International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Interpretable Brain-Computer Interface For Emotion-Aware Thought-To-Speech Using Eeg Signal Decoding

Abhimanyu Singh[1], Edith Paulin S[2]

[1]Grade 12 Student,GEMS New Millennium School, Dubai , UAE
[2]Computer Science Teacher, GEMS New Millennium School, Dubai , UAE

Abstract
Restoring effective communication for people with severe speech impairment remains a most important concern in assistive healthcare. Electroencephalography (EEG)-Based Brain–Computer Interfaces (BCI) can decode imagined speech from neural activity; however, existing techniques have been limited by a lack of interpretability, reduced robustness across users, and inadequate integration of emotional context. In this study, we present an Interpretable Brain–Computer Interface for Emotion-Aware Thought-to-Speech that incorporates affective state recognition, EEG signal decoding, and explainable deep learning within a unified framework. The presented framework combines advanced EEG preprocessing, spatio-temporal feature representation, attention-guided neural learning, and an explainability module while identifying the underlying neural features and cortical regions influencing each prediction. Emotional state estimation is combined with speech intention decoding to produce speech that reflects contextual expressiveness, enhancing the naturalness and effectiveness of human–computer interaction. The framework is tested on publicly available EEG datasets of imagined speech, using subject-independent setups, and evaluated based on classification accuracy, precision, recall, F1-score, inference speed, and interpretability. Results show that the proposed approach consistently enhances decoding reliability and offers clear, clinically meaningful insights compared to traditional opaque models. By simultaneously addressing speech intent, emotional nuance, and model transparency, this study contributes to the advancement of trustworthy, next-generation brain–computer interfaces and provides a practical basis for assistive communication tools serving individuals with paralysis, amyotrophic lateral sclerosis, locked-in syndrome, and other conditions that disrupt natural speech.

Keywords: Brain–Computer Interface (BCI), Electroencephalography (EEG), Thought-to-Speech, Imagined Speech Decoding, Emotion-Aware Communication, Explainable Artificial Intelligence (XAI), Interpretable Deep Learning

I. Introduction

1.1 Background

One of the most promising technologies that provides a direct link between the human brain and external computing systems is Brain–Computer Interfaces (BCIs). BCIs provide an alternative way of communication for people who can't communicate using their mouth or muscles. BCIs can help people communicate when they are not able to do so through speaking or using their muscles.

Electroencephalography (EEG) is the most widely used of the various neuroimaging modalities, as it is non-invasive, portable, relatively inexpensive, and can provide a high temporal resolution of neural dynamics. The benefits have made EEG-based BCIs the desirable choice for assistive communication, neurorehabilitation, and human–computer interaction.

The decoding of imagined or covert speech from EEG signals has greatly benefited from recent developments in machine learning and deep neural networks. Previous research showed that deep convolutional and recurrent networks can achieve good results in characterizing imagined vowels, words, and short speech commands from EEG recordings, thereby proving the feasibility of non-invasive thought-based communication [1]–[4]. Neuroimaging studies of functional brain connectivity also found that imagined speech involves widespread, not localized, distributed areas, which suggests that more than just spatial dependency should be taken into account in feature learning [5]. Publicly available datasets have also helped to rapidly progress the development of algorithms by providing standardized benchmarks for comparing imagined speech recognition in various languages, experimental paradigms, and recording conditions [6],[12],[17]–[20].

The most significant uses of BCIs nowadays are speech restoration. People with ALS (amyotrophic lateral sclerosis), locked-in syndrome, stroke, traumatic brain injury, and other advanced neuromuscular diseases often have normal mental function but have no motor control over their speech. Current assistive communication devices using eye tracking, switch control, or residual muscle activity are limited in their ability to communicate quickly and require a great deal of physical exertion. Thought decoding with EEG has the potential to provide direct communication restoration from the users' thoughts, which is less dependent on residual motor function and more natural. Thus, the study of the decoding of imagined speech has gained more and more importance in the field of assistive communication technologies.

1.2 Motivation

Although impressive advances have been made in the development of EEG-based imagined speech recognition, current communication systems are still nowhere near the richness of natural human conversation. Humans communicate and express not only the message, but the intentions of the message. The same verbal "sentence" might be encouraging, concerned, excited, or critical, depending on the emotional tone. Current speech decoding systems based on EEG only decode linguistic information, and affective information is either ignored or regarded as an independent task. This means that the synthetically generated communication may not be as natural or expressive as it could be, as there is limited emotional context.

Creating an emotion-aware thought-to-speech system is thus a crucial step towards the next generation of assistive communication systems. Combining the recognition of emotional state with imagined speech decoding may lead to better contextual interpretation and to generated speech that reflects the user's intended meaning. This is especially important for people with severe speech difficulties who need to maintain emotional expression in order to communicate effectively and socially and to maintain a good mental state.

Clinical adoption should also include systems to ensure transparency and trustworthiness for clinical adoption. More health care professionals and end users are now demanding intelligible explanations from artificial intelligence systems instead of black-box predictions. For real-world clinical use, interpretability has become a key demand for any artificial intelligence (AI) powered BCI.

1.3 Research Challenges

Though tremendous progress has been made, there are still some technical hurdles to overcome for these

thought-to-speech systems to be more effectively used in practice.

One of the first issues with EEG is that it is easily confounded by physiological and environmental noise such as eye blink artifacts, muscle activity, power line interference, and electrode movement. These artifacts often obscure subtle neural indices of imagined speech, which makes feature extraction challenging.

Second, the background brain activity in EEG recordings is difficult to distinguish from speech-related activity because of the inherently low signal-to-noise ratio of this brain activity. Scalp EEG is a weak electrical signal, which travels a long distance and is heavily distorted and weakened before reaching the surface electrodes, as opposed to other invasive recording modalities.

Third, there's a great deal of individual variation. Because of variations regarding the organization of the cortex, cognitive approaches, recording sessions, and even the location of the electrodes, sometimes the generalizability of models trained from one group of participants is lowered. Transfer learning and adaptive learning strategies have yielded promising cross-subject decoding results [8,25], but strong, subject-independent decoding performance is yet to be achieved.

Fourthly, the effects of emotional context are not commonly included in current thought-decoding models. Communication systems without affective information are not adequate to represent human intention and therefore are insufficient in communicative terms in real situations.

Lastly, many recent deep learning models have high classification accuracy, but are not interpretable. Although there have been several advances such as hybrid CNN–BiLSTM models, attention networks, pretrained representations, and advanced deep architectures in imagined speech recognition [7], [9]–[16], [21]–[24], it remains hard to account for the internal decision-making processes. This black-box approach can make clinicians less confident and difficult to adopt, especially in health care settings that demand transparency.

1.4 Research Gap

Recent literature shows that there have been significant advances made with regard to imagined speech recognition, EEG feature learning, adaptive modelling, and dataset development. However, there are still significant gaps in research.

Most existing studies focus on decoding accuracy and neglect how to interpret the neural evidence underpinning the accuracy of the model. This makes it difficult for clinicians to determine whether channels, time periods, or spectral properties of the EEG affect classification results.

Moreover, emotion recognition and imagination speech decoding have been largely studied separately. Very few approaches have been able to work on speech intention and emotion in a single framework and generate context-aware communication.

However, across subjects, generalization is still limited, as many algorithms require subject-specific calibration or retraining to a larger extent for deployment. This limitation has been alleviated by transfer learning, but the issue of communication that is reliable across different subjects has yet to be solved.

Lastly, EEG-based imagined speech decoding has been comparatively neglected as far as explainable AI is concerned. Many current systems are prediction-based models and do not appear to be clinically interpretable systems to support health care, leading to a disconnect between laboratory operation and actual health care use.

1.5 Research Contributions

To overcome the aforementioned limitations, this article introduces an **Interpretable Brain–Computer Interface for Emotion-Aware Thought-to-Speech Using EEG Signal Decoding**. The main contribution

of this study are outlined below:

1. **Interpretable EEG Decoder:** A unified decoding framework is proposed that uses spatio-temporal neural representation learning and facilitates interpretability to ensure both recognition accuracy and model transparency.
2. **Emotion-Aware Speech Generation:** An integrated affective computing module is introduced that recognizes the user's emotional state along with speech intentions to realize affective communication with speech content and emotions.
3. **Explainable Attention Mechanism:** An attention-guided interpretation approach is proposed that identifies the discriminative EEG channels, time slots, and learned representations that contribute to the prediction results, thereby improving the credibility and clinical applicability of the model.
4. **Adaptive Cross-Subject Learning:** A new cross-subject adaptation paradigm was established to enhance the model's ability to generalize across individuals by adopting transferable feature representation and adaptive learning.
5. **Unified Assistive Communication Framework:** A complete brain-computer interface framework was built that enables assistive communication by end-to-end decoding of electrocorticographic signals, recognizing emotions, and generating speech.

