EEG-Based Imagined-Speech Decoding: A Review

Hatem T M Duhair[1], Masrullizam Bin Mat Ibrahim[2], Mazen Farid[3]*,
Jamil Abedalrahim Jamil Alsayaydeh[4]*, Safarudin Gazali Herawan[5]
Department of Engineering Technology-Fakulti Teknologi Dan Kejuruteraan Elektronik Dan Komputer (FTKEK), Universiti Teknikal Malaysia Melaka (UTeM), 76100 Melaka, Malaysia[1, 2, 4]
Faculty of Information Science and Technology (FIST), Multimedia University, Melaka 75450, Malaysia[3]
Centre for Intelligent Cloud Computing-COE for Advanced Cloud, Multimedia University, Melaka 75450, Malaysia[3]
Industrial Engineering Department-Faculty of Engineering, Bina Nusantara University, Jakarta, Indonesia 11480[5]

***Abstract*—Non-invasive neural speech interfaces aim to reconstruct intended words from brain activity, offering critical communication options for individuals with severe dysarthria or locked-in syndrome. Among the available recording modalities, electroencephalography (EEG) remains the most accessible and cost-effective choice for long-term brain–computer interface (BCI) applications. Decoding imagined speech from EEG, however, remains difficult because of low signal-to-noise ratio, pronounced inter-subject variability, and the small, heterogeneous corpora that are currently available. This review adopts a narrative methodology to synthesise peer-reviewed studies on EEG-based imagined-speech decoding. Relevant articles were identified through keyword-based searches in major digital libraries and were included if they used non-invasive EEG, explicitly instructed imagined or covert speech, and reported quantitative decoding performance. The selected studies are organised along the processing pipeline, from experimental paradigms and data acquisition to preprocessing, feature extraction, representation learning, and classification. Across this body of work, binary imagined-speech tasks that rely on carefully designed time–frequency features and shallow classifiers often report accuracies above 80 percent, whereas multi-class word or phoneme recognition exhibits a much wider spread of performance and remains highly sensitive to dataset design and evaluation protocol. Recent trends favour convolutional and recurrent neural networks, temporal convolutional networks, and transfer learning strategies, which improve performance on some datasets but do not yet resolve fundamental issues of restricted vocabularies, inconsistent evaluation practices, and limited cross-subject generalisation. The review distils these observations into practical recommendations for dataset construction, model design, and evaluation protocols and outlines research directions aimed at more robust and clinically meaningful EEG-based imagined-speech BCIs.**

***Keywords*—Electroencephalography (EEG); Imagined speech; Brain–Computer Interfaces (BCIs); neural speech decoding; deep learning; transfer learning; time–frequency analysis; evaluation protocols**

I. Introduction

Human–computer interaction is increasingly seeking to bypass impaired neuromuscular pathways by directly reading intent from the brain. Brain–computer interfaces (BCIs) instantiate this vision by translating neural activity into actionable commands, opening communication channels for people with severe dysarthria or locked-in syndrome (e.g., ALS) [1]. Among BCI paradigms, imagined speech, which involves mentally articulating words without overt movement, provides a highly natural control signal that aligns closely with the cognitive actions users intend to perform.

Multiple neuroimaging modalities have been explored for imagined speech, such as electroencephalography (EEG) [2], [3], electrocorticography (ECoG) [4], and magnetoencephalography (MEG) [5]. EEG remains the most practical route for broad deployment due to its non-invasiveness, portability, and millisecond-scale temporal resolution [2], [3]. Yet decoding imagined speech from scalp potentials is intrinsically difficult: signals are low-amplitude and noisy, non-stationary across sessions, and highly variable across subjects; vocabularies are small; and existing corpora are heterogeneous and limited in size [6], [7]. These factors jointly confound generalisation and inflate reported accuracies under convenient but optimistic evaluation splits.

Early pipelines relied on handcrafted representations, band-limited filtering, and artefact suppression techniques (such as notch filtering and Independent Component Analysis, or ICA). This was followed by the extraction of features like Common Spatial Patterns (CSP), Power Spectral Density (PSD), wavelets, Hjorth parameters, and Riemannian mappings, which were then fed into conventional classifiers. Recent advancements have shifted towards deep learning methods for end-to-end representation and sequence modelling. Convolutional Neural Networks (CNNs) are now commonly used to capture spatial-spectral structures, while Bidirectional Long Short-Term Memory networks (BLSTMs) and Transformers are employed to model temporal dynamics. This approach is often supplemented by transfer learning and self- or contrastive pretraining to address challenges such as data scarcity and cross-subject variability [8], [9]. Although there has been encouraging progress, the field still lacks unified protocols for train/test splits (intra- vs. cross-subject), consistent metrics (accuracy/F1 vs. sequence-level WER), and transparent baselines with code releases. These limitations hinder reproducibility and fair comparisons across imagined speech datasets [5].

Several review articles have previously examined speech-related brain–computer interfaces and neural speech decoding from a broad perspective, often spanning both invasive and non-invasive modalities or merging imagined speech with overt, whispered, or attempted speech paradigms. These surveys are valuable for framing the overall landscape,

*Corresponding author.

particularly the neurophysiology of speech production and the evolution of decoding architectures, yet they often treat EEG-based imagined speech as a secondary case within a wider taxonomy of neural speech interfaces [45], [46], [47], [48]. As a result, key methodological details that strongly determine EEG performance, such as trial segmentation conventions, artefact handling, representation choices, and cross-subject evaluation practice, are frequently dispersed across sections rather than consolidated into an end-to-end pipeline view. In contrast, the present review is deliberately scoped to non-invasive EEG-based imagined-speech decoding and is organised along the full processing chain from acquisition and signal conditioning to representation learning and evaluation, with the aim of making both the current state of the art and its unresolved gaps more transparent to readers.

This article adopts a narrative review methodology with a defined scope and explicit selection criteria to improve transparency and replicability. Relevant literature was identified through keyword-based searches in major digital libraries, including IEEE Xplore, PubMed, Scopus, and Google Scholar, using combinations of terms such as “imagined speech”, “covert speech”, “silent speech”, “EEG”, “brain–computer interface”, and “neural speech decoding”. Studies were included if they used non-invasive EEG in human participants, employed explicit imagined or covert speech tasks at the level of phonemes, syllables, words, or short phrases, and reported quantitative decoding results. Studies were excluded if they relied solely on invasive recordings, did not include an imagined-speech component, or were non-empirical papers without original experimental evaluation. Screening was performed in two stages, first by title and abstract to remove clearly irrelevant work, then by full-text inspection to verify modality, task design, and reported outcomes. To reduce selection bias and ensure coverage of influential lines of work, citation chaining was applied to key papers, and the final set was prioritised toward peer-reviewed sources that reported sufficient experimental detail to permit methodological comparison, particularly regarding preprocessing, data splits, and evaluation metrics.

The objectives of this review are to:

- Summarise publicly available imagined-speech EEG datasets, their experimental paradigms, acquisition setups, preprocessing techniques, and data-structuring strategies for deep learning;
- Provide actionable recommendations for designing efficient and generalisable deep-learning architectures tailored to EEG-based imagined-speech decoding; and
- Examine the potential of EEG-driven deep learning for advancing neural speech interpretation and assistive communication technologies.

The remainder of this paper is organised as follows. Section II reviews the foundations of EEG, data acquisition protocols, and preprocessing methods. Section III presents feature engineering and deep learning approaches. Section IV discusses evaluation protocols and metrics. Section V synthesises results and discussion, and Section VI concludes with future research perspectives.

II. EEG Foundations, Data Acquisition, and Preprocessing

A. Electroencephalography (EEG)

Electroencephalography (EEG) is a non-invasive method for measuring brain electrical activity, where electrodes placed on the scalp detect voltage differences generated by brain transmissions, forming signals [10]. EEG systems typically use 14–64 electrodes, producing multidimensional signals, and are favoured for Brain-Computer Interfaces (BCIs) due to their non-invasiveness, simplicity, and high temporal resolution. However, EEG is susceptible to motion artefacts and myoelectric interference, especially during spoken language, posing challenges for Automatic Speech Recognition (ASR) [11]. Despite these challenges, EEG has been effectively used to analyse perceived speech and classify imagined phonics [12].

