Optimizing Deep Neural Networks for EEG-Based Speech Recognition: A Multimodal Approach to Assistive Communication

Anarghya Das, *Graduate Student Member, IEEE*, Puru Soni, Hubin Zhao, *Member, IEEE*, Ming-Chun Huang, and Wenyao Xu, *Senior Member, IEEE*

***Abstract*—Speech recognition for individuals with impairments remains a significant challenge due to atypical speech patterns that confound traditional acoustic-only models. This study introduces NeuroSpeech, a novel multimodal framework that integrates electroencephalography (EEG) with acoustic features to improve recognition accuracy, robustness, and efficiency. A large-scale random search identified optimal EEG encoder configurations and feature extraction parameters, with window size and overlap ($p < 0.001$) emerging as critical factors. Explainable AI (XAI) methods, specifically SHAP, provided insights into model decision-making, supporting interpretability and clinical translation. Evaluations were conducted on two publicly available datasets: Spanish commands and vowels (UNLP-CONICET) and English phonemes and words (KaraOne). Under clean conditions, NeuroSpeech achieved near-perfect accuracy ($F1 = 0.986$ on Spanish; 0.837 on English), while in noisy conditions (SNR $=$ 0.5) it maintained strong performance ($F1 = 0.92$ and 0.70), demonstrating EEG's role as a noise-robust complementary signal. In contrast, Whisper, a state-of-the-art ASR model, showed severe degradation under noise (e.g., $F1$ dropping from 0.81 to 0.46). Finally, complexity analysis showed that NeuroSpeech is lightweight (1–30M parameters) with inference latency of 10–18ms/sample (RTF < 1 on CPU and GPU), enabling near-real-time deployment. These results demonstrate NeuroSpeech's significant potential to leverage neural information to augment speech that is compromised, offering a promising advancement for assistive technologies and improved communication for individuals with speech disorders.**

***Index Terms*—EEG-based speech recognition, multimodal speech recognition, brain-computer interface, assistive communication, deep learning, explainable AI (XAI).**

Received 28 May 2025; revised 30 September 2025; accepted 5 October 2025. Date of current version 8 December 2025. *(Corresponding author: Anarghya Das.)*

Anarghya Das, Puru Soni, and Wenyao Xu are with the Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY 14260 USA (e-mail: anarghya@buffalo.edu).

Hubin Zhao is with the Department of Medical Physics and Biomedical Engineering, University College London, HA7 4LP London, U.K.

Ming-Chun Huang is with the Department of Data and Computational Science, Duke Kunshan University, Kunshan, Jiangsu 215316, China.

Digital Object Identifier 10.1109/JBHI.2025.3618998

I. Introduction

NEARLY 100 million individuals worldwide cannot rely on their natural speech for effective communication, often requiring specialized support from Speech-Language Pathologists (SLPs) [1]. This substantial population underscores the urgent demand for highly personalized AI-driven technologies that adapt to unique individual needs, moving beyond one-size-fits-all solutions. Automatic Speech Recognition (ASR) is a pivotal AI technology with the potential to bridge this communication divide. While modern ASR systems have achieved remarkable performance under ideal conditions, primarily due to advances in deep learning and the adoption of end-to-end architectures such as Transformers [2], their traditional reliance almost exclusively on acoustic inputs, typically time-frequency representations like log Mel spectrograms to map speech to text, presents a critical limitation. This deep dependence on clear, well-articulated acoustic signals renders ASR systems highly vulnerable when these signals are compromised. The vulnerability is starkly evident in clinical populations with speech impairments, such as dysarthria or apraxia, where motor or planning deficits lead to distorted phonemes, altered speech rates, disrupted prosody, and fragmented vocalizations [3], [4]. Such atypical acoustic patterns, often out of distribution for standard ASR models, result in sharply increased Word Error Rates (WERs) and significant communication breakdowns [5]. The challenge is compounded in noisy clinical environments where essential acoustic cues can be masked or distorted. Critically, when acoustic features are minimal or absent, as in cases of anarthria or severe apraxia, conventional ASR systems fundamentally fail, leaving many individuals without a viable means of speech-based technological interaction. Thus, the degraded and unpredictable nature of impaired speech challenges the core assumptions of acoustic-only ASR systems and highlights the need for supplementary modalities that capture speech intent upstream of articulation.

The human brain encodes rich linguistic and motor information, offering an untapped resource for speech-decoding systems. Recent Brain-Computer Interface (BCI) research has demonstrated that electroencephalography (EEG) recordings capture meaningful representations of speech-related neural

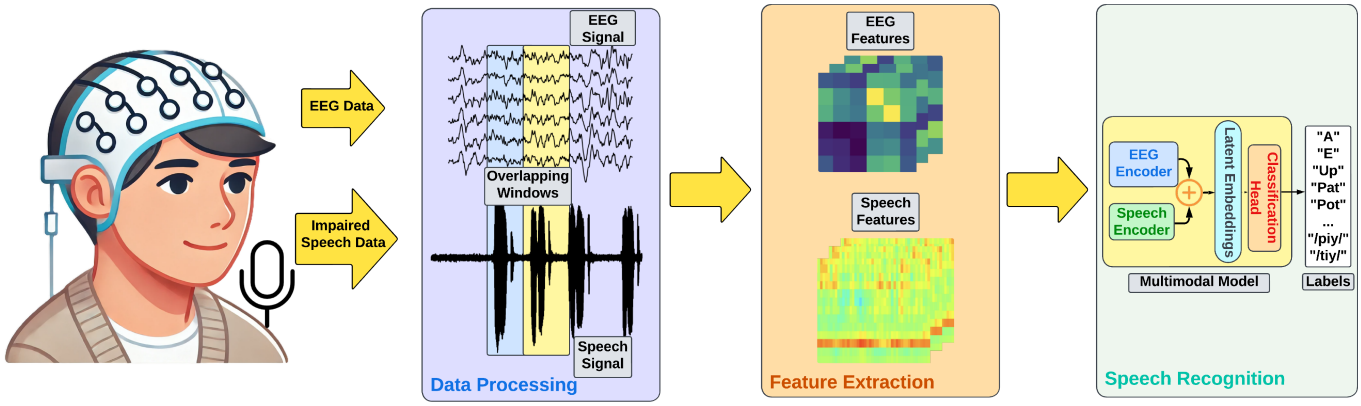

Fig. 1. Overview of the NeuroSpeech framework: Integrating EEG and speech data through feature extraction, encoding, and fusion to enhance speech classification accuracy and robustness when acoustic signals from impaired or noisy speech are compromised.