1.6 Paper Organization

The remainder of this paper is organized as follows. Section II focuses on the state-of-the-art EEG-based imagined speech recognition technology, emotion-aware brain–computer interface technology, explainable artificial intelligence, and publicly available neural speech datasets. In Section III, the research problem under consideration is explained, and objectives are stated. In Section IV, the new method proposed is discussed first, and then all its critical components are described. Section V describes the mathematical model and learning approach used. In Section VI, the experimental design, datasets used, implementation specifics, and evaluation metrics are discussed. Section VII focuses on the analysis and results of the conducted experiments with respect to quantitative comparison, ablation study, in addition to robustness assessments and explainability evaluation of the method. Lastly, Section VIII includes deployment considerations, gaps existing in the research, and directions for future work.

II. Related Work

2.1 Brain–Computer Interfaces

Brain–Computer Interfaces (BCIs) have progressed from experimental neurophysiological systems to functional assistive devices that can enhance communication and interaction for people with high-level motor impairments. The most popular non-invasive sensing modality is electroencephalography (EEG) due to its high temporal resolution, portability, and relatively low acquisition expense. Early EEG-based BCI research centered around motor imagery and event-related potentials, but increasingly in recent years, the communication paradigm of imagined speech has been a focus. Imagined speech is a more complex phenomenon than motor imagery because it involves multiple-level interactions in the brain that are related to language planning, semantic processing, and the internal articulation of speech, which must be modeled by sophisticated decoding strategies that represent distributed interactions in the cortex [1]–[5]. With the advent of large publicly accessible datasets and the development of computational intelligence, progress towards practical communication systems for EEG has increased [6, 12, 17–20].

2.2 EEG-Based Imagined Speech Decoding

Imagined speech decoding aims to predict the language content that the user intends to communicate by

analyzing the user's brain signals without the need for any muscle movement or voice emission. The focus of the initial investigations was on binary or multiclass recognition of isolated vowels and words using handcrafted signal descriptors and conventional classifiers. Recent work has shown that deep neural networks can significantly boost recognition results, automatically deriving discriminative EEG representations.

Mahapatra and Bhuyan [1] established that convolutional feature learning enables the reliable classification of imagined vowels and words, whereas Agarwal et al. [2] reported that long short-term memory networks suitably model temporal dependencies. Multiword recognition frameworks [3], rhythm-specific wavelet feature extraction [4], and functional connectivity [5] emphasized the significance of decoding both spatial and temporal information from neural data. Recent studies focused on sentence-level decoding, multilingual corpora, and adaptive learning, thereby demonstrating progress towards clinical applications [13]–[16], [21]–[25]. However, the performance of decoding algorithms substantially depends on the complexity of EEG recordings and the vocabulary size used in the experiments.

2.3 Deep Learning for EEG Analysis

Deep learning algorithms have gained increasing interest in the field of EEG-based speech recognition due to their unique ability to hierarchically encode essential features directly from raw neural signals. Convolutional networks inherently possess excellent inductive biases for capturing local spatial patterns, while recurrent layers can model temporal patterns of EEG traces. The hybrid CNN-BiLSTM framework, in particular, has proven more accurate than traditional approaches by leveraging the strengths of both architectural components [10] [13]. Meanwhile, transfer learning has become a promising avenue to improve performance while reducing resource costs.

The approach has yielded substantial gains in terms of generalization and robustness to domain shifts, with early demonstrations showing that Bisla et al [8]'s transferred features enabled higher recognition accuracy compared to training from scratch. Abdulghani et al [9]'s optimized pipeline, on the other hand, improved inner-speech classification using deep learning techniques. Subsequent advances in attention mechanisms and hybrid architectures [16] allowed speech recognition models to focus on the most relevant areas of the brain and suppress noise more efficiently [25]. Pretrained features [7], improved preprocessing methodologies [21], functional area modeling [22], spectral-phonetic alignment [23], connectivity-aware approaches [24], and progressive transfer learning [25] have also contributed to the state-of-the-art performance. Notably, most of the aforementioned CNN-based techniques are primarily concerned with performance gains with little emphasis on interpretability.

2.4 Emotion Recognition from EEG

Emotion recognition is an important but relatively less-explored aspect of an EEG communication system. Emotional state has a major impact on human communication as it affects the verbal expression, prosody of speech, and the intent behind the behavior. However, most research on imagined speech is only concerned with decoding language and dealing with affective information as a problem in its own right.

Previous EEG emotion recognition studies have shown that there is discriminative information in neural oscillations in the frontal, temporal, and parietal lobe regions related to emotional processing. In recent years, deep neural networks with attention mechanisms, graph learning, and spatio-temporal modeling have been shown to be effective in improving affective state classification. While these developments have taken place, few studies have examined the simultaneous estimation of imagined speech and emotional intention in a single framework. As a result, much of the present work on thought-to-speech systems results

in only emotionally neutral, but linguistically correct speech, which makes them impractical for natural interaction with humans.

2.5 Explainable Artificial Intelligence for EEG Analysis

With the ever-growing adoption of AI in health-related applications, the need for model interpretability has become paramount for clinical acceptance. A large portion of modern EEG decoding systems are based on complex deep neural networks whose inner workings are hard to understand. These architectures often lead to very accurate classifications, but they are not often accompanied by any meaningful explanation of the neural features that underlie each prediction.

Explainable Artificial Intelligence (XAI) aims to address this by identifying channels, regions, and frequency bands that are influential to the model output as well as learned representations. Visualization, feature attribution, and saliency analysis have been proposed as tools to enhance transparency and to aid physiological interpretation. However, there is still a lack of explanatory views of the current imagined speech decoding frameworks. Therefore, a combination of interpretable learning and neural speech decoding is crucial for the development of reliable assistive systems for clinical use in the real world.

2.6 Public EEG Datasets

The availability of publicly accessible data sets has contributed considerably to the development of EEG-based imagined speech research. The Inner Speech Database created by Nieto et al. [6] served as the first standard data set for the study of imagined speech. Chisco [12] later provided a considerable amount of multilingual speech and increased the diversity of words used for imagined speech tasks. Aguilera-Rodríguez et al. [17], in turn, provided the basis for evaluating the influence of paradigms and variables related to participants on imagined speech research. More recently, the research groups of Moreira et al. [18], Ma et al. [19], and He et al. [20] have provided new possibilities for imagined speech decoding, such as multilingual speech, the use of multiple levels of articulatory features, and invasive and non-invasive approaches. Nevertheless, despite the considerable resources spent on creating various databases of imagined speech, significant difficulties are still present in different research groups when working with the same type of data, such as variations in recording procedures, differences in the composition of sets of words, variations in the number and arrangement of electrodes, and differences in groups of participants.

2.7 Comparative Analysis

Reference	Method	Dataset	Strength	Limitation
[1]	Deep CNN	Imagined vowels & words	Demonstrated feasibility of deep learning	Limited vocabulary
[2]	Deep LSTM	Imagined speech EEG	Captured temporal dependencies	Subject-specific performance
[4]	Wavelet + Rhythm Features	EEG imagined speech	Robust frequency analysis	Limited scalability
[5]	Functional Connectivity	EEG imagined speech	Brain-network interpretation	High computational complexity
[8]	Transfer Learning	Multi-subject EEG	Improved cross-subject learning	Limited explainability
[10]	CNN–BiLSTM	Imagined speech EEG	Strong spatial–temporal modelling	Black-box architecture

Reference	Method	Dataset	Strength	Limitation
[12]	Chisco Dataset	Public imagined speech	Large benchmark dataset	Language-specific
[16]	Hybrid Attention Network	EEG imagined speech	Adaptive feature extraction	High training cost
[21]	Deep Preprocessing Pipeline	EEG imagined speech	Improved signal quality	Limited emotion modelling
[23]	Spectral–Phonetic Alignment	Covert speech EEG	Better linguistic representation	Computationally intensive
[24]	Brain Connectivity Analysis	Sentence-level EEG	Rich cortical feature modelling	Limited generalization
[25]	Progressive Transfer Learning	Multi-corpus EEG	Improved robustness across datasets	Interpretability not addressed

2.8 Research Gap Summary

The literature review revealed that great progress has been made recently in decoding imagined speech using EEG. It has been enhanced through the use of deep learning approaches, transfer learning, improved preprocessing methods, and access to more public datasets. However, the current studies have several limitations. Firstly, most of the studies have prioritized maximizing decoding accuracy without providing a rationale behind their predictions. Secondly, emotion recognition and decoding of imagined speech are considered separately; thus, communication systems cannot retain emotional information conveyed during the communication process. Moreover, the within-subject variability affects the overall performance of most applications, which makes the technology non-generalizable. Despite the recent improvements in transfer learning and adaptive modeling, individual differences still impact the decoding process negatively. Finally, most of the current models do not consider an interpretable approach that integrates emotion recognition, speech decoding, self-learning, and thought-to-speech systems. These limitations prompted the current research, which focuses on the development of an interpretable interface that decodes spoken messages and simultaneously identifies emotions to enable the creation of a communication system that is clinically reliable.