EEG signals are categorised into five frequency bands: gamma (>35 Hz), beta (12–35 Hz), alpha (8–12 Hz), theta (4–8 Hz), and delta (0.5–4 Hz) [13], each of which corresponds to distinct cognitive states. Gamma waves, for example, are linked to overt and covert speech, showing significant changes in various brain regions [14]. Beta waves are associated with muscle activity and speech generation, while alpha waves are crucial for auditory feedback and speech perception, with lower frequencies during covert speech [3]. Theta waves support phonemic restoration and the processing of coarticulation cues [14], aiding in consonant identification [15]. Lastly, delta waves play a role in intonation, rhythm, and other speech-related processes [16].

B. Data Acquisition

Acquiring high-quality EEG data is pivotal in developing Brain-Computer Interface (BCI) systems for deciphering imagined speech. The process involves recording the brain's electrical activity through electrodes on the scalp. This section outlines the key factors and methodologies for collecting EEG data for imagined speech recognition.

In the study, participants were presented with speech cues (either vowels or words) through visual, auditory, or audiovisual means. When these cues were shown before the imagery of speech, participants memorised them, which helped to differentiate the imagined speech task from reading or listening tasks. Participants engaged in imagined speech while simultaneously performing reading or listening tasks in scenarios where cues were provided concurrently. It is important to note that listening and reading activate different areas of the brain: the temporal lobes are involved in listening, while the occipital lobes are activated during reading. Therefore, the format of the cues and their timing in relation to imagined speech can impact brain activation patterns.

The acquisition protocol of the Arizona State University dataset [17] solely employed visual cues, presented simultaneously with an imagined speech recording. Subjects performed speech imagery at each "beep" until cue cessation (7 x T seconds), using short and long words. Common Spatial Pattern (CSP) was applied to identify active brain areas. Results highlighted activity primarily in the left frontal, middle, and parietal regions, corresponding to the motor cortex and

language areas. The KaraOne dataset [18] employing separate cue presentation aimed at distinguishing pronounced and imagined speech states. Central brain areas showed discriminative features. Without a specified rationale, Coretto et al. [19] employed a distinct protocol utilising audio and visual cues before imagined speech tasks. They used vowels and Spanish words, recording EEG data without spatial analysis. Further investigation is warranted to ascertain brain regions involved in cue processing.

Studies have elicited speech imagery at multiple linguistic levels, including vowels [17], phonemes [18], syllables [20], words [21], [22] and sentences [23] .As summarised in Fig. 1, word classification constitutes the largest share of reported EEG imagined-speech tasks, followed by vowel and phoneme or consonant–vowel targets, with a smaller fraction devoted to phrase-level decoding and symbol-like targets (digits or letters). This distribution is not only a matter of experimental convenience. Coarser targets such as isolated words reduce label ambiguity, shorten annotation pipelines, and often permit simpler evaluation, whereas phoneme-level and phrase-level settings impose stricter demands on temporal alignment, representation capacity, and cross-subject robustness. Consequently, comparisons across papers should be interpreted in light of the underlying task unit, because “high accuracy” in a small closed vocabulary does not imply comparable progress toward open-vocabulary decoding.

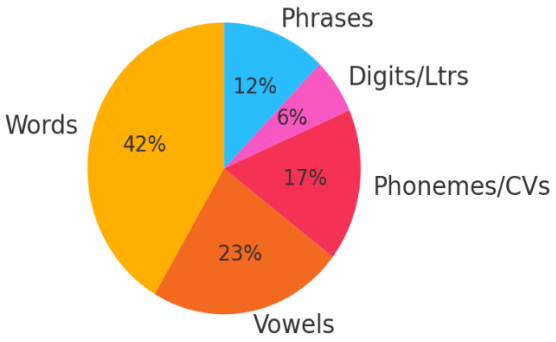

Fig. 1. Speech unit types across studies.

Sub-lexical units, such as vowels, phonemes, and syllables, are often used to focus on early planning and articulatory coding. In contrast, lexical items and binary-response questions (like yes/no) explore message preparation with a greater emphasis on syntactic and semantic complexity. From the literature we reviewed, we identified 28 EEG datasets related to imagined speech: eight are publicly available and twenty are private. Dataset selection is concentrated around a small number of publicly available corpora, which shapes what is routinely benchmarked and what remains underexplored. Fig. 2 shows that Kara One, the Coretto database, and the ASU dataset dominate the empirical evidence base, while other datasets appear only sporadically. This imbalance has practical implications. First, methodological “trends” can become dataset-specific, optimised to a narrow range of paradigms and recording setups. Second, generalisation claims are often constrained by repeated evaluation on the same few corpora, sometimes with inconsistent splitting conventions across studies. For readers, the figure clarifies why cross-dataset conclusions should be framed cautiously, and why broader benchmarking across heterogeneous corpora is essential for credible progress.

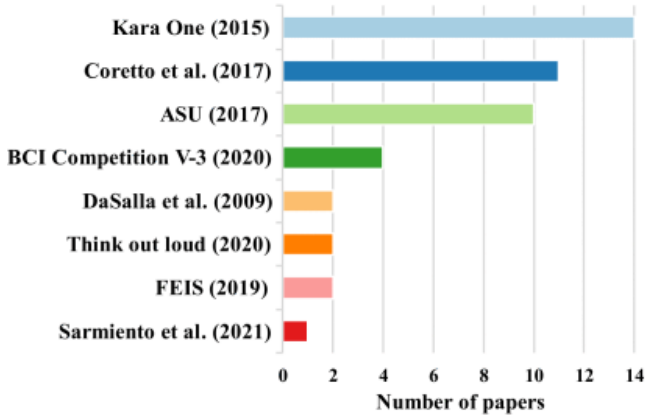

Fig. 2. EEG-based imagined speech public datasets.

In summary, collecting EEG data for imagined speech recognition is a complex process that involves configuring the EEG system, removing artefacts, designing experiments, and taking individual subject considerations into account. Advances in EEG technology and data acquisition methods hold promise for improving brain-computer interface (BCI) systems tailored for imagined speech applications.

C. Preprocessing Techniques

During imagined speech tasks, the raw EEG signals are prone to various artefacts and noise, which can significantly impair subsequent feature extraction and classification stages. Hence, suitable preprocessing methods are crucial for enhancing the signal-to-noise ratio and extracting pertinent information from the EEG data. Nonetheless, effectively cleaning the data without sacrificing important information or features for later analysis or pattern recognition remains challenging. It's crucial to remove noise before downsampling to avoid misinterpreting downsampled values as noise.

In cases where EEG acquisition involves a high sampling rate, such as 1000Hz, downsampling is often employed to balance computational efficiency and data integrity. However, it may lead to the loss of important features. Thus, using the original sample size could facilitate observing discriminative speech recognition features while considering available resources [24].

Standard preprocessing techniques include band-pass filtering, blind source separation (BSS), and subtracting mean values from each channel to remove high-frequency noise and focus on the frequency bands most relevant for speech processes [12]. Notch filters at 60 Hz and band-pass filters with various frequency ranges, such as 0.5-100 Hz [25], 1-50 Hz [26], and 2-40 Hz [27], have been applied to eliminate powerline interference, and the signals were segmented into 5-second epochs with a 0.5-second overlap to capture temporal dynamics. Other preprocessing methods include independent component analysis (ICA) [7], artefact removal [28], sliding window data augmentation [29], and baseline correction [28]. As a result, end-to-end learning methods that require minimal

preprocessing have gained interest in EEG classification. However, directly classifying nearly raw EEG signals remains challenging and requires further investigation [30].

