion, cue format, and speech types according to the purpose of the study, including the challenges that need answers to apply the decoding in the real world while maintaining the high accuracy achieved in a lab environment. These challenges deal with handling noises and artefacts, the trade-off between the number of channels and spatial features and onset time determination to gain discriminative temporal characteristics. Besides, variance shifts due to different recording sessions or users that demand an adaptive model and its validation need consideration. The inclusion of brain-impaired subjects and the potential of online learning could make the interface more applicable. To conclude, this review suggests that it is crucial to start by building the proper pipeline and taking problems in each step into consideration to overcome the challenges. High accuracy is insufficient to make the model applicable in the real world.

## References

[1] WHO, "World Report On Disability," 2011. [Online]. Available:https://www.who.int/disabilities/world_report/2011/report.pdf

[2] ILO and LPEM FEB UI, "Memetakan Penyandang Disabilitas (PD) di Pasar Tenaga Kerja Indonesia," 2017. [Online]. Available: https://www.ilo.org/jakarta/whatwedo/publications/WCMS_587668/lang--en/index.htm

[3] N. Bhattacharyya, "The prevalence of voice problems among adults in the United States," Laryngoscope, vol. 124, no. 10, pp. 2359–2362, Oct. 2014, doi: 10.1002/lary.24740.

[4] F. Bocquelet, T. Hueber, L. Girin, S. Chabardès, and B. Yvert, "Key considerations in designing a speech brain-computer interface," J. Physiol., vol. 110, no. 4, pp. 392–401, Nov. 2016, doi: 10.1016/j.jphysparis.2017.07.002.

[5] J. A. Gonzalez-Lopez, A. Gomez-Alanis, J. M. Martin Donas, J. L. Perez-Cordoba, and A. M. Gomez, "Silent Speech Interfaces for Speech Restoration: A Review," IEEE Access, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3026579.

[6] P. Suppes, Z.-L. Lu, and B. Han, "Brain wave recognition of words," Proc. Natl. Acad. Sci., vol. 94, no. 26, pp. 14965–14969, Dec. 1997, doi: 10.1073/pnas.94.26.14965.

[7] M. Wester, "Unspoken Speech: Speech Recognition Based On Electroencephalography," Universität Karlsruhe, 2006. [Online]. Available: https://www.researchgate.net/publication/36453500_Unspoken_Speech_-_Speech_Recognition_based_on_Electroencephalography

[8] A. Porbadnigk, M. Wester, J. P. Calliess, and T. Schultz, "EEG-Based Speech Recognition - Impact of Temporal Effects," in Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, 2009, no. January, pp. 376–381. doi: 10.5220/0001554303760381.

[9] M. D'Zmura, S. Deng, T. Lappas, S. Thorpe, and R. Srinivasan, "Toward EEG Sensing of Imagined Speech," in Proceedings of the 13th International Conference on Human-Computer Interaction, 2009, pp. 40–48. doi: 10.1007/978-3-642-02574-7_5.

[10] S. Deng, R. Srinivasan, T. Lappas, and M. D'Zmura, "EEG classification of imagined syllable rhythm using Hilbert spectrum methods," J. Neural Eng., vol. 7, no. 4, p. 046006, Aug. 2010, doi: 10.1088/1741-2560/7/4/046006.

[11] K. Brigham and B. V. K. V. Kumar, "Imagined Speech Classification with EEG Signals for Silent Communication: A Preliminary Investigation into Synthetic Telepathy," in 2010 4th International Conference on Bioinformatics and Biomedical Engineering, Jun. 2010, pp. 1–4. doi: 10.1109/ICBBE.2010.5515807.

[12] C. S. DaSalla, H. Kambara, M. Sato, and Y. Koike, "Single-trial classification of vowel speech imagery using common spatial patterns," Neural Networks, vol. 22, no. 9, pp. 1334–1339, Nov. 2009, doi: 10.1016/j.neunet.2009.05.008.

[13] B. M. Idrees and O. Farooq, "Vowel classification using wavelet decomposition during speech imagery," in 2016 3rd International Conference on Signal Processing and Integrated Networks (SPIN), Feb. 2016, pp. 636–640. doi: 10.1109/SPIN.2016.7566774.