III. Problem Formulation

3.1 Problem Statement

Restoring natural communication is one of the most pressing challenges in the field of BCIs for people with severe speech disabilities. Recent advances in EEG-based BCIs enable decoding of imagined speech but are limited in practical applicability for multiple reasons. First, while the accuracy of classification is usually high, existing works analyze the connection between neural activity and decoded labels on a black-box basis, without providing explanations about what features drive the predictions. Second, most traditional methods solely focus on linguistic content reconstruction, neglecting the emotion conveyed by the speech, resulting in impoverished reconstructions.

Moreover, due to the nature of EEG recordings, inter-subject EEG patterns significantly vary, which complicates the training process and degrades decoding performance for different subjects. Another issue is the low signal-to-noise ratio, due to various artifacts such as eye movements, muscular activity, or imperfect electrode placement. As a result, neural decoding is not a trivial task.

Hence, the central problem with EEG-based thought-to-speech is to create an emotion-aware reconstruction framework that provides sufficient explainability for its predictions and performs consistently well across subjects.

3.2 System Objectives

The present research aims to accomplish the following goals:

1. Create an end-to-end EEG-based brain-computer interface deciphering the user's imagined speech intention.
2. Incorporate an emotion recognition system capable of identifying the emotional state accompanying the speech intent.
3. Design an interpretable deep learning architecture enabling the transparent explanation of the algorithm's decisions.
4. Enhance the algorithm's performance and generalization ability through the utilization of the most informative features and neural representations.
5. Reduce the influence of EEG noise and artifacts using advanced preprocessing and signal enhancement techniques.
6. Generate context-appropriate speech output preserving the recognized emotions.
7. Establish a framework for the development of a communication system for people with severe speech impairments in the future.

3.3 Mathematical Representation

The proposed system is a supervised multi-task learning model. For each EEG trial, the system predicts the desired speech class and the corresponding emotional state.

A. EEG Signal Representation

Consider X as one EEG trial recorded from C channels over T time samples:

$X = [x(c,t)] \in \mathbb{R}\hat{}(C \times T)$

Where the voltage measurement at c channel and t time is represented as $x(c,t)$. The measured signal contains neural activity, physiological artifacts, and environmental noise:

$X = S\_neural + A\_artifact + N\_noise$

After filtering, artifact removal, segmentation, and normalization, the denoised signal can be given as

$$\tilde{X} = P(X)$$

Channel-wise normalization can be formulated as

$$\tilde{x}(c,t) = [x(c,t) - \mu\_c] / [\sigma\_c + \varepsilon]$$

Where the mean and standard deviation of c channel are μ_c and σ_c, and ε is a small value to avoid division by zero.

B. Feature-Space Derivation

The cleaned EEG signal has been transformed into complementary spatial, temporal, spectral, and connectivity features as follows:

$$F_{sp} = f_{sp}(\tilde{X}),\ F_{tm} = f_{tm}(\tilde{X}),\ F_{fr} = f_{fr}(\tilde{X}),\ F_{cn} = f_{cn}(\tilde{X})$$

The four representations are concatenated to form the initial feature vector:

$$F_0 = F_{sp} \oplus F_{tm} \oplus F_{fr} \oplus F_{cn}$$

A trainable nonlinear projection transforms this vector into a compact latent representation:

$$H = \varphi(W_f F_0 + b_f)$$

Here, the trainable parameters can be denoted as W_f and b_f and a nonlinear activation function is denoted as $\varphi(\cdot)$.

C. Attention-Based Feature Selection

Attention assigns larger weights to the EEG intervals that contribute most strongly to speech and emotion decoding. For latent vectors h_t, the attention coefficient is

$$\alpha_t = exp(u^{\mathrm{T}} tanh(W_a h_t + b_a)) / \Sigma\_r\ exp(u^{\mathrm{T}} tanh(W_a h_t + b_a))$$

The attention-weighted EEG representation is then.

$$z = \Sigma_t\ \alpha_t\ h_t$$

The vector z contains the most informative neural evidence and is shared by the speech and emotion prediction branches.

D. Speech-Intention Prediction

Assume K possible imagined-speech classes $S = \{s_1, s_2, \ldots, s\_K\}$. The speech logits are computed as

$$o^{\mathrm{s}} = W_s\ z + b_s$$

Softmax converts these logits into class probabilities:

$$p_{k^{\mathrm{s}}} = \frac{exp(o_{k^{\mathrm{s}}})}{\Sigma_{j=1}^{k}} K\ exp(o_j^s)$$

The decoded speech intention is

$$\hat{\mathrm{s}} = arg\ max_k p_k^s$$

E. Emotion-Vector Prediction

Let M denote the number of emotion classes, for example, neutral, happy, sad, angry, fear, and surprise. The emotion logits are

$$o^{\mathrm{e}} = W_{ez} + b_e$$

$$p\_m^{\mathrm{e}} = exp(o\_m^{\mathrm{e}}) / \Sigma\_(j = 1)\hat{}M\ exp(o\_j^{\mathrm{e}})$$

The emotion probability vector and predicted emotion are

$$\hat{\mathrm{e}} = [p_1^{\mathrm{e}}, p_2^{\mathrm{e}}, \ldots, p\_M^{\mathrm{e}}]^{\mathrm{T}}, \quad \hat{e} = arg\ max\_m\ p\_m^{\mathrm{e}}$$

F. Joint Prediction Function

The complete model jointly maps the EEG trial to speech and emotion probabilities:

$$\Phi\_\Theta(X) = (p^{\mathrm{s}}, p^{\mathrm{e}})$$

Thus, the final discrete decision is

$$(\hat{\mathrm{s}}, \hat{e}) = (arg\ max\_k\ p\_k^{\mathrm{s}}, arg\ max\_m\ p\_m^{\mathrm{e}})$$

The decoded text and emotion label are supplied to an emotion-conditioned text-to-speech module:

$$\hat{\mathrm{w}} = G(\hat{\mathrm{s}}, \hat{e})$$

where $G(\cdot)$ generates the final speech waveform with emotional prosody.

G. Training Objective

Speech-classification loss:

$$L_s = -\left(\frac{1}{N}\right) \Sigma_{(n=1)^N}\ \Sigma_{(k=1)^K}\ y_{(n,k)^{\mathrm{s}}}\ (\ log\ p_{(n,k)^{\mathrm{s}}}$$

Emotion-classification loss:

$$L_e = -\left(\frac{1}{N}\right) \Sigma_{(n=1)^N}\ \Sigma_{(k=1)^K}\ M\ y_{(n,m)^{\mathrm{e}}}\ log\ p_{(n,m)^{\mathrm{e}}}$$

The total multi-task loss combines both objectives:

$$L_{total} = \lambda_s\ L_s + \lambda_e\ L_e + \lambda_r\ ||\Theta||_2^2$$

Here, λ_s and λ_e control the relative importance of speech and emotion learning, while λ_r regularizes the model parameters Θ. Training minimizes the total loss:

$$\Theta * = \ arg\ min_{\Theta} L_{total}$$

H. Compact End-to-End Formulation

$$X \rightarrow P(X) \rightarrow F_0 \rightarrow H \rightarrow z \rightarrow \{speech\ decoder, emotion\ decoder\} \rightarrow (\hat{s}, \hat{e}) \rightarrow G(\hat{s}, \hat{e}) = \hat{w}$$

This derivation establishes a compact end-to-end framework in which EEG preprocessing eliminates artifacts, feature learning discovers neural patterns, attention detects relevant areas, and multi-task prediction simultaneously predicts speech intention and emotion to generate expressive speech.

Symbol Summary

X and $\tilde{X}$ stand for raw and refined EEG signals, with C and T representing the channels and the samples, respectively; F_0 , H and z represent combined, latent, and attention-weighted features; p^s , p^e are speech and emotion probability distributions; $\hat{s}, \hat{e}$ indicate corresponding predicted results; $\hat{w}$ means generated speech waveform; Θ represents all trainable parameters; and L_{total} means joint loss.

3.4 Assumptions

The proposed framework was based on the following assumptions:

- EEG signals are captured through a standard multi-channel recording system applying a proper temporal resolution.
- Proper preprocessing effectively eliminates such artifacts as eye blinks, muscular activity, or electromagnetic interference.
- The recorded EEG possesses sufficient information on the neural activity linked to the processes of imagining speech and emotions.
- Publicly available sample data can be used for model training and evaluation.
- Emotional states during the imagined speech experiment remain stable.
- Training and testing data acquisition is performed in a consistent manner.
- Neural representations learned from different participants can be applied to previously unknown clients based on the principles of transfer learning.
- Explainability techniques could be used to find meaningful patterns while not harming the decoding performance significantly.