III. Representation Learning and Decoding

In most imagined-speech studies, EEG processing follows a broadly similar structure. After artefact removal and band-limited filtering, continuous recordings are segmented into trials aligned with cue onsets or imagery windows. These trials are re-referenced or normalised, then transformed into representations that emphasise spatial patterns, spectral content, temporal dynamics, or connectivity structure. Handcrafted features compress each trial into a fixed-dimensional vector, which is then passed to a classifier, while deep models often operate directly on multichannel time series or time–frequency maps and learn discriminative features jointly with the decoder. The subsections below describe the main feature-extraction strategies and classification approaches used in this pipeline.

A. Feature Extraction

Feature extraction is a central step in decoding imagined speech from EEG, since it determines how rich but noisy neural activity is converted into stable and discriminative representations for BCI systems. The aim is to retain task-relevant information while suppressing background activity and artefacts. Commonly used transformations include Fourier, wavelet, and Hilbert–Huang decompositions, as well as spatial filtering techniques such as Common Spatial Patterns (CSP) and Principal Component Analysis (PCA). These families of methods, together with their typical application scenarios, can be conveniently summarised in a feature taxonomy table, for example Table I, to provide readers with a compact overview of the design space.

Because EEG is inherently a time-series signal, several studies have relied on time-domain models such as autoregressive (AR) coefficients [31] or borrowed representations from speech processing, such as Mel Frequency Cepstral Coefficients (MFCC) [32]. On the KARAONE dataset, MFCC features achieved higher performance than simple statistical and non-linear descriptors, with reported accuracies of 19.69 percent for MFCC, 15.91 percent for statistical features, and 14.67 percent for non-linear features on an 11-class task, where chance level is 9.09 percent [33]. This pattern is consistent with the intuition that MFCCs capture spectral envelopes and formant-related structure that are more tightly linked to articulatory and phonetic content than raw amplitude statistics. In a related line of work, some authors have treated EEG segments as sequences of local "visual words" and applied Bag of Features (BoF) models, effectively borrowing ideas from text and image representation to capture recurring temporal patterns [34].

Functional and effective connectivity features have received comparatively less attention in imagined-speech decoding, despite their potential to quantify coordinated activity across brain regions. Qureshi et al. [35] employed functional connectivity descriptors, including covariance and the maximum linear cross-correlation coefficient (MaxLCor), and reported 87.90 percent accuracy in a binary imagined-speech classification task. Pawar et al. [36] combined MaxLCor with Discrete Wavelet Transform (DWT) features and obtained 40.64 ± 2.45 percent accuracy. These results suggest that connectivity measures can enhance discrimination, particularly when combined with spectral or time–frequency features, but they have not yet been systematically explored for larger vocabularies or more challenging cross-subject settings.

A further design decision concerns whether features are extracted per channel or jointly across channels. Single-channel analysis is simpler and can highlight localised activity, but simultaneous extraction from multiple channels provides a more realistic view of distributed speech networks. Channel cross-covariance (CCV) matrices are a common way to encode such multichannel structure, since they aggregate relationships between electrodes into a compact form that can be processed by classical or deep models [37]. CCV can be computed in both time and frequency domains and over different window lengths, such as 0.25, 0.5, or 1 second, which allows the representation to trade temporal resolution for robustness [38].

Beyond these families, several works have employed Mel Frequency Cepstral Coefficients (MFCC) [25], Discrete Wavelet Transform (DWT) [21], Wavelet Packet Decomposition (WPD), Short-Time Fourier Transform (STFT) [28], and low-order statistical descriptors [15], [21]. In some cases, investigators have treated the raw microvolt values across channels as a high-dimensional feature vector without additional handcrafted compression [39]. This strategy maximises information content but places a heavier burden on the classifier and typically requires larger datasets or strong regularisation. Overall, the literature reflects a gradual evolution from hand-engineered spectral and spatial markers toward more structured, multichannel representations that are better suited for deep learning.

B. Classification Approaches

Once features have been extracted, the next step is to map each trial to its corresponding imagined-speech category, for example, a word, phoneme, or binary decision. The literature spans a spectrum of classifiers, from shallow machine learning models to deep neural architectures, each motivated by different assumptions about the structure and complexity of EEG data.

Early work relied primarily on traditional machine learning algorithms. Nguyen et al. [40] represented trials as tangent vectors on a Riemannian manifold of covariance matrices and employed a multiclass relevance vector machine to discriminate vowels and short words, achieving accuracies up to 49 percent. This approach leverages the geometry of covariance space to improve robustness but remains limited by linear decision boundaries in the tangent space. Sereshkeh et al. [41] combined autoregressive coefficients and DWT features with a support vector machine (SVM) and achieved 69.3 percent accuracy in online decoding of binary "yes" versus "no" decisions, highlighting the strength of SVMs when feature engineering is carefully tuned to the task. Cooney et al. [33] compared multiple feature sets and found that MFCC features paired with an SVM classifier produced the best performance on an 11-class imagined-speech task, with 22.7

percent accuracy, which exceeds chance but also illustrates the difficulty of multi-class decoding in realistic scenarios.

As datasets and vocabularies grew, many studies shifted toward deep learning to learn representations and decision boundaries jointly. Several efforts have combined Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) to capture spatial and temporal structure, respectively [42]. In some cases, feature vectors, for example channel covariance matrices, are passed through a Deep Autoencoder (DAE) that compresses them into low-dimensional latent codes, which are then used for classification. Siamese networks have been introduced to refine these latent spaces by enforcing that trials with the same label are mapped closer to one another than trials from different classes, which improves discriminability in settings with limited training examples [30]. Using the same Coretto dataset [19], such metric-learning extensions have been reported to outperform baseline architectures without Siamese constraints [27].

To clarify the methodological landscape before comparing individual architectures, Fig. 3 presents the distribution of model families used across the selected studies. The figure is created by categorising each paper according to its primary decoding approach- such as conventional machine learning with handcrafted features, CNN-based models, recurrent or sequence models, and transfer learning or hybrid methods- and then summing these categories throughout the survey. Two key observations emerge: first, deep learning currently dominates recent research, due to its ability to learn task-relevant features directly from noisy, high-dimensional EEG data; second, traditional pipelines remain common in small-data scenarios because their inductive biases and fewer parameters make them easier to train and interpret. This distribution guides the organisation of Section III.B, which compares model families based on the specific problem constraints they address, rather than treating architectures as interchangeable options.

Deep learning architectures explored in this context include pure CNNs, CNN combined with Long Short-Term Memory (LSTM) units, and Deep Autoencoders [7], [28]. Hierarchical designs, in which features are learned at multiple levels using stacked CNN, temporal CNN (TCNN), and DAE modules, have also been proposed [12]. Other classifiers include Random Forests (RF) [15], Support Vector Machines (SVM) [7], K-Nearest Neighbours (KNN), Naive Bayes [39], Deep Belief Networks (DBNs) [43], transfer learning schemes [29], Recurrent Neural Networks (RNNs) [43], Temporal Convolutional Networks (TCNs) [44], bimodal deep neural networks with fusion layers, Transformer-based models, and Capsule Networks. These architectures differ in how they trade off expressiveness, parameter count, and data requirements, but they share the objective of capturing complex, non-linear relationships that cannot be modelled by shallow classifiers.

Reported accuracies span a wide range and depend strongly on task design and dataset characteristics. For binary problems such as distinguishing vowels from consonants (C/V), detecting the presence or absence of nasality (±Nasal), identifying bilabial articulation (±Bilabial), or discriminating specific phonemes like /iy/ and /uw/, accuracies between 69 percent and 89 percent have been reported [26], [45]. Multi-class tasks that target larger phoneme inventories or full words achieve more variable performance, with accuracies reported from around 24.19 percent up to 97.34 percent [12], often under subject-dependent or session-specific splits that may be optimistic. Transfer learning approaches that initialise models from related EEG tasks or from other subjects have also been investigated, with accuracies of 65.65 percent [39] and 95.5 percent [29] reported in particular configurations. These figures demonstrate the potential of advanced classifiers but also underline the difficulty of comparing methods across studies with different protocols.