activity, opening new possibilities for augmenting ASR [6], [7]. EEG can capture brain activity associated with various stages of speech production, including intention formation, linguistic encoding, and motor planning, often before verbal articulation occurs or even in its absence. This capability to tap into pre-articulatory neural signatures means EEG can provide crucial information even when the acoustic signal is severely degraded, noisy, or altogether missing. Integrating EEG as an additional modality into ASR systems can support the work of SLPs and advance neural and rehabilitation engineering. SLPs utilize performance-based auditory perception, subjective clinical judgment, and patient-reported outcome measures to establish goals and assess treatment effects. However, the use of patient-reported outcome measures is limited by barriers such as time constraints and patient insight deficits [8]. EEG-augmented ASR can provide an invaluable source of objective, quantifiable data directly reflecting an individual's neural speech processing. Furthermore, such systems can facilitate personalized therapy planning and progress monitoring; SLPs could leverage decoded neural information to gauge a patient's cognitive engagement or track neurophysiological changes in response to specific interventions, allowing for more adaptive and effective therapeutic strategies tailored to individual recovery trajectories.

This work introduces a neuro-enhanced speech decoding framework that integrates brain-derived features into conventional ASR systems. We propose a neuro-enhanced ASR framework (Fig. 1) to overcome acoustic limitations and empower SLPs with a personalized tool. Using two publicly available datasets, we develop a hybrid ASR architecture that fuses neural and acoustic representations, demonstrating how brain-derived features can enhance traditional speech recognition using explainable AI methods. Key contributions of our study include:

- **EEG–Audio Fusion for Robust ASR**: NeuroSpeech improves speech recognition by combining EEG neural features with audio, showing strong gains under noisy conditions.
- **Optimized EEG Encoding**: A comprehensive design space search identified the Time-Distributed CNN as the most effective encoder for multimodal fusion.
- **Feature Parameter Insights**: Window size and overlap ($p < 0.001$), along with key hyperparameters (learning rate, dense units), were found to impact performance and inform future clinical adaptation significantly.
- **Robustness, SOTA Benchmarking & XAI**: Compared to Whisper [9], NeuroSpeech is more noise-resilient, and SHAP-based explainability enhances interpretability for clinical use.
- **Low-Latency Deployment**: Computational profiling shows low latency (10–18 ms/sample) and RTFs < 1 on CPU/GPU, supporting near real-time assistive applications.

The remainder of this paper is organized as follows. Section II reviews related work and motivates our multimodal approach. Section III details the NeuroSpeech framework and experimental setup, including the design space optimization strategy. Section IV presents results and analysis, while Section V discusses implications, limitations, and future directions. Finally, Section VI provides an overall summary, highlights areas for improvement, and outlines potential avenues for future research.

II. Background and Related Works

Speech is broadly categorized into overt speech, involving physical articulation and audible sound, and covert speech, the internal simulation of language without vocal output, relying purely on cognitive processes [10]. While decoding covert speech directly from EEG to potentially replace traditional ASR is a compelling prospect, its accuracy remains significantly lower than ASR standards. This is mainly due to challenges such as low signal-to-noise ratios and the inherently indirect nature of imagined speech signals, which complicate the reliable reconstruction of speech from neural activity alone [11], [12]. Existing works using the specific datasets employed in this study have predominantly focused on EEG-only decoding approaches [13]; however, the results often highlight inherent limitations, underscoring the need to explore different techniques to leverage EEG data and enhance performance. In our previous work [14], we systematically disentangled the contributions of the audio and EEG streams, demonstrating that each modality provides unique and non-redundant information. Our multimodal experiments demonstrated that EEG alone performed poorly in absolute terms; however, when fused with audio, it offered additional

discriminative information that stabilized recognition performance. Further embedding visualizations (t-SNE and silhouette analysis) confirmed that the multimodal model learned better class separability than either of the unimodal baselines, suggesting that EEG inputs enrich the latent representation space and sharpen class boundaries.

Beyond speech recognition, multimodal learning has been shown to enhance speech emotion recognition (SER), where deep bimodal models (audio+text) utilize multi-head attention and fusion strategies to leverage complementary signals [15], [16]. Yet, these works focus on affective state classification rather than linguistic decoding. Similarly, systems like ClinClip [17] align listener EEG with perceived speech to enhance medical transcription; however, these systems tackle perception rather than production. In contrast, our work uniquely positions EEG as a complementary signal on the production side in ASR, directly addressing gaps in noise robustness, assistive applications, and interpretability. By combining EEG with audio, we aim to demonstrate that neural activity provides resilience where acoustic channels falter, and that this integration yields not just higher accuracy but also new insights into modality-specific contributions.