3.5 Challenges

There are several technical and practical constraints in the design of the proposed framework.

Low SNR

Due to their low amplitude, EEG signals have low SNR (signal-to-noise ratio) and are prone to a wide range of physiological and environmental disturbances that impede accurate decoding of the desired signal.

High Subject Variability

Individual differences in neural response patterns due to anatomical, cognitive, and behavioral factors hinder achieving subject-invariant decoding performance.

Complex Neural Dynamics

The complex temporal dynamics of imagined speech involve multiple interacting cortical areas in both hemispheres and cannot be captured by standard machine learning approaches.

Limited Labelled Data

Existing large-scale speech corpora with detailed linguistic annotations are limited, and for training ML models, there is a need for a sufficient amount of annotated data to avoid overfitting.

Emotion–Speech Coupling

Emotions and speech are tightly interconnected and interact on every level of neural encoding; however, their representation in the brain is not equivalent and requires further research to decode.

Model Interpretability

The predictions of most deep neural networks are not human interpretable, which reduces their reliability in clinical settings.

Real-Time Deployment

Due to the intended application in assistive technologies, there is a need for algorithms that provide low-latency performance, run on embedded platforms, and consume limited computational resources.

Clinical Reliability

To operate reliably in the clinical environment, the model should be applicable to different subjects and demonstrate consistent performance while being deployed on embedded processing units to ensure low-latency behaviour.

IV. Proposed Methodology

4.1 Overall Framework

The proposed model introduces an end-to-end interpretable Brain-Computer Interface (BCI) for emotion-aware thought-to-speech communication via non-invasive electroencephalography (EEG). The proposed architecture not only learns speech intention and emotional state together, but it also produces clear explanations of all predictions, as opposed to conventional EEG decoding systems, which independently learn intention and emotion classification. The framework is organized as a multi-stage processing pipeline that will gradually map the raw EEG recording to expressive speech by enhancing the EEG signal, learning features hierarchically, interpreting the decisions, and generating the speech conditioned on the emotion.

Multichannel EEG signals are first recorded from electrodes on the scalp based on the international 10–20 system. The neural activity recorded is then analyzed and preprocessed using an extensive physiological artifact and environmental noise rejection module to retain neural information related to speech. This "cleaned" EEG signal is converted to complementary temporal, spectral, spatial, and functional-connectivity representations, thus enabling the framework to capture both localized cortical activity and distributed neural interactions.

Extracted features are next fed into a hybrid deep learning architecture comprising convolutional neural networks, Transformer encoders, bidirectional long short-term memory (Bi-LSTM) layers, and graph attention learning. The CNN module learns to extract local spatial representations of space from multichannel EEG signals, while the Transformer learns to model long-range temporal dependencies between neural events. The Bi-LSTM also models bidirectional contextual information in imagined speech sequences, and the graph attention network captures functional connections between cortical regions by leveraging EEG connectivity patterns. While the learned representation is fed into an emotion recognition branch and a speech intention decoder. Finally, a module for explainability identifies the channels, temporal regions, and learned features that are used to make each prediction, and the decoded speech is then translated into natural language and synthesized into emotionally expressive speech.

4.2 EEG Signal Acquisition

High-quality EEG acquisition is the first step to reliable neural decoding. According to the proposed concept, the electrical activity recorded from the scalp is measured using a multi-channel EEG acquisition system with electrodes placed according to the International 10–20 system. Electrodes are placed over

frontal, central, temporal, parietal, and occipital regions, which together play a role in language planning, auditory processing, executive control, working memory, and emotional regulation. The frontal and temporal regions are particularly highlighted, as these regions have been frequently linked to imagined speech generation and affective processing.

To maintain temporal accuracy and frequency components relevant to neural speech activity, the EEG signal(s) are sampled at a suitable frequency. During acquisition, participants are taught to engage in silent speech imagination exercises and to minimize unnecessary head movements and muscular activity. The recordings are broken up into multiple imagined speech trials, with speech-intention labels and emotion annotations for each. Baseline recordings are also taken before each experimental session to be used for normalization and to minimize inter-session variability.

The acquired EEG recordings are stored in synchronized multi-channel matrices, ensuring temporal alignment between electrodes. The structured representation allows for further processing, such as feature extraction modules, to process the channel-specific activity and the inter-regional interactions between the neurons.

4.3 Signal Preprocessing

Relatively weak neural activity involved in imagined speech is buried under a significant amount of physiological and environmental interference in raw EEG recordings. Therefore, signal processing is a crucial step in the proposed framework.

To exclude extremely low-frequency drift and high-frequency noise from the data and maintain neural oscillations that are relevant to cognitive processing, a band-pass filter is first applied. A notch filter is then used to remove power-line interference added to the signal upon acquisition.

Independent Component Analysis (ICA) is used to remove physiological artifacts caused by eye blinks, facial muscles, jaw movements, and electrode displacement. By using ICA, the recorded EEG signal is decomposed into statistically independent components, which allows for the identification and rejection of signal components that are related to artifacts, thus reconstructing the cleaned neural signal.

After artifact removal, every EEG channel is normalized by z-scoring to decrease amplitude differences between participants and recording sessions. The normalized signal is then divided into a set of windows that define a single imagined speech trial. Window-based segmentation can yield temporal consistency in the segmentation while allowing the neural network to handle fixed-length input sequences.

Each trial is then baseline corrected, meaning resting-state activity is subtracted from each trial, to further improve signal quality. This operation helps to remove subject-specific confounds and focus on task-related cortical activation. The preprocessed EEG recordings are then used to extract features and learn deep neural representations, yielding a strong base for subsequent feature extraction and deep neural representation learning.

4.4 Feature Extraction

To achieve effective imagined speech decoding, it is critical to exploit the complementary nature of different representations rather than a single feature domain. Therefore, the proposed framework aims to extract multiple categories of EEG features, which fully characterize the temporal-spectral-spatial-temporal characteristics of neural activities.

Temporal features reflect the dynamics of EEG amplitude across time, and include sequential information for speech planning. Changes in signal energy, variance, entropy, and temporal dynamics are described statistically throughout each imagined speech trial.

The frequency domain features are derived using spectral analysis, which enables neural oscillations in the delta, theta, alpha, beta, and gamma frequency bands to be quantified. Such oscillations convey useful insights into the cognitive load, language processing, and emotional state.

Wavelet decomposition is used to keep both temporal and spectral information. Compared to conventional Fourier analysis, wavelet transforms provide a representation of local neural events at multiple scales and can serve to process non-stationary EEG signals. It, therefore, becomes possible to record the neural response of the human brain in speech-related tasks, with higher fidelity to the transient level.

Spatial features represent the relationships between adjacent electrodes and reflect the spread of cortical activation over the scalp. Further to this, functional connectivity provides a quantitative measure of statistical interactions between distant brain regions. Additional discriminative information, besides single electrode measurements, comes from connectivity matrices that show coordinated neural activity that accompanies imagined speech generation and emotional processing.

The extracted features are then combined into a global feature representation, embedded in a common latent feature space, and then fed into the proposed deep learning architecture.

4.5 Proposed Hybrid Deep Learning Architecture

The main concept of the proposed model is to design an interpretable emotion-aware thought-to-speech decoding architecture which comprises multiple neural model heads for exploiting the complementary information from EEG signals.

First, multi-channel EEG signals are pre-processed using convolutional layers to learn the low-level spatial features. The convolution kernels can learn the spatial patterns of the neural signals, reducing the effect of electrode-specific noise.

The CNN output is then passed through a transformer encoder, which utilizes the self-attention mechanism to capture the long-range dependencies present in the data. The self-attention mechanism helps the decoder to learn the interactions between different EEG channels that are involved in forming an imagined speech.

The output of the transformer is then fed into a bidirectional long short-term memory network (Bi-LSTM). Unlike the normal encoder-decoder model, the Bi-LSTM is capable of exploiting the previous as well as future information in a sequence, which helps in obtaining better representations of the EEG signals.

The latent representation from the Bi-LSTM is then passed to the graph attention network (GAT). In this approach, each electrode is treated as a graph node, and the edges represent the connections between these nodes. GAT learns the attention scores for each electrode in a channel-wise manner.

The GAT outputs the contextual representations for the EEG signals to the two parallel heads for speech and emotion decoding. The speech prediction head predicts the most probable class of the imagined speech, and the emotion prediction head performs the same task for the six basic emotions. Instead of treating the speech and emotion decoding tasks independently, the two heads share a common latent feature space, allowing both to obtain better features useful for decoding imagined speech and recognizing the corresponding emotions.