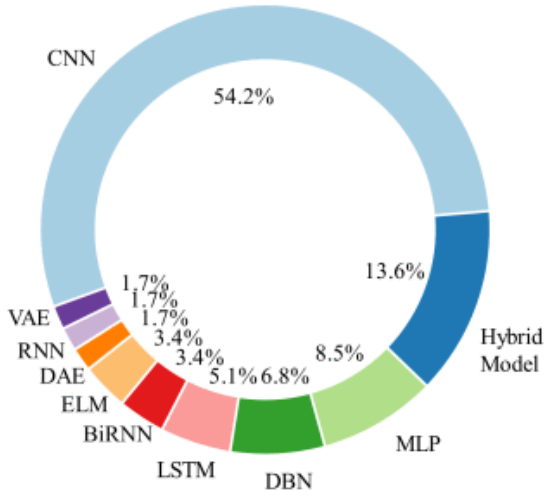

Fig. 3. Distribution of model families used for EEG imagined-speech decoding in our survey.

Despite these advances, several challenges remain for classification in imagined-speech EEG. Inter-subject variability and non-stationarity across sessions make it difficult to design models that generalise reliably beyond the conditions in which they were trained. The limited size of most labelled datasets constrains deep architectures and increases the risk of overfitting, particularly in multi-class problems and when moving toward phrase or sentence-level decoding. Current results therefore, provide important proof-of-concept evidence, but they also point to the need for larger and more diverse datasets, more rigorous cross-subject and cross-session evaluations, and classifier designs that explicitly address variability and uncertainty in real-world settings. These issues are closely linked to the evaluation protocols and metrics discussed in the following section.

IV. Evaluation Protocols and Metrics

The thorough evaluation of imagined-speech decoders relies on three key components: the metric, the validation approach, and the subject segmentation. These factors collectively influence the comparability and reliability of the reported findings across different studies. Reported evaluation practice is dominated by a narrow set of metrics, which partly explains why headline results can be difficult to compare across papers. As summarised in Fig. 4, accuracy is by far the most frequently reported metric, while complementary

measures that reveal class imbalance effects and decision reliability, such as sensitivity, specificity, F-score, and AUC, appear less consistently. Statistical testing and effect-size reporting are rare, despite the known variance induced by subject identity, session effects, and split choice. The practical consequence is that two studies may report similar accuracies while differing substantially in error structure and operational usefulness, particularly in multi-class settings where confusion patterns matter. This motivates the use of a minimal, standard metric set that pairs accuracy with class-sensitive and uncertainty-aware reporting, and it reinforces the need to publish confusion matrices and split protocols alongside aggregate numbers.

In the literature reviewed, accuracy emerges as the most common outcome measure, used in 96.6% of publications, primarily because of its straightforward interpretation for established vocabularies. To assess agreement beyond random chance and reduce issues related to class imbalance, numerous studies also present Cohen's κ (13.6%). Many researchers enhance a single scalar metric with a confusion matrix (27.1%), from which precision/PPV (11.9%), recall/sensitivity/TPR (16.9%), specificity/TNR (2%), and the F-score (15.3%) are calculated. When a classifier's function is based on a continuous decision variable or a threshold score, ROC curves and AUC serve as suitable summaries.

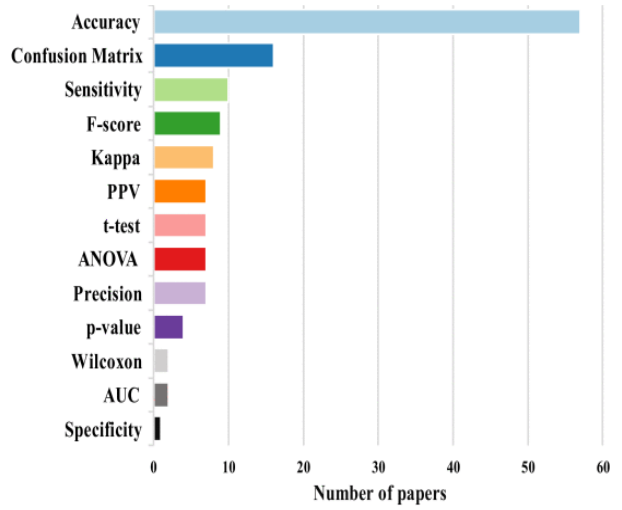

Fig. 4. Frequency of evaluation metrics reported in EEG imagined-speech studies.

Statistical testing is not yet routine in imagined-speech EEG studies, although it is essential for separating genuine methodological gains from variance induced by subject identity and split choice. Across the surveyed papers, roughly one third report inferential evidence, using p-values, t-tests, ANOVA, or non-parametric alternatives such as the Wilcoxon signed-rank test to quantify improvement over baselines. Validation practice is similarly uneven. Hold-out evaluation remains common, but cross-validation is more frequently adopted, and a substantial subset of studies explicitly reserves a validation set for hyperparameter selection, which is particularly important for deep models where tuning decisions can dominate reported performance. These methodological choices interact directly with the most consequential design decision in this literature, namely whether evaluation is subject-dependent or subject-independent. Most studies still follow subject-dependent testing, which typically yields higher scores because the model is assessed on participants represented during training. Consistent with this pattern, stronger results under subject-dependent settings have been reported in [82] and [77]. Subject-independent evaluation is more difficult because inter-subject variability is large and non-stationary, but it is the setting that best reflects practical deployment. For example, although [62] reported that a new deep model improves over conventional baselines, performance across unseen subjects remains weaker than within-subject testing. Against this methodological background, Fig. 5 summarises reported accuracies by target type. Short word sets tend to produce higher central performance than phoneme-level decoding, with vowel classification typically intermediate, a hierarchy that follows expected differences in temporal granularity and class separability. Fig. 5 also includes a sentence-level reference group from high-density ECoG, which provides an upper-bound context given its higher signal fidelity and should not be interpreted as directly comparable to scalp EEG. Overall, the figure reinforces a practical interpretation principle: progress should be judged not only by peak accuracy, but by robustness under stricter splits and by performance retention when moving from simplified targets toward linguistically richer decoding tasks.

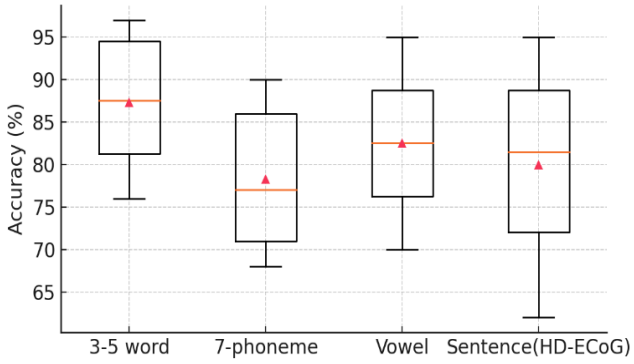

Fig. 5. Reported accuracy ranges per task granularity.

V. Results and Discussion

The surveyed literature shows that EEG-based imagined-speech decoding has moved from proof-of-concept demonstrations to increasingly sophisticated pipelines that combine advanced preprocessing, feature learning, and deep classifiers. Yet, even the strongest reported systems remain some distance from robust, general-purpose communication tools. The gap arises from a combination of signal-level limitations, constrained datasets, substantial inter-subject variability, and heterogeneous evaluation practices that make it difficult to compare methods or to judge their readiness for real-world use.