III. Methods

A. Dataset Information

This study utilizes two publicly available EEG-speech datasets: Dataset 1 (UNL-CONICET) [18] and Dataset 2 (Kara One) [19]. Dataset 1 includes recordings from 15 healthy Argentinian college students (mean age 25; 8 male, 7 female), all native Spanish speakers without hearing impairments. EEG signals were captured using six Ag-AgCl electrodes (F3, F4, C3, C4, P3, P4) at 1024 Hz, alongside 44.1 kHz audio, during tasks involving five vowels and six Spanish command words. Each trial included rest, visual prompt, 4-second imagined rehearsal, and 4-second overt speech, repeated approximately 50 times per word (total session duration: 3.5 hours). Dataset 2 consists of 12 healthy English-speaking participants (8 male, 4 female) from the University of Toronto, recorded with a 64-channel Neuroscan Quick-Cap at 1000 Hz and 12 kHz audio. Participants responded to seven phonemes and four English words in a sequence of imagined and overt speech, repeated 12 times per prompt. Unlike Dataset 1's continuous vocalizations within trials, Dataset 2 included only single utterances, allowing for a cleaner isolation of overt versus imagined states. Only the EEG and audio segments aligned with spoken speech were used in this study. These datasets provide complementary conditions: low-density vs. high-density EEG, repeated vs. single utterances, and Spanish vs. English prompts, supporting analysis of multimodal speech decoding across distinct neural and linguistic settings.

B. Data Preprocessing

For Dataset 1 (1974 trials, 15 participants; multiple word repetitions, prolonged vowels), EEG data was already bandpass filtered (2-40 Hz) to isolate relevant frequencies and remove 50 Hz line noise. Muscle and electrode artifacts were manually removed; blink artifacts were annotated but retained for further analysis. Our analysis used 4-second epochs, matching the original EEG and audio recording lengths, thus requiring no trimming or padding. Dataset 2 (1647 trials, 12 participants; single word/phoneme repetitions) contained raw EEG data that was not pre-processed. Preprocessing steps included filtering 60 Hz line noise, a 1 Hz high-pass filter to reduce low-drift noise, and a 1-50 Hz band-pass filter. All trials were standardized to 2 s by trimming or padding.

EEG data and corresponding audio recordings were segmented into overlapping windows to facilitate the training and testing of multimodal models. Various window sizes and overlap values were explored during our experiments, the details of which are provided in the Experiments section. This segmentation enhances feature extraction by capturing time-dependent signal variations, allowing the model to detect subtle correlations between brain activity and speech patterns. Overlapping windows increase the training data, improving the model's ability to generalize and decode speech more accurately. Furthermore, synchronized windowing ensures alignment between the EEG signals and audio recordings. This is crucial for learning meaningful correlations between brain activity and the speech produced.

C. Feature Extraction

This study explored two types of EEG features: time-domain features and a hybrid approach combining time-domain and time-frequency domain features. It aimed to evaluate which method contributed most effectively to speech recognition performance.

1) Time-Domain Features: We utilized widely used statistical measures for temporal features, including mean, variance, skewness, kurtosis, standard deviation, average amplitude change, and zero crossing rate [20]. These temporal features were chosen because they thoroughly represent the fundamental dynamics of EEG signals during auditory processing. They encapsulate critical aspects of the signal's central tendency, variability, and distribution shape, allowing us to model fluctuations in neural activity over time in response to speech. By concentrating on these statistical attributes, we aim to capture the intricate patterns of brain activity that correspond with the temporal structure of speech, such as rhythm, intensity, and transitions between phonemes or syllables. This method enhances our ability to decode speech, as these features underscore the stability and variability of neural responses to auditory stimuli. When integrated with audio data, these temporal features facilitate the mapping of brain activity to specific speech components, streamlining the decoding process in a data-driven yet comprehensible manner.

2) Time-Frequency Domain Features: We utilized a power spectrum cross-covariance matrix for the time-frequency domain, an approach that has shown promise in imagined speech applications [21]. We initially computed the power spectrum of each EEG channel in decibels using the Fast Fourier Transform (FFT) ((1)), retaining only the first half of the frequency

spectrum to account for the symmetry of real-valued signals.

$$\mathrm{y} = 20 \cdot \log_{10}\left(|\mathrm{FFT(x)}|^2\right) \tag{1}$$

Here, x represents the EEG data from each channel after preprocessing windows and overlaps, and y denotes the corresponding power spectrum. We constructed cross-covariance matrices to further capture inter-channel relationships, providing a structured representation of spectral power and spatial dependencies within each EEG window. The final EEG feature representation had a $channel \times channel$ dimension, resulting in 6×6 feature matrices for Dataset 1 and 64×64 for Dataset 2.

3) Speech Features: Mel Frequency Cepstral Coefficients (MFCCs) stand out for audio features because they mimic human auditory sensitivity by compressing frequency information into the Mel scale. This reduction in complexity makes MFCCs widely used in speech and audio recognition applications [22]. Thus, we used MFCCs for audio feature extraction in our research. We extracted 13 MFCC features from the speech data. To facilitate a comprehensive cross-modal analysis between the MFCC and EEG features, we aligned these modalities by making temporal adjustments. Recognizing the inherent differences in sampling rates between audio and EEG signals, we truncated the MFCC sequence to match the EEG windows. This decision aimed at achieving temporal coherence and synchronization between the modalities. We ensured the time points synchronized by aligning the MFCC sequence with the EEG data.

D. Multimodal Models

We propose a multimodal framework for speech decoding that combines processed EEG and audio features. This framework leverages complementary information from brain and speech signals to classify spoken phrases effectively. The architecture comprises three variations of EEG encoders and a consistent audio encoder, followed by a late fusion strategy and a classification head, as illustrated in Fig. 2. The EEG encoder varies across three configurations, each designed to evaluate the effectiveness of different EEG feature representations in decoding speech.