[14] A. Riaz, S. Akhtar, S. Iftikhar, A. A. Khan, and A. Salman, "Inter comparison of classification techniques for vowel speech imagery using EEG sensors," in The 2014 2nd International Conference on Systems and Informatics (ICSAI 2014), Nov. 2014, no. Icsai, pp. 712–717. doi: 10.1109/ICSAI.2014.7009378.

[15] X. Chi, J. B. Hagedorn, D. Schoonover, and M. D. Zmura, "EEG-Based Discrimination of Imagined Speech Phonemes," Int. J. Bioelectromagn., vol. 13, no. 4, pp. 201–206, 2011, [Online]. Available: https://pdfs.semanticscholar.org/b74f/c325556d1a7b5eb05fe90cde1f0c891357a3.pdf

[16] C. M. Spooner, E. Viirre, and B. Chase, "From Explicit to Implicit Speech Recognition," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 8027 LNAI, 2013, pp. 502–511. doi: 10.1007/978-3-642-39454-6_54.

[17] T. Kim, J. Lee, H. Choi, H. Lee, I.-Y. Kim, and D. P. Jang, "Meaning based covert speech classification for brain-computer interface based on electroencephalography," in 2013 6th International IEEE/EMBS Conference on Neural Engineering (NER), Nov. 2013, pp. 53–56. doi: 10.1109/NER.2013.6695869.

[18] L. Wang, X. Zhang, X. Zhong, and Y. Zhang, "Analysis and classification of speech imagery EEG for BCI," Biomed. Signal Process. Control, vol. 8, no. 6, pp. 901–908, Nov. 2013, doi: 10.1016/j.bspc.2013.07.011.

[19] Y. Song and F. Sepulveda, "Classifying speech related vs. idle state towards onset detection in brain-computer interfaces overt, inhibited overt, and covert speech sound production vs. idle state," in 2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings, Oct. 2014, pp. 568–571. doi: 10.1109/BioCAS.2014.6981789.

[20] M. Salama, L. Elsherif, H. Lashin, and T. Gamal, "Recognition of Unspoken Words Using Electrode Electroencephalograhic Signals," in COGNITIVE 2014 : The Sixth International Conference on Advanced Cognitive Technologies and Applications, 2014, pp. 51–55. [Online]. Available: https://bu.edu.eg/portal/uploads/Engineering, Shoubra/Electrical Engineering/3513/publications/May ahmed salama mohamed_Recog of unspoken words.pdf

[21] B. Min, J. Kim, H. Park, and B. Lee, "Vowel Imagery Decoding toward Silent Speech BCI Using Extreme Learning Machine with Electroencephalogram," Biomed Res. Int., pp. 1–11, 2016, doi: 10.1155/2016/2618265.

[22] A. Balaji et al., "EEG-based classification of bilingual unspoken speech using ANN," in 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jul. 2017, pp. 1022–1025. doi: 10.1109/EMBC.2017.8037000.

[23] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2015, pp. 992–996. doi: 10.1109/ICASSP.2015.7178118.

[24] C. Cooney, R. Folli, and D. Coyle, "Mel Frequency Cepstral Coefficients Enhance Imagined Speech Decoding Accuracy from EEG," in 29th Irish Signals and Systems Conference (ISSC), Jun. 2018, pp. 1–7. doi: 10.1109/ISSC.2018.8585291.

[25] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "Decoding Imagined Speech using Wavelet Features and Deep Neural Networks," in 2019 IEEE 16th India Council International Conference (INDICON), Dec. 2019, pp. 1–4. doi: 10.1109/INDICON47234.2019.9028925.

[26] C. H. Nguyen, G. K. Karavas, and P. Artemiadis, "Inferring imagined speech using EEG signals: a new approach using Riemannian manifold features," J. Neural Eng., vol. 15, no. 1, p. 016002, Feb. 2018, doi: 10.1088/1741-2552/aa8235.