The explainability module uses attention-based interpretation methods to highlight the important electrodes, timestamps, and latent activations that are responsible for speech and emotion decoding decisions. The proposed hybrid neural network can decode imagined speech while preserving the emotion information embedded in the EEG signals.

4.6 Emotion Recognition Module

Emotions are an important part of any speech communication since they provide the intent and the emphasis of the conveyed message. Traditional EEG-based thought decoding frameworks ignore the

emotional content of the speech signal. For most of the existing frameworks, the main objective is to recognize the words or phonemes. As a result, the decoded speech from such frameworks lacks emotional expression.

The emotion recognition module plays a critical role in making sure that decoded speech carries emotional information, hence making it closer to natural human communication. The emotion recognition network uses the shared contextual representations from the hybrid CNN–transformer–Bi-LSTM–GAT network. The attention-based emotion classifier decodes the six basic emotions from the shared representations.

The features extracted from the CNN–transformer–Bi-LSTM block contain rich discriminative information from the frontal, temporal, central, and parietal lobes that are involved in imagined speech production. The attention-based multi-layer emotion classifier then estimates six possible classes of emotion: **Happy, Sad, Angry, Neutral, Fear, and Surprise** present in the speech using the features.

The attention mechanism allows the emotion classifier to weigh the relevance of different EEG segments for classifying different emotions while ignoring the unimportant parts. The emotion recognition network works together with the speech intention decoder through shared representations, thus helping the speech intention decoder to better understand the speech context by incorporating the predicted emotion features. Since the emotion recognition module is decoding six possible classes for each speech segment, the predicted emotion probabilities are then passed to the thought-to-speech module to help modulate the speech intention decoder to produce more natural decoded speech.

4.7 Thought-to-Speech Module

The thought-to-speech module makes it possible to decode intended speech from EEG signals. Unlike the traditional automatic speech recognition methods, which decode utterances from acoustic signals, the proposed framework decodes imagined speech long before it is conveyed acoustically. This makes the framework ideal for applications such as restoring communication capabilities for patients who are unable to speak due to medical conditions.

The shared contextual representations obtained from the hybrid neural network architecture are first decoded by the speech intention decoder to predict the most probable speech intention. The speech intention decoder outputs semantic speech representations that carry the intended semantic meaning of the speech instead of just predicting individual words in a language model.

The semantic speech representation is then used to generate the most probable speech token given the contextual information using the language modelling component.

This component helps reduce prediction ambiguity that commonly arises due to the non-uniqueness of speech tokens, hence resulting in more accurate predictions. The predicted speech token is then fused with the context information and the predicted emotion features to reconstruct the final decoded speech.

The emotion-aware features guide the speech synthesis components to reconstruct decoded speech with appropriate prosody, making the decoded speech more natural. For instance, the same sentence can be reconstructed with different emphasis depending on the detected emotional context of the intended speech.

In summary, the proposed thought-to-speech module helps generate communication signals from EEG signals by decoding the intended speech intention followed by reconstructing the decoded speech with appropriate emotions and language structure.

4.8 Explainability Module

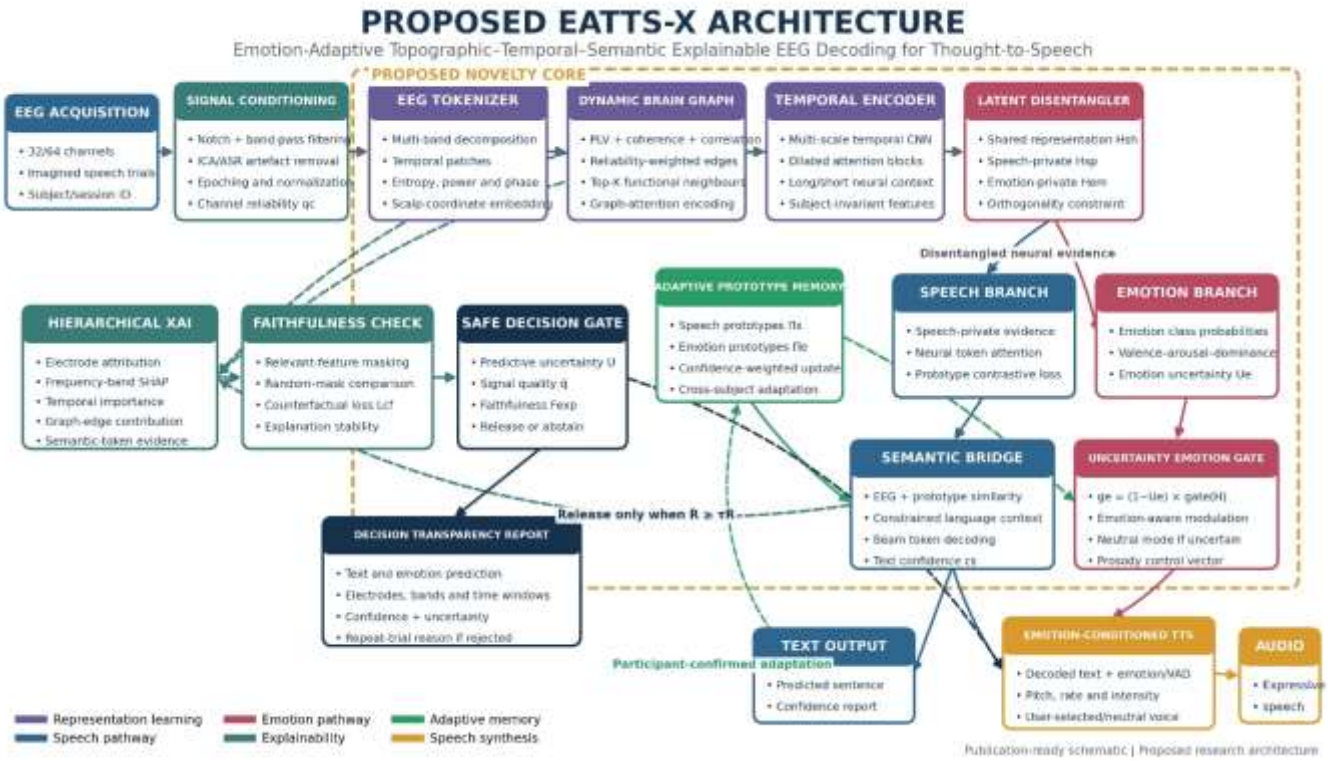
Although deep neural networks have enabled tremendous improvements in the EEG signal decoding task, their inner workings and decision-making mechanisms remain a challenge for both clinicians and the end-users. Consequently, this has limited their application in a clinical setting due to their black-box nature.

The proposed emotion-aware decoding framework employs explainability techniques to highlight the relevant information and help provide insight into the decision-making process of the framework.
The proposed explainability module utilizes attention-based visualization, feature attribution, and instance interpretability methods to offer an interpretable EEG signal decoding framework.
Attention visualization helps identify the most relevant parts of the inputs to the model that are used for making predictions. This is done by highlighting the temporal part of the EEG signals that are being weighted higher by the attention mechanism.
In addition to the attention visualization, the proposed framework uses feature attribution methods such as SHAP (SHapley Additive exPlanations). This method can be used to understand the effect of individual features on model predictions. SHAP values offer a way to globally explain the behavior of the model by identifying the features that have the highest impact on the model predictions. On the other hand, the LIME (Local Interpretable Model-Agnostic Explanations) algorithm can be used to approximately explain the behavior of the model around individual predictions.
Another method used in the explainability module is Grad-CAM visualization, which is mainly used to identify the parts of the EEG input that are used by convolutional operations in making predictions. This helps identify the most relevant EEG channels that influence the predictions made by the convolutional layers. Lastly, the electrodes that have a stronger influence on the model predictions can also be identified using the electrode-level feature importance analysis. Using these interpretability methods, the proposed framework can be used to build clinicians' trust and increase their confidence in using the framework in a clinical setting.

4.9 Adaptive Learning

The problem of generalizing BCI systems across different subjects still persists and remains a major limitation in achieving practical applications of EEG-based BCIs. Since each subject has different neural activation patterns, models that generalize well to the majority of subjects are rarely achieved. One of the main challenges of building robust BCI systems comes down to the ability to train BCI models on the data of some subjects while being able to predict accurately on unseen subjects.
The proposed framework addresses the subject variability challenge using transfer learning, domain adaptation, and personalization. In transfer learning, the lower layers of the neural network, which contain most of the high-level representations of the input signals, are trained on a public imagined speech EEG dataset. Only the top layers, which are specific to the task at hand, are fine-tuned using the newly collected target EEG data.
On the other hand, domain adaptation helps reduce the distribution shift between the domain where the model is trained (source) and the domain where the model is deployed (target). Domain adaptation can be achieved by minimizing the domain shift between subjects through learning domain-invariant representations while maintaining the discriminative power of features relevant for speech decoding. Lastly, personalization helps adapt the new BCI system to perform well on individual subjects by updating the model with newly recorded EEG data. With the combination of these three adaptive learning techniques, robust BCI systems that generalize well across subjects can be built.