1) Gated Recurrent Unit (GRU) for Statistical Temporal Features: The GRU-based EEG encoder (Fig. 2(A)) is designed to process time-domain statistical features derived from EEG signals, leveraging the temporal structure of the data to decode speech-related patterns. The input consists of EEG data divided into temporal windows, each containing statistical features calculated across the EEG channels.

2) Convolution LSTM (ConvLSTM) for Frequency Domain and Temporal Dependencies: This EEG encoder (Fig. 2(B)) combines convolutional layers for spatial feature extraction with long short-term memory (LSTM) layers to model temporal dependencies, capturing spatiotemporal dynamics inherent in multichannel EEG signals. The input to this encoder consists of EEG cross-covariance FFT feature windows.

3) Time-Distributed Convolutional Neural Network (TD-CNN) for Frequency-Based Spatial Features: The TD-CNN EEG encoder (Fig. 2(C)) leverages convolutional operations applied to temporal segments of cross-covariance matrices, capturing spatial features critical for decoding speech-related patterns in EEG signals. The input to this encoder consists of EEG cross-covariance FFT feature windows. Each temporal window is processed independently, preserving temporal structure while learning spatial relationships across EEG channels. The convolutional layers are parameterized to extract hierarchical spatial features, where the first layer applies 64 filters and the second layer uses 128 filters, each followed by a rectified linear unit (ReLU) activation function.

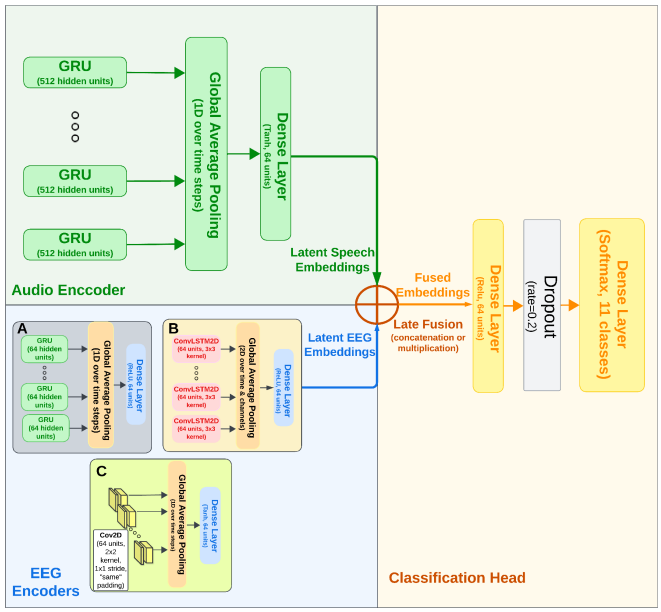

Fig. 2. NeuroSpeech Multimodal Framework: Explores three EEG encoder architectures: (A) GRU, (B) ConvLSTM, and (C) Time-Distributed CNN-based. Audio and EEG embeddings are fused via late fusion, followed by a classification head to predict 11 speech classes.

Each EEG encoder outputs a latent EEG representation z_{eeg}, following its respective architecture. These encoders use a Global Average Pooling layer to aggregate temporal information into a fixed-size representation. After obtaining the fixed-length neural embedding, the framework processes the synchronized speech waveform through a dedicated audio encoder, yielding an analogous acoustic embedding z_{audio}.

4) Audio Encoder: The GRU-based audio encoder processes speech audio features in MFCCs to produce a latent representation of speech in a reduced-dimensional feature space. This model encodes speech audio into a latent space while preserving the essential temporal and spectral information. Inspired by insights from [23], it comprises three stacked GRU layers, each with 512 units, which model temporal relationships within the MFCC input.

After obtaining the latent representations z_{audio} and z_{eeg}, a late fusion strategy combines them by concatenation using (2):

$$z_{fused} = concat(z_{audio}, z_{eeg}) \tag{2}$$

Where z_{fused} is the fused feature vector representing both modalities. Late fusion combines EEG and audio embeddings after independent processing, ensuring efficient integration while preserving modality-specific features [24]. The fused vector is passed through dense layers with ReLU and dropout for refinement and regularization, followed by a softmax layer for class prediction.

TABLE I
DESIGN SPACE OF HYPERPARAMETERS EXPLORED IN NEUROSPEECH

Category	Range / Values
Feature Extraction	
Window size	10–500 samples (Low: 10–100, Medium: 100–250, High: 250–500)
Window overlap	0–0.75 (Low: 0–0.25, Medium: 0.25–0.50, High: 0.50–0.75)
Global Hyperparameters	
Dropout rate	0–2
Learning rate	1×10^{-6} – 1×10^{-3}
Batch size	16, 32, 64
Fusion strategy	Concatenation, Multiplication
Encoder-Specific	
GRU units (GRU)	128, 256, 512
Conv filters (TD-CNN)	32, 64, 128
ConvLSTM filters	16, 32, 64
Dense units	128–1024

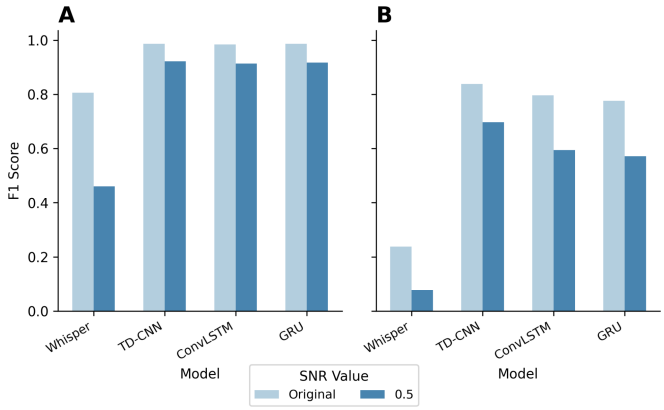

Fig. 3. Comparison of F1 Scores Across Models Under Original and Noisy Conditions (SNR = 0.5). The chart compares the performance of Whisper and the three NeuroSpeech EEG encoders in (A) Dataset 1 and (B) Dataset 2.