[27] P. Saha and S. Fels, "Hierarchical Deep Feature Learning for Decoding Imagined Speech from EEG," in Proceedings of the AAAI Conference on Artificial Intelligence, Jul. 2019, vol. 33, pp. 10019–10020. doi: 10.1609/aaai.v33i01.330110019.

[28] M. Parhi and A. H. Tewfik, "Classifying Imaginary Vowels from Frontal Lobe EEG via Deep Learning," in 2020 28th European Signal Processing Conference (EUSIPCO), Jan. 2021, pp. 1195–1199. doi: 10.23919/Eusipco47968.2020.9287599.

[29] J. T. Panachakel, A. G. Ramakrishnan, and T. V. Ananthapadmanabha, "A Novel Deep Learning Architecture for Decoding Imagined Speech from EEG," IEEE Austria Int. Biomed. Eng. Conf. (AIBEC 2019), Mar. 2020, [Online]. Available: http://arxiv.org/abs/2003.09374

[30] M. Jiménez-Guarneros and P. Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," Pattern Recognit. Lett., vol. 141, pp. 54–60, Jan. 2021, doi: 10.1016/j.patrec.2020.11.013.

[31] A. A. Torres-García, C. A. Reyes-García, and L. Villaseñor-Pineda, "Toward a Silent Speech Interface Based on Unspoken Speech," in Proceedings of the International Conference on Bio-inspired Systems and Signal Processing, 2012, pp. 370–373. doi: 10.5220/0003769603700373.

[32] A. A. Torres-García, C. A. Reyes-García, L. Villaseñor-Pineda, and G. García-Aguilar, "Implementing a fuzzy inference system in a multi-objective EEG channel selection model for imagined speech classification," Expert Syst. Appl., vol. 59, pp. 1–12, Oct. 2016, doi: 10.1016/j.eswa.2016.04.011.

[33] J. S. García-Salinas, L. Villaseñor-Pineda, C. A. Reyes-García, and A. A. Torres-García, "Transfer learning in imagined speech EEG-based BCIs," Biomed. Signal Process. Control, vol. 50, pp. 151–157, Apr. 2019, doi: 10.1016/j.bspc.2019.01.006.

[34] G. A. Pressel Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in 12th International Symposium on Medical Information Processing and Analysis, Jan. 2017, p. 1016002. doi: 10.1117/12.2255697.

[35] C. Cooney, A. Korik, R. Folli, and D. Coyle, "Evaluation of Hyperparameter Optimization in Machine and Deep Learning Methods for Decoding Imagined Speech EEG," Sensors, vol. 20, no. 16, p. 4629, Aug. 2020, doi: 10.3390/s20164629.

[36] C. Cooney, A. Korik, R. Folli, and D. H. Coyle, "Classification of Imagined Spoken Word-pairs using Convolutional Neural Networks," in Proceedings of the 8th Graz Brain Computer Interface Conference 2019, 2019, pp. 338–343. doi: 10.3217/978-3-85125-682-6-62.

[37] 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), Oct. 2019, pp. 1311–1316. doi: 10.1109/SMC.2019.8914246.

[38] D. Y. Lee, M. Lee, and S. W. Lee, "Classification of Imagined Speech Using Siamese Neural Network," in 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Oct. 2020, pp. 2979–2984. doi: 10.1109/SMC42975.2020.9282982.

[39] Y. Wang, P. Wang, and Y. Yu, "Decoding English Alphabet Letters Using EEG Phase Information," Front. Neurosci., vol. 12, p. 62, Feb. 2018, doi: 10.3389/fnins.2018.00062.

[40] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech Recognition with No Speech or with Noisy Speech," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 1090–1094. doi: 10.1109/ICASSP.2019.8683453.

[41] G. Krishna, C. Tran, Y. Han, M. Carnahan, and A. H. Tewfik, "Speech Synthesis Using EEG," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2020, pp. 1235–1238. doi: 10.1109/ICASSP40776.2020.9053340.

[42] 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, 2020, pp. 387–400. doi: 10.1007/978-3-030-41299-9_30.