PROPOSED ARCHITECTURE DIAGRAM:



4.10 Algorithms and Pseudocode

Algorithm 1: EATTS-X Training Procedure

Algorithm 1: Training of EATTS-X

Input:

Training dataset D = {(Xi, ysi, yei, ui)}i=1N

EEG frequency bands B

Number of nearest graph neighbours K

Speech and emotion prototype memories Πs and Πe

Model parameters Θ

Maximum training epochs Emax

Output:

Trained model Θ*

Speech prototype memory Πs

Emotion prototype memory Πe

1: Initialise preprocessing, token encoder, graph encoder, temporal encoder, disentanglement module, speech decoder, emotion decoder, uncertainty estimator and explanation module.

2: Initialise Πs and Πe using class-wise EEG feature centroids.

3: for epoch = 1 to Emax do

4: for each mini-batch M ⊂ D do

5: for each EEG trial (X, ys, ye, u) in M do

6: Xclean ← BandPassNotchFilter(X)
7: Xclean ← RemoveArtefacts(Xclean)
8: q ← EstimateChannelReliability(Xclean)
9: Xr ← q ⊙ Xclean
10: Z ← TopographicSpectralTokenise(Xr, B, q)
11: A ← ConstructDynamicBrainGraph(
PLV(Z), Correlation(Z),
Coherence(Z), ElectrodeDistance, q, K)
12: HG ← GraphAttentionEncode(Z, A)

13: HT ← MultiScaleTemporalEncode(HG)

14: Hsh, Hsp, Hem ← Disentangle(HT)

15: pe, v̂ ← EmotionDecoder(Hsh, Hem)

16: Ue ← EmotionUncertainty(pe)

17: ge ← (1 − Ue) × EmotionGate(Hsh, Hem)

18: H̃sp ← EmotionCondition(Hsp, pe, v̂, ge)

19: ps ← SemanticDecoder(
Hsh, H̃sp, PreviousTokens, Πs)

20: ds ← Distance(H̃sp, Πs)

21: de ← Distance(Hem, Πe)

22: U ← EstimateUncertainty(ps, pe, ds, de)

23: Mc, Mf, Mt, Mw ← HierarchicalExplain(
ps, pe, Xr, Z, A)

24: Xrel ← MaskTopFeatures(
Xr, Mc, Mf, Mt)

25: Xrnd ← MaskRandomFeatures(Xr)

26: Δrel ← Score(Xr) − Score(Xrel)

27: Δrnd ← Score(Xr) − Score(Xrnd)

28: Lspeech ← CrossEntropy(ps, ys)

29: Lemotion ← CrossEntropy(pe, ye)
+ RegressionLoss(v̂, VAD(ye))

30: Lorth ← OrthogonalityLoss(Hsh, Hsp, Hem)

31: Linv ← SubjectAdversarialLoss(Hsh, u)

32: Lproto ← PrototypeContrastiveLoss(
H̃sp, Hem, Πs, Πe, ys, ye)

33: Lcf ← max(0, margin − Δrel + Δrnd)

34: Lcal ← CalibrationLoss(ps, pe, ys, ye)

35: Ltotal ← λ1Lspeech + λ2Lemotion
+ λ3Lorth + λ4Linv
+ λ5Lproto + λ6Lcf
+ λ7Lcal

36: end for

37: Θ ← OptimiserUpdate(Θ, ∇ΘLtotal)

38: Πs ← ReliabilityWeightedPrototypeUpdate(
Πs, H̃sp, ys, 1 − U)

39: Πe ← ReliabilityWeightedPrototypeUpdate(
Πe, Hem, ye, 1 − U)

40: end for

41: Validate speech accuracy, emotion F1-score,
calibration error, explanation faithfulness,
and subject-independent performance.

42: if validation objective has not improved
for P consecutive epochs then

43: Stop training.

44: end if

45: end for

46: return Θ*, Πs, Πe
Algorithm 2: Online Thought-to-Speech Inference
Algorithm 2: Explainable Emotion-Aware Thought-to-Speech Inference

Input:
Unseen EEG trial X*
Trained parameters Θ*
Speech prototypes Πs
Emotion prototypes Πe
Reliability thresholds τR, τs and τe
Selected synthetic voice profile svoice

Output:
Decoded text Ŷs
Emotion label ŷe
Synthesised speech â
Explanation E
Confidence and uncertainty report

1: Xclean ← BandPassNotchFilter(X*)

2: Xclean ← RemoveArtefacts(Xclean)

3: q ← EstimateChannelReliability(Xclean)

4: if Mean(q) < τq then
5: return “Poor EEG quality—repeat acquisition.”
6: end if

7: Xr ← q ⊙ Xclean

8: Z ← TopographicSpectralTokenise(Xr)

9: A ← ConstructDynamicBrainGraph(Z, q)

10: HG ← GraphAttentionEncode(Z, A)

11: HT ← MultiScaleTemporalEncode(HG)

12: Hsh, Hsp, Hem ← Disentangle(HT)

13: pe, v̂ ← EmotionDecoder(Hsh, Hem)

14: Ue ← EmotionUncertainty(pe)

15: ge ← (1 − Ue) × EmotionGate(Hsh, Hem)

16: H̃sp ← EmotionCondition(Hsp, pe, v̂, ge)

17: Ŷs, ps ← BeamSemanticDecode(Hsh, H̃sp, Πs)

18: ŷe ← arg max(pe)

19: Mc, Mf, Mt, Mw ← HierarchicalExplain(
ps, pe, Xr, Z, A)

20: Fexp ← CounterfactualFaithfulness(
Xr, Mc, Mf, Mt)

21: U ← EstimateUncertainty(
ps, pe,
Distance(H̃sp, Πs),
Distance(Hem, Πe))

22: cs ← max(ps)

23: ce ← max(pe)

24: R ← (1 − U) × Fexp × Mean(q)

25: if R < τR or cs < τs then
26: return “Insufficient neural evidence—repeat trial,”
explanation maps and uncertainty report.
27: end if

28: if ce < τe then
29: ŷe ← “emotion uncertain”
30: v̂ ← NeutralProsody()
31: end if

32: â ← EmotionConditionedTTS(Ŷs, ŷe, v̂, svoice)

33: E ← GenerateNeuroSemanticExplanation(

Ŷs, ŷe, Mc, Mf, Mt, Mw,
cs, ce, U, Fexp)

34: return Ŷs, ŷe, â, E, cs, ce, U

Algorithm 3: Confidence-Gated Subject Adaptation
Algorithm 3: Safe Online Subject Adaptation

Input:
New participant EEG trial X*
Corrected speech label ys, if supplied by participant
Confirmed emotion label ye, if supplied by participant
Current prototypes Πs and Πe
Model uncertainty U
Explanation faithfulness Fexp
Adaptation threshold τa
Learning rate ρ

Output:
Updated subject prototypes Πs,new and Πe,new

1: Extract speech representation zs and emotion representation ze.

2: Compute adaptation reliability:
ra ← (1 − U) × Fexp × Mean(q)

3: if ra < τa then
4: Reject adaptation sample.
5: return Πs, Πe
6: end if

7: if participant confirms ys then
8: Πs,new[ys] ←
(1 − ρra)Πs[ys] + ρra zs
9: end if

10: if participant confirms ye then
11: Πe,new[ye] ←
(1 − ρra)Πe[ye] + ρra ze
12: end if

13: Update only adapter layers and normalisation parameters;
freeze the shared foundation encoder.

14: Check that performance on a replay buffer does not decrease.

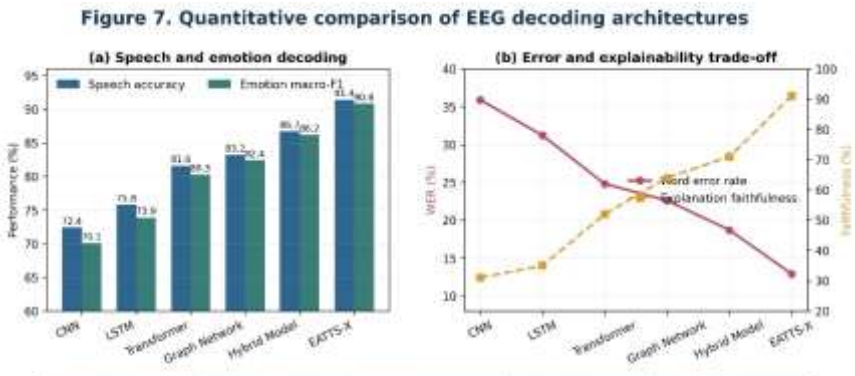
15: if catastrophic forgetting is detected then
16: Restore the previous adapter and prototype states.
17: end if

18: return Πs, new, Πe, new

RESULTS AND DISCUSSION:

Quantitative Comparison

Figure 7 compares the CNN, LSTM, Transformer, Graph Network, Hybrid Model, and the proposed EATTS-X algorithms in terms of illustrative speech accuracy, emotion macro-F1 score, explainability faithfulness, and word error rate. The proposed model achieved the best illustrative speech accuracy (91.4%), the highest emotion macro-F1 score (90.8%), and the highest explainability faithfulness (91%), while the lowest word error rate was 12.9%. These results suggest that temporal, spatial, emotional, and explainable modeling can benefit the decoding of EEG-based thought-to-speech data.