IV. EXPERIMENTS

To comprehensively evaluate the NeuroSpeech framework, we established a design space spanning feature extraction, model hyperparameters, and artificial noise augmentation. The goal was to optimize speech classification accuracy, identify top configurations, and determine the most effective EEG encoder. We benchmarked NeuroSpeech against Whisper (large V3 model) to assess whether EEG provides additional contextual information for enhanced speech decoding.

A. Exploration of the Design Space

The optimal EEG encoder among three variations, along with its hyperparameters, was identified through an extensive random search, averaging 9420 iterations. Each configuration was evaluated using F1 Score and Cohen's Kappa on the 11-class prompt classification task with participant-independent 5-fold cross-validation to ensure generalization.

Table I summarizes the explored design space, covering feature extraction parameters (window size and overlap), model-level hyperparameters (dropout, learning rate, batch size), encoder-specific settings (GRU units, Conv filters, ConvLSTM filters, dense units), and fusion strategies. This organization allows a clear view of both the global and encoder-specific hyperparameters optimized in our framework.

1) Model Hyperparameters: Model hyperparameter selection focused on key parameters impacting performance across our multimodal framework's three EEG encoder configurations. Core global hyperparameters, such as dropout rate, batch size, and learning rate, were selected to influence the overall model's generalization, optimization, and prevention of overfitting. Specific to each EEG encoder, we varied the number of GRU units for the GRU-based encoder, the convolutional filters for the Time-Distributed CNN encoder, and the ConvLSTM filters for the ConvLSTM encoder. Each encoder also included configurable dense layers to refine EEG representations. We parameterized the fusion strategy (concatenation vs. multiplication) for multimodal fusion and the final dense layers that combine EEG and speech embeddings before classification.

2) Effect of Artificial Noise: To simulate challenging real-world conditions and evaluate the robustness of the NeuroSpeech framework, we introduced artificially generated white noise into the speech recordings. This augmentation enabled us to assess the effectiveness of incorporating EEG signals under varying acoustic conditions and investigate whether neural features improve speech decoding when the audio signal is degraded. Artificial white noise was added at different difficulty levels, parameterized by the signal-to-noise ratio (SNR), where lower SNR values indicate higher noise levels. The SNR values ranged from 0.5 (extremely noisy) to 50 (original clean recordings), allowing us to span a realistic noise design space. This procedure ensured precise control over the SNR for each recording, enabling consistent augmentation across the dataset.

V. RESULTS & DISCUSSION

This section presents the outcomes of our design space experiments, the analysis of hyperparameter importance using SHAP, and an investigation into the effects of windowing strategies.

A. NeuroSpeech Performance and Robustness in Simulated Challenging Conditions

To benchmark NeuroSpeech, we compared the performance of three EEG encoders (TD-CNN, ConvLSTM, and GRU) with that of Whisper (large-v3). Results across both datasets are summarized in Fig. 3, which shows that all EEG–audio models substantially outperform Whisper, particularly in the presence of noise. Among the tested encoders, TD-CNN consistently provided the best balance of accuracy and robustness, with Table II listing the optimal configurations that yielded these results. Accordingly, we focused subsequent validation and analysis on this model.

On Dataset 1, TD-CNN achieved near-perfect clean audio performance ($F1 = 0.986$, $\kappa = 0.985$), while Whisper reached only 0.81. Under noisy conditions (SNR = 0.5), Whisper degraded sharply to 0.46, whereas TD-CNN maintained $F1 = 0.92$, showing EEG's critical role when audio is corrupted

TABLE II
OPTIMAL TD-CNN CONFIGURATIONS IDENTIFIED FOR EACH DATASET

Dataset	Best Configuration
Dataset 1	W = 50, Ov = 0, Conv = 64, Dense = 512, Drop = 0.2, LR = 3×10^{-5}, Fusion = Concat
Dataset 2	W = 100, Ov = 0.75, Conv = 128, Dense = 256, Drop = 0.2, LR = 3×10^{-4}, Fusion = Concat

Abbrev.: W = window size (ms), Ov = overlap, Conv = Conv2D filters, Dense = encoder dense units, Drop = dropout, LR = learning rate.

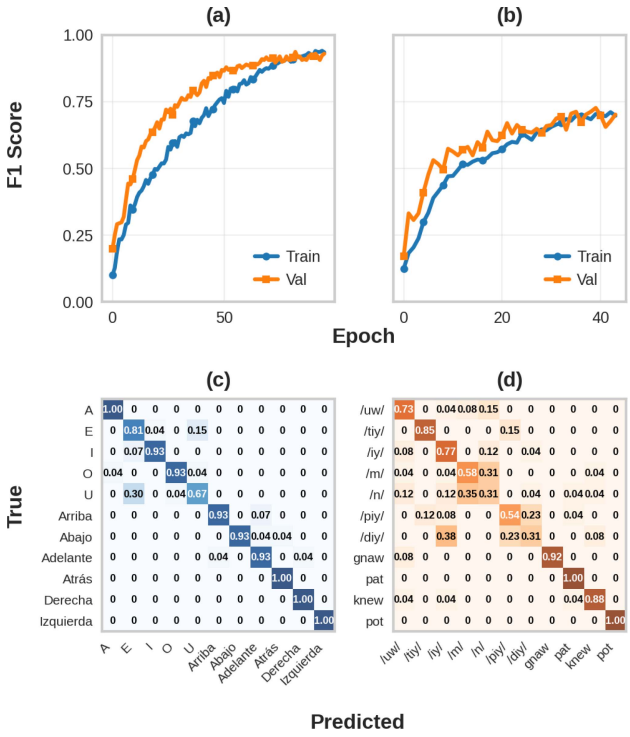

Fig. 4. Performance of the TD-CNN model under noisy conditions (SNR = 0.5). Panels (a) and (b) show the model learning curves (train vs. validation F1 scores) for Dataset 1 and Dataset 2, respectively. Panels (c) and (d) present the corresponding confusion matrices for Dataset 1 and Dataset 2, illustrating class-by-class performance under noisy input conditions.