[43] M. N. I. Qureshi, B. Min, H.-J. Park, D. Cho, W. Choi, and B. Lee, "Multiclass Classification of Word Imagination Speech With Hybrid Connectivity Features," IEEE Trans. Biomed. Eng., vol. 65, no. 10, pp. 2168–2177, Oct. 2018, doi: 10.1109/TBME.2017.2786251.

[44] D. Pawar and S. Dhage, "Imagined Speech Classification using EEG based Brain-Computer Interface," in 2022 IEEE 11th International Conference on Communication Systems and Network Technologies (CSNT), Apr. 2022, pp. 662–666. doi: 10.1109/CSNT54456.2022.9787644.

[45] S. Chengaiyan, A. S. Retnapandian, and K. Anandan, "Identification of vowels in consonant–vowel–consonant words from speech imagery based EEG signals," Cogn. Neurodyn., vol. 14, no. 1, pp. 1–19, Feb. 2020, doi: 10.1007/s11571-019-09558-5.

[46] R. D. Kent, G. Weismer, J. F. Kent, and J. C. Rosenbek, "Toward Phonetic Intelligibility Testing in Dysarthria," J. Speech Hear. Disord., vol. 54, no. 4, pp. 482–499, Nov. 1989, doi: 10.1044/jshd.5404.482.

[47] D. Pal, S. Palit, and A. Dey, "Brain Computer Interface: A Review," in Lecture Notes in Electrical Engineering, vol. 786, 2022, pp. 25–35. doi: 10.1007/978-981-16-4035-3_3.

[48] B. Onaral and A. Cohen, "Biomedical Signals," in Medical Devices and Systems, 3rd ed., J. D. Bronzino, Ed. CRC Press, 2006, pp. 1-1-1–22. doi: 10.1201/9781420003864.sec1.

[49] S. Vorobyov and A. Cichocki, "Blind noise reduction for multisensory signals using ICA and subspace filtering, with application to EEG analysis," Biol. Cybern., vol. 86, no. 4, pp. 293–303, Apr. 2002, doi: 10.1007/s00422-001-0298-6.

[50] S. Cruces, L. Castedo, and A. Cichocki, "Novel blind source separation algorithms using cumulants," in 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.00CH37100), Dec. 2002, vol. 5, no. 1–4, pp. 3152–3155. doi: 10.1109/ICASSP.2000.861206.

[51] A. Mognon, J. Jovicich, L. Bruzzone, and M. Buiatti, "ADJUST: An automatic EEG artifact detector based on the joint use of spatial and temporal features," Psychophysiology, vol. 48, no. 2, pp. 229–240, Feb. 2011, doi: 10.1111/j.1469-8986.2010.01061.x.

[52] S. D. Muthukumaraswamy, "High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations," Front. Hum. Neurosci., vol. 7, 2013, doi: 10.3389/fnhum.2013.00138.

[53] P. Georgieva, F. Silva, M. Milanova, and N. Kasabov, "EEG Signal Processing for Brain–Computer Interfaces," in Springer Handbook of Bio-/Neuroinformatics, no. June 2016, Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 797–812. doi: 10.1007/978-3-642-30574-0_46.

[54] S. Sanei and J. A. Chambers, "Introduction to EEG," in EEG Signal Processing, West Sussex, England: John Wiley & Sons Ltd, 2013, pp. 1–34. doi: 10.1002/9780470511923.ch1.

[55] A. van Boxtel, "Optimal signal bandwidth for the recording of surface EMG activity of facial, jaw, oral, and neck muscles," Psychophysiology, vol. 38, no. 1, p. S004857720199016X, Jan. 2001, doi: 10.1017/S004857720199016X.

[56] G. Hesslow, "Conscious thought as simulation of behaviour and perception," Trends Cogn. Sci., vol. 6, no. 6, pp. 242–247, Jun. 2002, doi: 10.1016/S1364-6613(02)01913-7.

[57] E. F. Chang, K. P. Raygor, and M. S. Berger, "Contemporary model of language organization: an overview for neurosurgeons," J. Neurosurg., vol. 122, no. 2, pp. 250–261, Feb. 2015, doi: 10.3171/2014.10.JNS132647.

[58] G. Hickok and D. Poeppel, "Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language," Cognition, vol. 92, no. 1–2, pp. 67–99, May 2004, doi: 10.1016/j.cognition.2003.10.011.