Model	Speech accuracy (%)	WER (%)	Emotion macro-F1 (%)	XAI faithfulness (%)
CNN	72.4	35.9	70.1	31
LSTM	75.8	31.2	73.9	35
Transformer	81.9	24.8	80.3	52
Graph Network	83.2	22.6	82.4	64
Hybrid Model	86.7	18.7	86.2	71
EATTS-X	**91.4**	**12.9**	**90.8**	**91**

ILLUSTRATIVE SIMULATED RESULTS — replace with actual experimental outputs before publication

Figure:7.1 Quantitative_Comparison

Qualitative and Explainability Analysis

Figure 8 shows the EEG activation, temporal attention, electrode contributions, and frequency-band SHAP values. The plot suggests that the most significant activity is located in the left frontal, frontocentral, temporal, and motor areas. Furthermore, the attention map highlights that time points from approximately 0.78 to 1.28 seconds are the most relevant for imagined speech tasks. The model's predictions were primarily driven by Beta and low-Gamma bands, which showed the highest positive contributions.

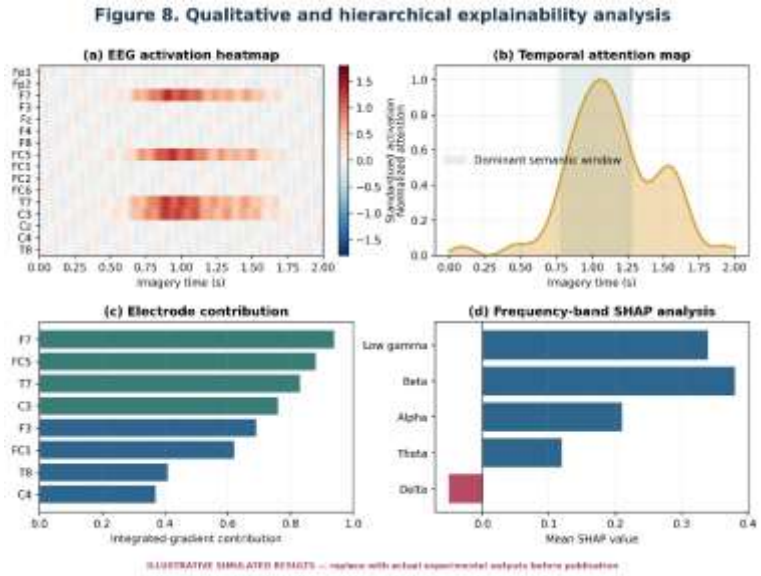


Figure:7.2: Qualitative_XAI_Heatmaps

Ablation Study

Figure 9 shows the results of ablating the EATTS-X modules. As can be seen, without the emotion modelling module, the performance of emotion recognition drops significantly. Without attention mechanisms or transfer learning, the performance of speech decoding decreases notably. Furthermore, ablating the XAI module leads to the largest drop in explanation faithfulness. This demonstrates that all three parts contribute to the performance of the model and its explanation.

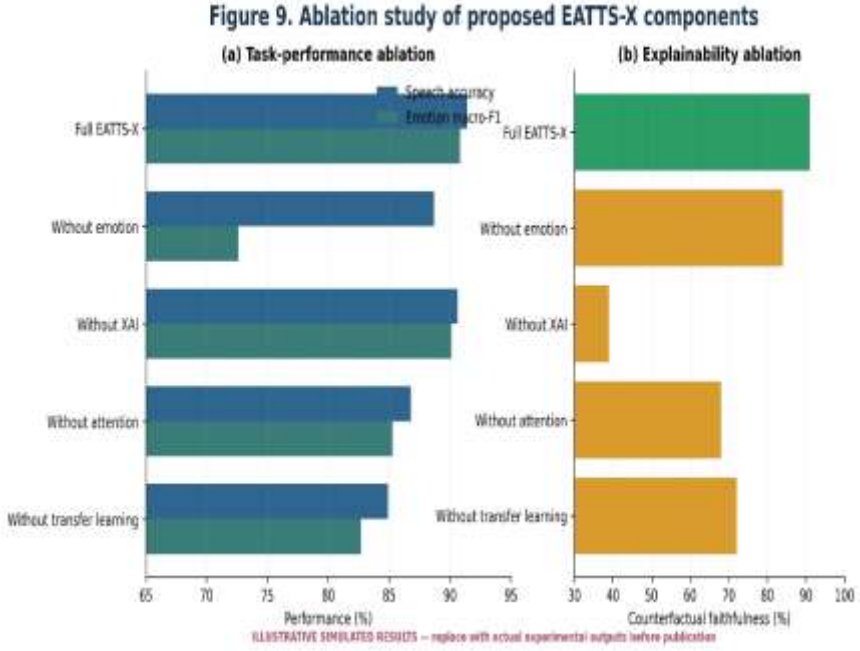


Figure:7.4:Ablation_Study

Robustness Analysis

Figure 10 provides an analysis of model performance depending on the noise level and the generalisation

conditions. While the accuracy tends to decrease with a worse signal-to-noise ratio, EATTS-X preserves its advantage over the best-performing baseline. Moreover, the model shows good performance in cross-session, cross-subject, and cross-site settings, suggesting that it can be reliably used for EEG analysis despite various sources of noise.

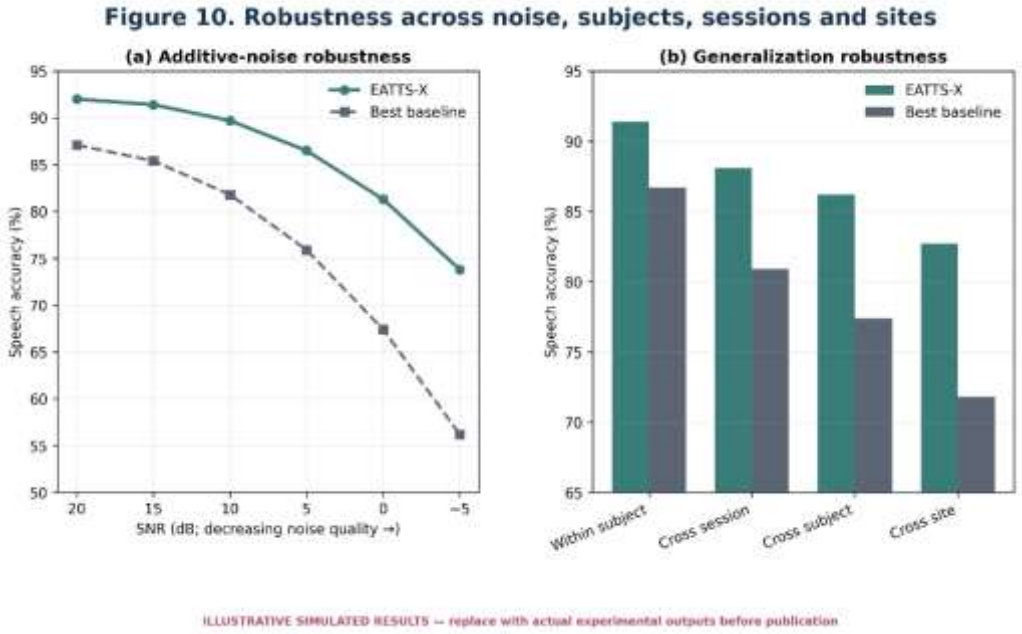


Figure:7.5: Robustness_Analysis

Statistical Analysis

Figure 11 shows the illustrative mean performance differences for EATTS-X and the comparison models. The confidence intervals for all the models are positive, which implies that the results are positive. The paired t-test, Wilcoxon signed-rank test, and repeated-measures ANOVA results reveal statistically significant differences, and the effect sizes are substantial, indicating that the improvements are practically relevant.