(Fig. 3(A)). The learning curve (Fig. 4(a)) confirms smooth convergence and stable validation performance, while the confusion matrix (Fig. 4(c)) shows almost complete class separability, with only minor confusions among vowels (e.g., 'E' and 'U'). On Dataset 2, TD-CNN achieved $F1 = 0.837$ in clean audio and retained $F1 = 0.70$ under noise, compared to Whisper's 0.24 and 0.08, respectively (Fig. 3(B)). The learning curve (Fig. 4(b)) again indicates steady convergence without divergence between training and validation curves. The confusion matrix (Fig. 4(d)) highlights strong recognition of distinct word-level stimuli such as pat and pot. However, higher overlap between acoustically similar phoneme-level units (e.g., /m/ and /n/, or /piy/ and /diy/) reflects the intrinsic challenge of short phonemic stimuli, since these sounds share similar spectral and temporal cues and evoke overlapping neural responses. In contrast, the results in Fig. 4(c) show much stronger overall performance because the longer,

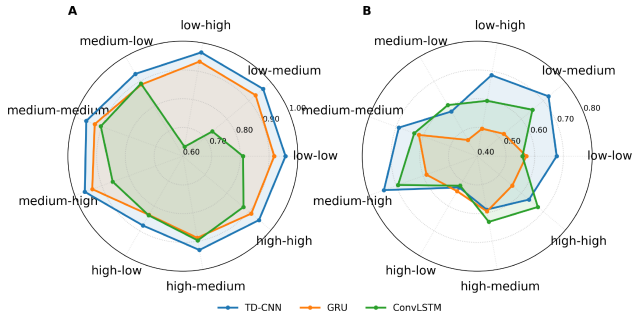

Fig. 5. Performance comparison of TD-CNN (blue), GRU (orange), and ConvLSTM (green) EEG encoders under various window size-window overlap configurations. (A) Model performance on Dataset 1. (B) Model performance on Dataset 2. Each axis on the radar plots represents a specific 'window size-overlap' pairing. The radial distance from the center along each axis reflects the achieved model performance for that particular configuration.

more acoustically distinct word stimuli provide richer temporal context and more separable EEG patterns. This suggests that even with multimodal input, recognition accuracy depends strongly on stimulus length and distinctiveness: short, similar phonemes remain harder to disentangle than longer, context-rich words.

In terms of computational complexity, TD-CNN remains lightweight compared to large ASR backbones. The best-performing configuration for Dataset 1 contained only 1.18 million parameters (≈ 4.7 MB), while the Dataset 2 configuration used 30.4 million parameters (≈ 122 MB). Measured inference times per sample were 18.5ms (CPU) / 1.0ms (GPU) for Dataset 1 (50ms window, no overlap) and 10.8ms (CPU) / 1.4ms (GPU) for Dataset 2 (100ms window, 75% overlap). Both setups achieved a real-time factor (RTF) well below 1 on CPU and orders of magnitude below 1 on GPU, even under overlapping conditions, thereby enabling near-real-time communication. The RTF was calculated as the ratio of average inference latency to input window duration and overlap:

$$\text{RTF} = \frac{\text{Inference Time (ms/sample)}}{\text{Window Size (ms)} \times (1 - \text{Overlap})}$$

All measurements were performed on a system equipped with NVIDIA L40S GPU and an Intel Xeon Platinum 8562Y+ CPU, ensuring reproducibility of complexity and latency metrics.

B. Effect of Window Size and Window Overlap

To assess the influence of window size and window overlap on the performance of different EEG encoders, we performed 2-way Aligned Rank Transform (ART) ANOVAs [25]. This method was chosen because the results from the design space experiments for window size and overlap did not follow a normal distribution, making parametric ANOVA unsuitable. The ART ANOVA methodology effectively allowed us to isolate the independent and interactive effects of these crucial temporal segmentation parameters. Our analysis consistently revealed that window size and temporal overlap are pivotal, jointly steering the performance of every encoder. However, the strength of their

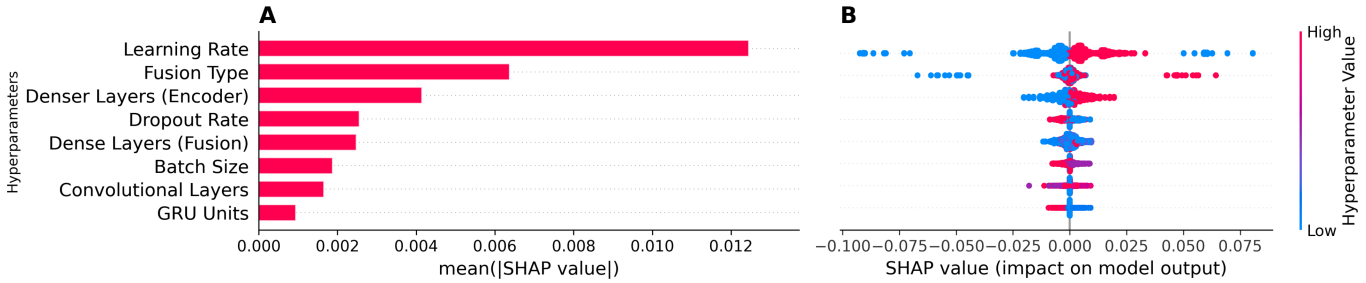

Fig. 6. (A) Overall hyperparameter importance ranking based on the mean absolute SHAP value across all encoder configurations and both datasets. (B) SHAP summary plot illustrating the impact of individual hyperparameter values on model performance. Each point represents a Shapley value for a specific instance of a hyperparameter, with color indicating whether the hyperparameter's value was high or low.

interplay varied depending on the dataset, as shown in Fig. 5(A) and (B).