The corpus of studies considered in this review was drawn from major digital libraries and includes work that focuses on artificial intelligence techniques, feature extraction methods, signal filtering, and data acquisition strategies for speech-related EEG decoding. Most of the selected articles are peer-reviewed journal publications [45], [46], [47], [48], with a

noticeable concentration of contributions from research groups in Korea and India [30], [47], [21], [26]. Foundational work appeared between 2013 and 2016, but the number of publications increased sharply in 2020 and 2021, reflecting a recent shift toward deep learning and transfer learning approaches in this domain.

Table I provides a compact overview of representative studies on EEG-based imagined-speech decoding. It lists, for each work, the task design, preprocessing steps, feature extraction methods, classifiers, and corresponding accuracies. Several patterns emerge. Binary imagined-speech tasks, such as distinguishing a small set of directional commands or phonological contrasts, frequently report accuracies above 80 percent [21], [44], while multi-class settings display wider variation, with values ranging from about 24 percent to above 90 percent depending on vocabulary size, feature choices, and the evaluation protocol [12], [20], [59], [60], [61], [63]. Methods that rely on rich time–frequency representations, including DWT, RADWT, SPWVD, and other time–frequency images coupled with CNN or TCN architectures, account for many of the top-performing models [21], [44], [59]–[61]. Transfer learning, often with ResNet or DenseNet backbones, yields additional gains when labelled data are limited [20], [29], [54].

In comparative terms, EEG-based decoding still trails invasive or high-field modalities in maximum performance. ECoG studies have reported accuracies of 88.3 percent for pairwise imagined-speech classification [49] and 98.8 percent for syllable recognition using spatio-spectral feature clustering [50]. MEG-based systems have achieved phrase classification accuracies around 95 percent [51]. These results illustrate what is possible when spatial resolution and signal-to-noise ratio are high, but they depend on invasive surgery or bulky laboratory hardware. EEG, by contrast, is non-invasive, relatively inexpensive, and portable, which makes it attractive for eventual clinical and home use despite its lower ceiling on accuracy. In practice, current EEG systems are most promising when restricted to modest vocabularies, such as command sets including “up”, “down”, “left”, “right”, “forward”, “backward”, and “select”, which can control screens, mobile devices, or prosthetic devices [52].

To turn this heterogeneous body of work into a coherent roadmap, the remaining discussion is organised into four closely related themes: signal quality and representation, dataset scale and linguistic diversity, individual variability and generalisation, and evaluation protocols and practical deployment. Each theme is framed as an open challenge, then linked to existing studies that partly address it, and finally used to motivate specific directions for future research.

A. Signal Quality and Representation

Low signal-to-noise ratio is an intrinsic limitation of non-invasive EEG. Imagined speech generates weak and spatially diffuse activity that must be recovered in the presence of ocular, muscular, and environmental artefacts. Many of the works in Table I address this by combining classical denoising, such as bandpass and notch filtering, Independent Component Analysis, and Common Average Referencing, with carefully designed features [20], [21], [60], [61]. Time–frequency methods, including DWT, RADWT, SPWVD, and other time–frequency images, further concentrate energy into discriminative bands. When these representations are fed to CNNs, TCNs, or hybrid deep architectures, substantial performance gains are obtained [21], [44], [59]–[61].

These approaches, however, remain tailored to controlled imagery windows and predefined frequency ranges. They are rarely tested in more naturalistic, continuous settings, and their performance may degrade when task timing is variable or when artefact statistics change between sessions. Moreover, most feature pipelines are hand-tuned to particular datasets, which complicates reuse across tasks. One natural extension of the current literature is to replace fixed front-ends with adaptive or learned ones that are trained jointly with the classifier, for example, denoising autoencoders, learnable filterbanks, or self-supervised objectives that encourage separation of neural signal and artefacts. Such work would build directly on the existing time–frequency and connectivity-based representations, but would aim to make them more robust and less dependent on manual feature engineering.

B. Dataset Scale, Vocabulary, and Linguistic Diversity

A second, closely related challenge concerns dataset size and linguistic coverage. Many studies in Table I employ relatively small subject pools and limited vocabularies, often focusing on a handful of vowels, short words, or simple commands [21], [44], [59]–[61], [63]. This is understandable, since imagined-speech experiments are demanding and EEG recording sessions are time-consuming. Nevertheless, small datasets restrict the expressive power that can be used in models, and they make it hard to separate genuine algorithmic improvements from overfitting to narrow tasks. A few works have attempted to move beyond this constraint by combining datasets such as FEIS and KARAONE [20], [29], [59] or by constructing tasks involving subject, verb, and object words [62]. These studies demonstrate that richer vocabularies are technically feasible, but they also highlight the need for more extensive collections of trials if vocabulary size and linguistic variability are to increase.

Language and phoneme variability adds another layer of complexity that has not yet been fully explored. Most experiments are carried out in a single language, with phoneme sets chosen for convenience or for strong acoustic–articulatory contrasts. FEIS and KARAONE provide broader phonetic inventories and are valuable in this respect [20], [59], [60], but there is little systematic work on how language-specific phonology or prosody influences EEG decodability. Future datasets that are designed from the outset to be multi-lingual and multi-phonemic, with shared vocabularies and standardised splits, would enable these questions to be addressed. They would also provide a more solid foundation for the transfer learning strategies already investigated in [20], [29], [54], which rely on diverse data to learn generalisable representations.

C. Individual Variability and Generalisation

Inter-subject and inter-session variability remains one of the most significant barriers to robust imagined-speech decoding. Anatomical differences, variations in electrode placement, and individual cognitive strategies all contribute to

changes in the recorded signal. Many of the highest accuracies in Table I come from subject-dependent or mixed-trial evaluations, where training and test sets share data from the same individuals and recording sessions [44], [59]–[61], [63]. In such settings, complex models can learn idiosyncratic patterns and achieve strong performance, but this success does not guarantee generalisation to new users or days.

TABLE I. SUMMARY OF REPRESENTATIVE STUDIES ON EEG-BASED IMAGINED SPEECH DECODING

Ref.	Task	Preprocessing Method	Feature Extraction	Classification Method	Accuracy
[21]	Muti-class and binary classification of “left,” “right,” “up,” and “down.”	ICA	DWT	- Kernel ELM - Statistical features	Multi-class: 49% Binary: 85%
[54]	Classification of five vowels and six words	Transfer learning	CNN		24%
[55]	addition of new classes to a pre-trained classifier with few trials.	instantaneous frequency and spectral entropy	Deep metric learning framework		-6-class accuracy of 45.00 ± 3.13%. - 5-class accuracy of 48.10 ± 3.68%
[29]	Vowels, Short and long words	- Sliding window data augmentation - Band-pass filtering	MPC MSC	Transfer learning with ResNet50 pre-trained	- 79.7% (min) for vowels - 95.5% (max) for short-long words
[44]	Vowels and words	FastICA	DWT	TCN and CNN	96.49%
[56]	words "rock", "paper", "scissors", and rest state	Classical signal processing	GSP/GL	SVM	50.10%
[57]	long words, short words, vowels	Operational Architectonics	network metrics	Naive Bayes	-62.55% (long words), 66.44% (long vs short), 53.47% (short words), 48.13% (vowels)
[58]	kinesthetic imagery of the “left hand” and “right hand.” . visualise (“split” and “fall in”), imagine the pronunciations (“go” and “stop”),	multiscale convolutional transformer	t-SNE	multiscale convolutional transformer	- 0.62 on the private EEG dataset. - 0.70 on BCI competition IV 2a dataset. - 0.72 on the Arizona State University dataset.
[59]	Vowels, Short and long words classification	SPWVD	TFR images	CNN	- Long words: 94.82% - Short-long words: 94.26% - Short words: 94.68% - Vowels: 84.50%
[60]	Vowels, Short and long words classification	Butterworth bandpass filter, Notch filter, Common average referencing	RADWT/PSO	SVM, KNN, Random Forest, Rotation Forest, Bagging, AdaBoos	87.26% 89.23%, 95.5%, and 92.16% for long words, short–long words, short words, and vowels, respectively.
[61]	Vowels, Short and long words classification	Butterworth bandpass filter, Notch filter, Common average referencing	SPWVD	CNN	Binary: 79.82% to 82.04% Multiclass: 49.93% to 51.44%
[62]	The words included “I” and “partner” as subject words, “move,” “have,” and “drink” as verb words, and “box,” “cup,” and “phone” as object words	Downsampling from 1000 to 500 Hz, Band-pass filtering (30-125 Hz), and Embedding of spectrograms	CNNs and ground truth (mel-spectrogram), (GRU)-based regression	Structural Similarity Index Measure (SSIM)	Subject words 79.2%, Verb words 82.5%, Object words 81.1%
[3]	Four words: UP, DOWN , LEFT, right	Bandpass filtering (10-100 Hz)	WST	LSTM model with L2 regularisation	92.50%
[20]	FEIS : /i/, /u:/, /æ/, /ɔ:/, /m/, /n/, /ŋ/, /f/, /s/, /ʃ/, /v/, /z/, /ʒ/, /p/, /t/, /k/ .KaraOne : /iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/,pat, pot, knew, and gnaw	ICA and bandpass filtering	DWT	Deep transfer learning (DenseNet, ResNet)	- KaraOne: 82.35% - FEIS: 89.01%
[63]	/iy/, /uw/, /piy/, /tiy/, /diy/, /m/, /n/,pat, pot, knew, and gnaw	filter up to 1 kHz bandpass	EM-CSP	Ensemble stacking learning	97.34%
[21]	Multi-class and binary classification of “left,” “right,” “up,” and “down.”	ICA	DWT	- Kernel ELM - Statistical features	Multi-class: 49% Binary: 85%