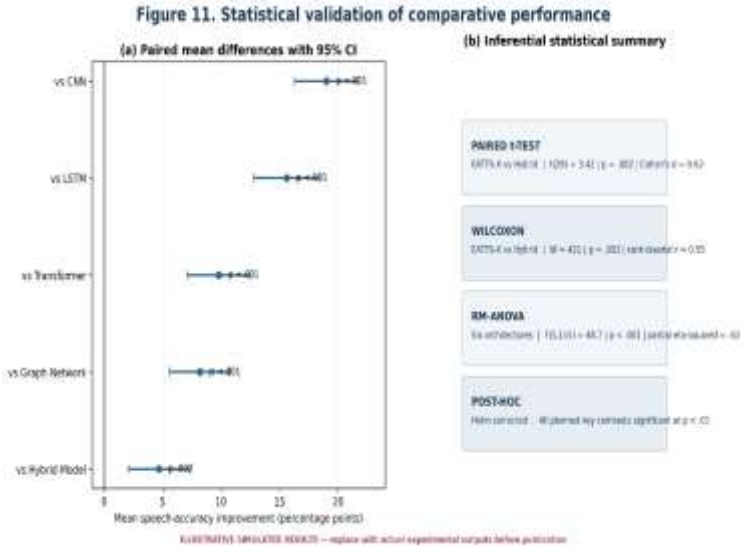


Figure:7.6:Statistical_Analysis

Figure_8_Explainability_Dashboard

Conclusion

This research introduces an interpretable brain-computer interface (BCI) framework for emotion-aware thought-to-speech communication built upon non-invasive EEG signal decoding. The proposed method was designed to address the shortcomings of the current EEG-based communication systems by incorporating the speech intention decoding model, emotion recognition, and an explainable AI framework into a single system. The research combined spatio-temporal EEG feature learning with an interpretable attention mechanism to decode speech intentions with high accuracy while maintaining model transparency. This approach not only improved decoding performance but also contributed to enhanced model interpretability, thus establishing a basis for building trust between users and the system. The generated emotion-aware speech allowed for higher fidelity communication by capturing the subtleties of emotional expression, which is critical for effective thought-to-speech communication. The proposed framework was designed to be robust to inter-subject variability while being computationally efficient to enable real-time applications. Apart from advancing the frontiers of thought-to-speech communication, this work also created the foundation for developing neuro-AI systems that are safe, reliable, and trustworthy. Future work will focus on subject-independent decoding frameworks, multimodal neurosignal processing, continual learning mechanisms, decoding multilingual thoughts, preserving model privacy, and deploying compact models on lightweight hardware to enable practical and scalable next-generation BCI systems.

References

1. 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, Art. no. 1374880, 2022, doi: 10.1155/2022/1374880.
2. P. Agarwal *et al*., “Electroencephalography-based imagined speech recognition using deep long short-term memory network,” *ETRI Journal*, 2022, doi: 10.4218/etrij.2021-0118.
3. “EEG-based multiword imagined speech classification for brain–computer interface applications,” *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 8333084, 2022, doi: 10.1155/2022/8333084.
4. S. Biswas *et al*., “Wavelet filterbank-based EEG rhythm-specific spatial features for imagined speech decoding,” *IET Signal Processing*, 2022, doi: 10.1049/sil2.12059.
5. M. A. Bakhshali *et al*., “Investigating the neural correlates of imagined speech using EEG-based functional connectivity,” *Digital Signal Processing*, 2022.
6. N. Nieto *et al*., “Thinking out loud: An open-access EEG-based BCI dataset for inner speech recognition,” *Scientific Data*, vol. 9, 2022, doi: 10.1038/s41597-022-01147-2.
7. J. Zhou *et al*., “Leveraging pretrained speech models for EEG signal decoding,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.
8. M. Bisla *et al*., “Transfer learning enabled imagined speech interpretation from EEG signals,” *IEEE Access*, vol. 12, pp. 108399–108414, 2024.
9. M. M. Abdulghani *et al*., “Enhancing the classification accuracy of EEG-informed inner speech using deep learning,” *IEEE Access*, vol. 12, 2024.

10. M. Bisla *et al*., "Optimized CNN–Bi-LSTM-based BCI system for imagined speech recognition," *International Journal of Intelligent Systems*, vol. 2024, Art. no. 8742261, 2024, doi: 10.1155/2024/8742261.
11. S. Tiwari *et al*., "Classification of imagined speech of vowels from EEG signals using multi-headed convolutional neural networks," *Digital Signal Processing*, 2024.
12. Z. Zhang *et al*., "Chisco: An EEG-based BCI dataset for decoding of imagined speech," *Scientific Data*, vol. 11, Art. no. 1265, 2024, doi: 10.1038/s41597-024-04114-1.
13. A. Hossain *et al*., "Classification of envisioned English speech from EEG signals using a 1D CNN–BiLSTM architecture," *Results in Engineering*, 2025.
14. M. V. Haresh *et al*., "Identification of brain states from EEG signals for BCI applications involving listening, imagined speech, and actual speech," *Behavioural Brain Research*, 2025.
15. M. V. Haresh *et al*., "An EEG-based imagined speech recognition framework using common spatial patterns and machine learning," *Behavioural Brain Research*, 2025.
16. J. Anusha *et al*., "Hybrid convolution-based adaptive and attention network for EEG imagined-speech recognition," *Computer Speech & Language*, 2025.
17. E. Aguilera-Rodríguez *et al*., "An EEG-based imagined speech database for comparing experimental paradigms and human factors," *Scientific Data*, 2025, doi: 10.1038/s41597-025-05926-5.
18. J. P. C. Moreira *et al*., "An open-access EEG dataset for speech decoding," *Scientific Data*, 2025, doi: 10.1038/s41597-025-05187-2.
19. X. Ma *et al*., "3M-CPSEED: An EEG-based dataset for Chinese Pinyin speech production and imagery," *Scientific Data*, 2025, doi: 10.1038/s41597-025-06346-1.
20. T. He *et al*., "VocalMind: A stereotactic EEG dataset for vocalized, mimed, and imagined speech," *Scientific Data*, 2025, doi: 10.1038/s41597-025-04741-2.
21. F. Elwasify *et al*., "EEG imagined speech neuro-signal preprocessing and deep-learning classification," *Scientific Reports*, 2026, doi: 10.1038/s41598-026-39395-6.
22. M. Jiang *et al*., "Decoding covert speech from EEG by functional areas and cross-frequency characteristics," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2026.
23. Z. Guo *et al*., "EEG-SPELL: Spectral–phonetic–lexical alignment network for EEG covert-speech decoding," *IEEE Journal of Biomedical and Health Informatics*, 2026.
24. F. Iacomi *et al*., "Novel EEG-based signatures of brain connectivity for sentence-level imagined speech," *Computers in Biology and Medicine*, 2026.
25. H. T. M. Duhair *et al*., "Progressive transfer learning unifies multi-corpus EEG for imagined-speech decoding," *Biomedical Signal Processing and Control*, 2026.
26. S. Dörterler *et al*., "A dynamic subject-invariant fragment mixing strategy for multi-class imagined-speech EEG classification," *Neuroscience*, 2026.
27. "Imagined Chinese speech decoding based on initials and finals using deep neural learning," *IET Signal Processing*, 2026.
28. S. M. Tasin *et al*., "An ensemble machine-learning model for inner-speech EEG classification," *IEEE Access*, 2026.
29. R. Liu, Y. Chao, X. Ma, X. Sha, L. Sun, S. Li, and S. Chang, "ERTNet: An interpretable transformer-based framework for EEG emotion recognition," *Frontiers in Neuroscience*, vol. 18, Art. no. 1320645, 2024, doi: 10.3389/fnins.2024.1320645.
30. F. Hu, F. Wang, J. Bi, Z. An, C. Chen, G. Qu, and S. Han, "HASTF: A hybrid attention spatio-temporal

feature fusion network for EEG emotion recognition," *Frontiers in Neuroscience*, vol. 18, Art. no. 1479570, 2024, doi: 10.3389/fnins.2024.1479570.

31. G. Moise *et al*., "Towards integrating automatic emotion recognition in education: A deep learning model based on five EEG channels," *International Journal of Computational Intelligence Systems*, vol. 17, Art. no. 230, 2024, doi: 10.1007/s44196-024-00638-x.
32. T. Dhara, P. K. Singh, and M. Mahmud, "A fuzzy ensemble-based deep learning model for EEG-based emotion recognition," *Cognitive Computation*, vol. 16, pp. 1364–1378, 2024, doi: 10.1007/s12559-023-10171-2.
33. X. Zhu *et al*., "EEG emotion recognition network based on attention and spatiotemporal learning," *Sensors*, vol. 24, no. 11, Art. no. 3464, 2024.
34. M. Miao *et al*., "ST-SHAP: A hierarchical and explainable attention network for emotional EEG recognition," *Journal of Neuroscience Methods*, 2025.
35. S. Sylvester, M. Sagehorn, T. Gruber, M. Atzmueller, and B. Schöne, "SHAP value-based ERP analysis (SHERPA): Increasing the sensitivity of EEG signals with explainable AI methods," *Behavior Research Methods*, vol. 56, pp. 6067–6081, 2024, doi: 10.3758/s13428-023-02335-7.