1) GRU Encoder: For the GRU encoder, Dataset 1 highlighted a strong synergy: window size, overlap, and their interaction were all highly significant ($p < .001$). Optimal performance ($Kappa = 0.935$) was achieved with a small window and medium overlap, while a large window with low overlap yielded the poorest results (0.836). On Dataset 2, the landscape shifted: only overlap significantly impacted performance ($p < .01$), with window size ($p = .207$) and interaction effects ($p = .192$) receding in importance. Here, a medium window with high overlap proved most fruitful (0.616), with low-overlap settings notably underperforming (lowest at 0.497).

2) ConvLSTM Encoder: The ConvLSTM encoder echoed the complex dynamics seen on Dataset 1, with significant main effects for window size and overlap ($p < .001$), alongside a strong interaction effect ($p < .001$). A medium window paired with high overlap delivered the highest Kappa (0.904), whereas a small window with high overlap was least effective (0.633). For Dataset 2, while the interaction was not significant ($p = .270$), the main effects of window size ($p < .001$) and overlap ($p < .001$) remained influential. The optimal strategy mirrored that of Dataset 1: a medium window with high overlap (0.693), with large, low-overlap windows at the bottom (0.518).

3) TD-CNN Encoder: On Dataset 1, the TD-CNN encoder demonstrated acute sensitivity to all factors: window size, overlap, and their interaction ($p < .001$) were all profoundly significant. This led to the highest score in this dataset (0.966) with a small window and high overlap; conversely, a large window with low overlap produced the weakest result (0.878). In Dataset 2, these factors acted more independently (interaction $p = .285$), though both window size and overlap were crucial ($p < .001$). The best configuration emerged as a medium window with high overlap (0.745).

Across all encoders, a clear trend favored larger overlaps (around 50–75%), consistently boosting Kappa scores. However, the optimal window length was distinctly dataset-specific. Smaller windows proved advantageous for TD-CNN and GRU in the repetitive, utterance-rich Dataset 1. In contrast, medium windows were generally preferred for the ConvLSTM across both datasets and for all models when processing the shorter, single-instance data of Dataset 2.

C. Multimodal Hyperparameter Importance

We used the Shapley Additive Explanations (SHAP) method to evaluate the importance of the hyperparameters. This method assigns an importance value to each feature based on its contribution to a model's output. Our SHAP analysis (Fig. 6(A) and (B)) shows consistent hyperparameter effects across the EEG encoders. The learning rate emerged as the primary driver of performance for all architectures. A broad range of values between $1e^{-6}$ and $1e^{-2}$ was tested, and values in the narrower window of roughly $3e^{-5}$ to $3e^{-4}$ were most favorable, balancing fast convergence with stability. The ConvLSTM additionally benefited from slightly lower rates. Fusion type was the next most influential feature, indicating that preserving complete modality-specific information yields stronger multimodal representations. Switching from element-wise multiplication (blue, negative SHAP) to concatenation (red, positive SHAP) consistently raised performance. More dense units in the encoder improved accuracy, highlighting the need for sufficient representational capacity before multimodal fusion. The dropout rate also played a critical role, with a value of 0.2 consistently emerging as the most favorable and being used in the best-performing configurations. Fig. 6(B) shows that larger dense layers and moderate dropout (0–0.2) have a positive effect, potentially mitigating overfitting, whereas very small dense layers or no dropout hurt performance. Dense units in the fusion head and batch size followed in importance. Convolutional depth and GRU unit count had minimal impact, implying that expanding the dense layers at the end of the encoder yields greater benefits than adding complexity to the early sequence modeling stages.

VI. Conclusion

This work introduced NeuroSpeech, a multimodal EEG–audio framework optimized through extensive random search across feature extraction and model hyperparameters. The TD-CNN encoder consistently performed the best. On Dataset 1, it achieved $F1 = 0.986$ (clean) and 0.92 (noisy, SNR 0.5), while on Dataset 2, it obtained $F1 = 0.837$ (clean) and 0.70 (noisy). In both cases, NeuroSpeech substantially outperformed Whisper, particularly under noise, underscoring EEG's role as a complementary, noise-robust modality. Learning curves demonstrated smooth convergence, and confusion matrices confirmed strong

class separability, with only minor confusions among acoustically similar units. Computational analysis further showed low inference latency (10–18ms/sample) and RTFs < 1 on both CPU and GPU, supporting near-real-time feasibility.

Despite these promising results, the work has limitations. The datasets feature a restricted 11-class vocabulary and lack continuous speech recordings, constraining evaluation in natural conversational contexts. Performance was also lower on Dataset 2 due to its shorter, single-instance phonemic prompts, which have similar acoustic and neural patterns, making them harder to distinguish than longer, more distinct utterances in Dataset 1. Moreover, the experiments relied on publicly available datasets from healthy participants, rather than data collected directly from individuals with speech impairments, the target users of such systems. Future work should therefore expand to sentence-level decoding and larger vocabularies, and crucially, validate the framework on populations with impaired speech in real-world clinical settings to assess its scalability and impact fully.

REFERENCES

[1] D. R. Beukelman and J. C. Light, *Augmentative & Alternative Communication: Supporting Children and Adults With Complex Communication Needs*, 5th ed. Baltimore MD, USA: Paul H. Brookes Publishing Co, 2020.