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, 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, t-SNE= t–stochastic neighbor embedding, CAR=Common Average Reference, RVM = relevance vector machine (VM), STFT = Short Time FT, SVM = support VM, TL = transfer learning.

Several studies have started to reduce this gap. Transfer learning approaches initialise models on one dataset or group of subjects and then fine-tune them on a smaller target set. For example, [29] uses ResNet-based transfer with sliding-window augmentation, and [20] explores knowledge transfer between FEIS and KARAONE, reporting accuracies above 80 percent for some configurations. Metric-learning strategies, such as Siamese networks and deep metric learning frameworks [30],

[55], encourage models to place trials with the same label closer together in latent space and trials with different labels further apart. These methods help to reduce subject-specific variability and have outperformed simpler baselines on the same data [27], [30].

Even so, most of these works still require subject-specific fine-tuning and have not yet been evaluated on large, independent cohorts. A promising direction is the development of architectures and training regimes that explicitly disentangle subject and task factors, for instance, through adversarial domain adaptation, subject embedding layers, or factorised latent spaces. The existing transfer and metric-learning literature provides concrete starting points, but more ambitious cross-subject and cross-session benchmarks will be needed to identify which strategies lead to genuine gains in generalisation rather than to improvements confined to a single dataset.

D. Evaluation Protocols and Practical Deployment

The way performance is measured is as important as the models themselves. Evaluation protocols in the imagined-speech literature vary widely. Many studies employ subject-dependent or random cross-validation schemes that mix trials from the same recording session in training and test sets. This practice often yields optimistic estimates of accuracy, because nuisance factors such as noise statistics and electrode placement are shared across splits. Only a subset of studies explicitly report cross-subject or cross-session results, and for widely used datasets such as KARAONE, FEIS, or the Arizona State University corpus, there is still no consensus on standard training and test partitions [3], [20], [58], [59], [63].

Some authors have already taken steps toward more transparent evaluation. For example, [21] distinguishes between binary and multi-class performance on directional commands, [59] and [60] report separate accuracies for vowels, short words, and long words, and [62] analyses subject, verb, and object categories separately. These distinctions make it easier to understand how a model behaves under different task demands. However, further progress will likely require community-wide agreement on recommended protocols, such as fixed cross-subject splits, reporting of both accuracy and F1-score, and the introduction of sequence-level metrics, such as word error rate, when moving toward continuous decoding.

Practical deployment also raises questions that are only beginning to be addressed. Online or real-time experiments remain rare, although [41] demonstrates that binary “yes/no” decisions can be decoded in an online setting with reasonable accuracy. Most of the work summarised in Table I is still offline. Bringing systems into real-time use will require efficient preprocessing, models that can run with limited hardware, and calibration procedures that minimise user burden. The command-based applications discussed in [52] suggest that small-vocabulary control may already be within reach, provided that stability over time and across sessions can be established.

E. Synthesis

Taken together, the available evidence suggests a field that is technically vibrant but structurally constrained. Time–frequency and connectivity features coupled with CNNs, RNNs, and transfer learning have demonstrated that imagined-speech decoding from EEG is feasible and that accuracies can be high on carefully controlled tasks [20], [21], [44], [59]–[63]. At the same time, low signal-to-noise ratio, limited and linguistically narrow datasets, strong individual variability, and heterogeneous evaluation protocols restrict the generalisability of these findings. The four themes discussed above are tightly interconnected. Better representations depend on richer and more diverse data. Robust generalisation relies on both sophisticated modelling and realistic cross-subject evaluations. Practical deployment in turn depends on all three: signal quality, data and model scale, and credible performance estimates.

If future work can align efforts along these lines, for example, by constructing larger multi-site datasets, adopting shared benchmark protocols, and exploiting representation learning techniques that explicitly tackle subject and session variability, EEG-based imagined-speech interfaces may evolve from laboratory prototypes into reliable communication tools for individuals with severe motor and speech impairments [53]. In that sense, the current literature should be viewed not only as a catalogue of methods and accuracies, but also as a foundation on which a more standardised and clinically relevant generation of imagined-speech BCIs can be built.

VI. Conclusion

EEG-based imagined-speech decoding is a promising yet challenging area in non-invasive brain-computer interface research. Despite significant progress in recent years, the field is still limited by low signal-to-noise ratios, high variability among subjects, and a lack of large, standardised datasets. This review consolidates the various methodologies employed throughout the decoding process, covering EEG acquisition, preprocessing, feature extraction, and deep learning-based representation learning, to offer a comprehensive overview of the current state of the art. The synthesis of existing literature indicates a clear shift from traditional handcrafted signal-processing techniques to data-driven deep learning architectures. These new methods can learn spatial, spectral, and temporal representations concurrently. However, methodological inconsistencies—especially regarding dataset selection, evaluation protocols, and performance metrics—persist and hinder reproducibility and direct comparisons between studies. To accelerate advancements in this field, future research should focus on establishing unified benchmark datasets and cross-subject evaluation frameworks. Additionally, incorporating multimodal signals, as well as contrastive and self-supervised learning strategies, and domain adaptation techniques may enhance the generalizability of the models. Furthermore, interdisciplinary collaboration between neuroscientists, linguists, and machine learning researchers is vital to ensure that the models developed are physiologically interpretable and ethically sound. Ultimately, the combination of EEG and deep learning in imagined-speech decoding has the potential to transform assistive communication technologies. Achieving this goal will necessitate not only innovative algorithms but also rigorous experimental standardisation and open, collaborative research practices that promote transparency, reproducibility, and equitable progress across the field.

ACKNOWLEDGMENT

The authors would like to express their sincere gratitude to Multimedia University and Universiti Teknikal Malaysia Melaka (UTeM) for their valuable support in this research.

AUTHORS' CONTRIBUTIONS

The authors' contributions are as follows: "Conceptualization, Jamil Abedalrahim Jamil Alsayaydeh and Mazen Farid; methodology, Hatem T M Duhair; software, Jamil Abedalrahim Jamil Alsayaydeh; validation, Hatem T M Duhair; formal analysis, Masrullizam Bin Mat Ibrahim; investigation, Jamil Abedalrahim Jamil Alsayaydeh; resources, Masrullizam Bin Mat Ibrahim; writing—original draft preparation, Jamil Abedalrahim Jamil Alsayaydeh and Safarudin Gazali Herawan; writing—review and editing, Mazen Farid; funding acquisition, Hatem T M Duhair and Safarudin Gazali Herawan.