[59] T. Schultz, M. Wand, T. Hueber, D. J. Krusienski, C. Herff, and J. S. Brumberg, "Biosignal-Based Spoken Communication: A Survey," IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 25, no. 12, pp. 2257–2271, Dec. 2017, doi: 10.1109/TASLP.2017.2752365.

[60] J. Prescott and G. Andrews, "Early and late components of the contingent negative variation prior to manual and speech responses in stutterers and non-stutterers," Int. J. Psychophysiol., vol. 2, no. 2, pp. 121–130, Nov. 1984, doi: 10.1016/0167-8760(84)90005-9.

[61] K. K. Ang, Z. Y. Chin, H. Zhang, and C. Guan, "Filter Bank Common Spatial Pattern (FBCSP) in Brain-Computer Interface," in 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), Jun. 2008, pp. 2390–2397. doi: 10.1109/IJCNN.2008.4634130.

[62] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Hum. Brain Mapp., vol. 38, no. 11, pp. 5391–5420, Nov. 2017, doi: 10.1002/hbm.23730.

[63] K. J. Friston, "Functional and Effective Connectivity: A Review," Brain Connect., vol. 1, no. 1, pp. 13–36, 2011, doi: 10.1089/brain.2011.0008.

[64] P. Babaeeghazvini, L. M. Rueda-Delgado, J. Gooijers, S. P. Swinnen, and A. Daffertshofer, "Brain Structural and Functional Connectivity: A Review of Combined Works of Diffusion Magnetic Resonance Imaging and Electro-Encephalography," Front. Hum. Neurosci., vol. 15, no. October, 2021, doi: 10.3389/fnhum.2021.721206.

[65] K. E. Stephan and K. J. Friston, "Analyzing effective connectivity with functional magnetic resonance imaging," WIREs Cogn. Sci., vol. 1, no. 3, pp. 446–459, May 2010, doi: 10.1002/wcs.58.

[66] C. Meisel and C. Kuehn, "Scaling Effects and Spatio-Temporal Multilevel Dynamics in Epileptic Seizures," PLoS One, vol. 7, no. 2, p. e30371, Feb. 2012, doi: 10.1371/journal.pone.0030371.

[67] R. W. Thatcher, D. North, and C. Biver, "EEG and intelligence: Relations between EEG coherence, EEG phase delay and power," Clin. Neurophysiol., vol. 116, no. 9, pp. 2129–2141, Sep. 2005, doi: 10.1016/j.clinph.2005.04.026.

[68] L. A. Baccalá and K. Sameshima, "Partial directed coherence: a new concept in neural structure determination," Biol. Cybern., vol. 84, no. 6, pp. 463–474, May 2001, doi: 10.1007/PL00007990.

[69] M. J. Kaminski and K. J. Blinowska, "A new method of the description of the information flow in the brain structures," Biol. Cybern., vol. 65, no. 3, pp. 203–210, 1991, doi: 10.1007/BF00198091.

[70] T. Schreiber, "Measuring Information Transfer," Phys. Rev. Lett., vol. 85, no. 2, pp. 461–464, Jul. 2000, doi: 10.1103/PhysRevLett.85.461.

[71] D. Dash, P. Ferrari, and J. Wang, "Spatial and Spectral Fingerprint in the Brain: Speaker Identification from Single Trial MEG Signals," in Interspeech 2019, Sep. 2019, vol. 2019-Septe, pp. 1203–1207. doi: 10.21437/Interspeech.2019-3105.

[72] J. S. Brumberg, E. J. Wright, D. S. Andreasen, F. H. Guenther, and P. R. Kennedy, "Classification of intended phoneme production from chronic intracortical microelectrode recordings in speech-motor cortex," Front. Neurosci., vol. 5, no. 65, May 2011, doi: 10.3389/fnins.2011.00065.

[73] N. Birbaumer et al., "The thought translation device (TTD) for completely paralyzed patients," IEEE Trans. Rehabil. Eng., vol. 8, no. 2, pp. 190–193, Jun. 2000, doi: 10.1109/86.847812.