[2] M. Malik, M. K. Malik, K. Mehmood, and I. Makhdoom, "Automatic speech recognition: A survey," *Multimedia Tools Appl.*, vol. 80, no. 6, pp. 9411–9457, Mar. 2021, doi: 10.1007/s11042-020-10073-7.

[3] I. Calvo et al., "Evaluation of an automatic speech recognition platform for dysarthric speech," *Folia Phoniatrica et Logopaedica: Official Organ Int. Assoc. Logopedics Phoniatrics (IALP)*, vol. 73, no. 5, pp. 432–441, 2021.

[4] S. Alharbi et al., "Automatic speech recognition: Systematic literature review," IEEE Access, vol. 9, pp. 131858–131876, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9536732

[5] K. Mokgosi, C. Ennis, and R. Ross, "Automatic speech recognition models for pathological speech: Challenges and insights," in *Proc. 32nd Ir. Conf. Artif. Intell. Cogn. Sci.*, Dec. 2024, pp. 63–74. [Online]. Available: https://ceur-ws.org/Vol-3910/aics2024_p63.pdf

[6] V. Viswanathan, H. M. Bharadwaj, and B. G. Shinn-Cunningham, "Electroencephalographic signatures of the neural representation of speech during selective attention," *eNeuro*, vol. 6, no. 5, Sep. 2019, doi: 10.1523/ENEURO.0057-19.2019. [Online]. Available: https://www.eneuro.org/content/6/5/ENEURO.0057-19.2019

[7] X.-Y. Pan, J.-J. Zou, P.-Q. Jin, and N. Ding, "The neural encoding of continuous speech - recent advances in EEG and MEG studies," *Acta Physiologica Sinica*, vol. 71, no. 6, pp. 935–945, Dec. 2019.

[8] F. Stagge, M. L. Cohen, A. L. Johnson, and A. M. Lanzi, "Speech-language pathologists' use of patient-reported outcome measures for adult patients with cognitive-communication disorders: A survey study," *Amer. J. Speech- Lang. Pathol.*, vol. 34, no. 2, pp. 798–817, Mar. 2025, doi: 10.1044/2024_AJSLP-24-00285.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proc. 40th Int. Conf. Mach. Learn.*, Jul. 2023, pp. 28492–28518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[10] C. Cooney, R. Folli, and D. Coyle, "Opportunities, pitfalls and trade-offs in designing protocols for measuring the neural correlates of speech," *Neurosci. Biobehavioral Rev.*, vol. 140, Sep. 2022, Art. no. 104783. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S014976342200272X

[11] D. Lopez-Bernal, D. Balderas, P. Ponce, and A. Molina, "A state-of-the-art review of EEG-Based imagined speech decoding," *Front. Hum. Neurosci.*, vol. 16, Apr. 2022, Art. no. 867281. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9086783/

[12] J. T. Panachakel and A. G. Ramakrishnan, "Decoding covert speech from EEG-A comprehensive review," *Front. Neurosci.*, vol. 15, 2021, Art. no. 642251.

[13] V. R. Carvalho, E. M. A. M. Mendes, A. Fallah, T. J. Sejnowski, L. Comstock, and C. Lainscsek, "Decoding imagined speech with delay differential analysis," *Front. Hum. Neurosci.*, vol. 18, May 2024, Art. no. 1398065. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11140152/

[14] A. Das, P. Soni, M.-C. Huang, F. Lin, and W. Xu, "Multimodal speech recognition using EEG and audio signals: A novel approach for enhancing ASR systems," *Smart Health*, vol. 32, Jun. 2024, Art. no. 100477. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352648324000333

[15] S. Kakuba, A. Poulose, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, vol. 11, pp. 113769–113789, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10286495/

[16] R. P, R. K. Dhaman, and A. Poulose, "Feature importance and model performance in deep learning for speech emotion recognition," in *Proc. 11th Int. Conf. Adv. Comput. Commun. (ICACC)*, Nov. 2024, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/10845528/

[17] G. Sun, "ClinClip: A multimodal language pre-training model integrating EEG data for enhanced english medical listening assessment," *Front. Neurosci.*, vol. 18, Jan. 2025, Art. no. 1493163, doi: 10.3389/fnins.2024.1493163/full.

[18] G. A. P. Coretto, I. E. Gareis, and H. L. Rufiner, "Open access database of EEG signals recorded during imagined speech," in *Proc. 12th Int. Symp. Med. Inf. Process. Anal.*, Jan. 2017, doi: 10.1117/12.2255697.

[19] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech," in *Proc. 2015 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 992–996.

[20] D. Chen, H. Huang, X. Bao, J. Pan, and Y. Li, "An EEG-based attention recognition method: Fusion of time domain, frequency domain, and non-linear dynamics features," *Front. Neurosci.*, vol. 17, Jul. 2023, Art. no. 1194554, doi: 10.3389/fnins.2023.1194554/full.

[21] A.-L. Rusnac and O. Grigore, "CNN architectures and feature extraction methods for EEG imaginary speech recognition," *Sensors*, vol. 22, no. 13, Jan. 2022, Art. no. 4679. [Online]. Available: https://www.mdpi.com/1424-8220/22/13/4679

[22] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980. [Online]. Available: https://ieeexplore.ieee.org/document/1163420

[23] G. Krishna, C. Tran, J. Yu, and A. H. Tewfik, "Speech recognition with no speech or with noisy speech," in *Proc. 2019 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, May 2019, pp. 1090–1094. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8683453

[24] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, Jan. 2021, Art. no. 104042. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885620301748

[25] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins, "The aligned rank transform for nonparametric factorial analyses using only anova procedures," in *Proc. ACM Conf. Hum. Factors Comput. Syst.*, New York: ACM Press, 2011, pp. 143–146. [Online]. Available: https://depts.washington.edu/acelab/proj/art/