DATA AVAILABILITY STATEMENT

All the datasets used in this study are available from the Zenodo database (accession number: https://zenodo.org/records/17313452).

REFERENCES

[1] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A State-of-the-Art Review of EEG-Based Imagined Speech Decoding," Frontiers in Human Neuroscience, vol. 16, 2022, Accessed: Nov. 13, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fnhum.2022.867281

[2] D. Jaipriya and K. C. Sriharipriya, "Brain Computer Interface-Based Signal Processing Techniques for Feature Extraction and Classification of Motor Imagery Using EEG: A Literature Review," Biomedical Materials & Devices, May 2023, doi: 10.1007/s44174-023-00082-z.

[3] M. M. Abdulghani, W. L. Walters, and K. H. Abed, "Imagined Speech Classification Using EEG and Deep Learning," Bioengineering, vol. 10, no. 6, 2023, doi: 10.3390/bioengineering10060649.

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

[5] J. S. García-Salinas, A. A. Torres-García, C. A. Reyes-Garćia, and L. Villaseñor-Pineda, "Intra-subject class-incremental deep learning approach for EEG-based imagined speech recognition," Biomedical Signal Processing and Control, vol. 81, 2023, doi: 10.1016/j.bspc.2022.104433.

[6] Z. Khademi, F. Ebrahimi, and H. M. Kordy, "A review of critical challenges in MI-BCI: From conventional to deep learning methods," Journal of Neuroscience Methods, vol. 383, p. 109736, Jan. 2023, doi: 10.1016/j.jneumeth.2022.109736.

[7] 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, 2020, doi: 10.3390/s20164629.

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

[9] H. Chang, J. Han, C. Zhong, A. M. Snijders, and J.-H. Mao, "Unsupervised Transfer Learning via Multi-Scale Convolutional Sparse Coding for Biomedical Applications," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 5, pp. 1182–1194, May 2018, doi: 10.1109/TPAMI.2017.2656884.

[10] A. Singh and A. Gumaste, "Decoding Imagined Speech and Computer Control using Brain Waves," Journal of neuroscience methods, vol. 358, p. 109196, Apr. 2021, doi: 10.1016/j.jneumeth.2021.109196.

[11] M. P.P., T. Thomas, and R. Gopikakumari, "Wavelet feature selection of audio and imagined/vocalized EEG signals for ANN based multimodal ASR system," Biomedical Signal Processing and Control, vol. 63, p. 102218, Jan. 2021, doi: 10.1016/j.bspc.2020.102218.

[12] P. Saha, M. Abdul-Mageed, and S. Fels, "Speak Your Mind! Towards Imagined Speech Recognition with Hierarchical Deep Learning," in Proc. Interspeech, 2019, doi: 10.21437/Interspeech.2019-3041.

[13] J. A. J. Alsayaydeh, Irianto, M. F. Ali, M. N. M. Al-Andoli and S. G. Herawan, "Improving the Robustness of IoT-Powered Smart City Applications Through Service-Reliant Application Authentication Technique," in IEEE Access, vol. 12, pp. 19405-19417, 2024, doi: 10.1109/ACCESS.2024.3361407.

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

[15] M. N. Al-Andoli, Irianto, J. A. Alsayaydeh, I. M. Alwayle, C. K. N. Che Ku Mohd and F. Abuhoureyah, "Robust Overlapping Community Detection in Complex Networks With Graph Convolutional Networks and Fuzzy C-Means," in IEEE Access, vol. 12, pp. 70129-70145, 2024, doi: 10.1109/ACCESS.2024.3399883.

[16] V. J. Boucher, A. C. Gilbert, and B. Jemel, "The role of low-frequency neural oscillations in speech processing: revisiting delta entrainment," Journal of Cognitive Neuroscience, vol. 31, no. 8, pp. 1205–1215, 2019.

[17] 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.

[18] 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), IEEE, 2015, pp. 992–996. Accessed: Nov. 21, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/7178118/?casa_token=NMZgHVxYQdsAAAAA:KEDP1GPpEHwzRdSenoo96oF05i4y8f6uzrqxtDNq4-5fihTYHPMoKIB5wIULTTzYkdpDobpQmg

[19] G. A. P. 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, SPIE, 2017, p. 1016002. Accessed: Nov. 21, 2023. [Online]. Available: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10160/1016002/Open-access-database-of-EEG-signals-recorded-during-imagined-speech/10.1117/12.2255697.short

[20] N. C. Mahapatra and P. Bhuyan, "Decoding of imagined speech electroencephalography neural signals using transfer learning method," Journal of Physics Communications, vol. 7, no. 9, 2023, doi: 10.1088/2399-6528/ad0197.

[21] D. Pawar and S. Dhage, "Multiclass covert speech classification using extreme learning machine," Biomedical Engineering Letters, vol. 10, no. 2, pp. 217–226, 2020.

[22] V. Shkarupylo, J. A. J. Alsayaydeh, M. F. B. Yusof, A. Oliinyk, V. Artemchuk and S. G. Herawan, "Exploring the Potential Network Vulnerabilities in the Smart Manufacturing Process of Industry 5.0 via the Use of Machine Learning Methods," in IEEE Access, vol. 12, pp. 152262-152276, 2024, doi: 10.1109/ACCESS.2024.3474861.

[23] P. D. Barua et al., "Automated EEG sentence classification using novel dynamic-sized binary pattern and multilevel discrete wavelet transform techniques with TSEEG database," Biomedical Signal Processing and Control, vol. 79, p. 104055, Jan. 2023, doi: 10.1016/j.bspc.2022.104055.

[24] L. M. Martinon, J. Smallwood, D. McGann, C. Hamilton, and L. M. Riby, "The disentanglement of the neural and experiential complexity of self-generated thoughts: A users guide to combining experience sampling with neuroimaging data," NeuroImage, vol. 192, pp. 15–25, May 2019, doi: 10.1016/j.neuroimage.2019.02.034.

[25] A.-L. Rusnac and O. Grigore, "Generalized Brain Computer Interface System for EEG Imaginary Speech Recognition," in 2020 24th International Conference on Circuits, Systems, Communications and Computers (CSCC), IEEE, 2020, pp. 184–188.

[26] R. A. Sharon and H. A. Murthy, "Correlation based Multi-phasal Models for Improved Imagined Speech EEG Recognition," in Proc. ISCA Speech, Music and Mind Workshop (SMM), 2020.

[27] J. A. J. Alsayaydeh, Irianto, M. Zainon, H. Baskaran, and S. G. Herawan, "Intelligent interfaces for assisting blind people using object recognition methods," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 5, pp. 84–92, 2022, doi: 10.14569/IJACSA.2022.0130584.

[28] S. Datta and N. V. Boulgouris, "Recognition of grammatical class of imagined words from EEG signals using convolutional neural network," Neurocomputing, vol. 465, pp. 301–309, Nov. 2021, doi: 10.1016/j.neucom.2021.08.035.

[29] J. T. Panachakel and R. A. Ganesan, "Decoding Imagined Speech From EEG Using Transfer Learning," IEEE Access, vol. 9, pp. 135371–135383, 2021, doi: 10.1109/ACCESS.2021.3116196.

[30] J. A. J. Alsayaydeh, Irianto, A. Aziz, C. K. Xin, A. K. M. Z. Hossain, and S. G. Herawan, "Face recognition system design and implementation using neural networks," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 13, no. 6, pp. 63–69, 2022, doi: 10.14569/IJACSA.2022.0130663.

[31] 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, IEEE, 2014, pp. 568–571. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/6981789/?casa_token=pYAsqfKrCl4AAAAA:eK_kCeT5P50iwWBMfzU8aMVgXiZEvpkZohNjRXZl8rQ1fMe4I2_m8pGdrwE-g9lSBCCCGJ8Ctw

[32] 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), IEEE, 2020, pp. 1235–1238. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9053340/?casa_token=dxhgPvVLni4AAAAA:I_Qh9ChhOxIAsC9lUXjKreFgUsU-4T0vVaKWQ9k_SoUQ9xUR6bcU53AADBjkYaqJyFOrJwPMPw

[33] C. Cooney, R. Folli, and D. Coyle, "Mel frequency cepstral coefficients enhance imagined speech decoding accuracy from EEG," in 2018 29th Irish Signals and Systems Conference (ISSC), IEEE, 2018, pp. 1–7.

[34] M. Jiménez-Guarneros and P. Gómez-Gil, "Standardization-refinement domain adaptation method for cross-subject EEG-based classification in imagined speech recognition," Pattern Recognition Letters, vol. 141, pp. 54–60, 2021.

[35] M. N. I. Qureshi, B. Min, H. Park, D. Cho, W. Choi, and B. Lee, "Multiclass classification of word imagination speech with hybrid connectivity features," IEEE Transactions on Biomedical Engineering, vol. 65, no. 10, pp. 2168–2177, 2017.

[36] 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), IEEE, 2022, pp. 662–666. Accessed: Dec. 05, 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9787644/?casa_token=q5YZSt-70HgAAAAA:3jTrKlxnQOJjSbaPu6dC9Ilr1B1hhSbia7GrKko6EdG9lSa-blnvzPpn7EBaWnD0JjEkF1fd7A

[37] P. Saha, S. Fels, and M. Abdul-Mageed, "Deep learning the EEG manifold for phonological categorization from active thoughts," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, pp. 2762–2766.

[38] A.-L. Rusnac and O. Grigore, "CNN Architectures and Feature Extraction Methods for EEG Imaginary Speech Recognition," Sensors, vol. 22, no. 13, p. 4679, 2022.

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

[40] A. Hossain, K. Das, P. Khan, and Md. F. Kader, "A BCI system for imagined Bengali speech recognition," Machine Learning with Applications, vol. 13, p. 100486, Sept. 2023, doi: 10.1016/j.mlwa.2023.100486.

[41] A. R. Sereshkeh, R. Trott, A. Bricout, and T. Chau, "Online EEG Classification of Covert Speech for Brain–Computer Interfacing," Int. J. Neur. Syst., vol. 27, no. 08, p. 1750033, Dec. 2017, doi: 10.1142/S0129065717500332.

[42] P. Saha and S. Fels, "Hierarchical deep feature learning for decoding imagined speech from EEG," in Proceedings of the AAAI Conference on Artificial Intelligence, 2019, pp. 10019–10020.

[43] 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.

[44] N. C. Mahapatra and P. Bhuyan, "Multiclass Classification of Imagined Speech Vowels and Words of Electroencephalography Signals Using Deep Learning," Advances in Human-Computer Interaction, vol. 2022, 2022.

[45] V. Shkarupylo, I. Blinov, A. Chemeris, V. Dusheba, J. A. J. Alsayaydeh and A. Oliinyk, "Iterative Approach to TLC Model Checker Application," 2021 IEEE 2nd KhPI Week on Advanced Technology (KhPIWeek), Kharkiv, Ukraine, 2021, pp. 283-287, doi: 10.1109/KhPIWeek53812.2021.9570055.

[46] 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, 2020.

[47] D. -Y. Lee, M. Lee, and S. -W. Lee, "Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1363–1374, 2021, doi: 10.1109/TNSRE.2021.3096874.

[48] F. Li et al., "Decoding imagined speech from EEG signals using hybrid-scale spatial-temporal dilated convolution network," Journal of Neural Engineering, vol. 18, no. 4, p. 0460c4, 2021.

[49] S. Martin et al., "Word pair classification during imagined speech using direct brain recordings," Scientific reports, vol. 6, no. 1, pp. 1–12, 2016.

[50] M. Matsumoto and J. Hori, "Classification of silent speech using support vector machine and relevance vector machine," Applied Soft Computing, vol. 20, pp. 95–102, 2014.

[51] D. Dash, P. Ferrari, D. Heitzman, and J. Wang, "Decoding speech from single trial MEG signals using convolutional neural networks and transfer learning," in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), IEEE, 2019, pp. 5531–5535.

[52] M. R. Asghari Bejestani, M. Khani, V. R. Nafisi, and F. Darakeh, "Eeg-based multiword imagined speech classification for Persian words," BioMed Research International, vol. 2022, 2022.

[53] A. Rezazadeh Sereshkeh, R. Yousefi, A. T. Wong, F. Rudzicz, and T. Chau, "Development of a ternary hybrid fNIRS-EEG brain–computer interface based on imagined speech," Brain-Computer Interfaces, vol. 6, no. 4, pp. 128–140, 2019, doi: 10.1080/2326263X.2019.1698928.

[54] M.-O. Tamm, Y. Muhammad, and N. Muhammad, "Classification of vowels from imagined speech with convolutional neural networks," Computers, vol. 9, no. 2, p. 46, 2020.

[55] D. -Y. Lee, M. Lee, and S. -W. Lee, "Decoding Imagined Speech Based on Deep Metric Learning for Intuitive BCI Communication," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 29, pp. 1363–1374, 2021, doi: 10.1109/TNSRE.2021.3096874.

[56] A. Einizade, M. Mozafari, S. Jalilpour, S. Bagheri, and S. Hajipour Sardouie, "Neural decoding of imagined speech from EEG signals using the fusion of graph signal processing and graph learning techniques," Neuroscience Informatics, vol. 2, no. 3, p. 100091, Sept. 2022, doi: 10.1016/j.neuri.2022.100091.

[57] A. C. Iliopoulos and I. Papasotiriou, "Functional Complex Networks Based on Operational Architectonics: Application on EEG-based Brain–computer Interface for Imagined Speech," Neuroscience, vol. 484, pp. 98–118, 2022, doi: 10.1016/j.neuroscience.2021.11.045.

[58] H. -J. Ahn, D. -H. Lee, J. -H. Jeong, and S. -W. Lee, "Multiscale Convolutional Transformer for EEG Classification of Mental Imagery in Different Modalities," IEEE Transactions on Neural Systems and Rehabilitation Engineering, vol. 31, pp. 646–656, 2023, doi: 10.1109/TNSRE.2022.3229330.

[59] A. Kamble, P. H. Ghare, and V. Kumar, "Deep-Learning-Based BCI for Automatic Imagined Speech Recognition Using SPWVD," IEEE

Transactions on Instrumentation and Measurement, vol. 72, pp. 1–10, 2023, doi: 10.1109/TIM.2022.3216673.

[60] A. Kamble, P. H. Ghare, and V. Kumar, "Optimized Rational Dilation Wavelet Transform for Automatic Imagined Speech Recognition," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–10, 2023, doi: 10.1109/TIM.2023.3241973.

[61] A. Kamble, P. H. Ghare, V. Kumar, A. Kothari, and A. G. Keskar, "Spectral Analysis of EEG Signals for Automatic Imagined Speech Recognition," IEEE Transactions on Instrumentation and Measurement, vol. 72, pp. 1–9, 2023, doi: 10.1109/TIM.2023.3300473.

[62] J. -H. Jeong, J. -H. Cho, B. -H. Lee, and S. -W. Lee, "Real-Time Deep Neurolinguistic Learning Enhances Noninvasive Neural Language Decoding for Brain–Machine Interaction," IEEE Transactions on Cybernetics, vol. 53, no. 12, pp. 7469–7482, Dec. 2023, doi: 10.1109/TCYB.2022.3211694.

[63] D. Alizadeh and H. Omranpour, "EM-CSP: An efficient multiclass common spatial pattern feature method for speech imagery EEG signals recognition," Biomedical Signal Processing and Control, vol. 84, 2023, doi: 10.1016/j.bspc.2023.104933.