Research Statement

Do Tri Nhan

Proposed PhD Research in Biosignal-Enabled Speech Synthesis

Motivation and Applications

The proposal research focuses on enabling spoken communication by predicting the acoustic speech signal from biosignals

This research is particularly motivated by:

- Individuals who have undergone total laryngectomy and permanently lost vocal fold function;
- Patients with severe motor speech impairments, such as those diagnosed with amyotrophic lateral sclerosis (ALS) or locked-in syndrome;
- Individuals affected by neurological or neuromuscular disorders resulting in the loss of overt speech production.

This work has the potential to contribute to the development of next-generation biosignal-driven speech communication devices and rehabilitation technologies.

1 Introduction and Research Scope

Biosignal-enabled spoken communication is an emerging research trend that intersects the domains of biosignal processing and speech technology. The applications of biosignals in speech tasks encompass speech recognition, synthesis, enhancement, voice conversion, and auditory attention detection. This research proposal primarily focuses on **Biosignal-Enabled Speech Synthesis** [4].

1.1 Comparative Taxonomy: Neural vs. Peripheral Speech Interfaces

The following table summarizes the key differences between Brain-Computer Interfaces (BCI) and Silent Speech Interfaces (SSI):

Table 1: Comparative Taxonomy of Neural vs. Peripheral Speech Interfaces

Interface	**Brain–Computer Interface (BCI)**	**Silent Speech Interfaces (SSI)**
Signal Input	EEG (non-invasive), ECoG, fMRI, MEG, Utah array (invasive), Neuralink	EMA, EMG, High-Speed Nasopharyngoscopy (HSN), Ultrasound Tongue Imaging (UTI), Lip video
Data Attribute	Very low amplitude (10–100 μV), low SNR, artifacts (blinking, motion)	Higher amplitude than EEG (0.1–5 mV), superior SNR
Physiological Origin	Central nervous system (Neurological activities)	Peripheral nervous system (Respiratory/laryngeal/articulatory activity)
Frequency	Non-invasive: 0.5–40 Hz; Invasive: kHz	20–500 Hz

1.2 Research Directions

- **Silent Speech Interfaces (SSI):** This proposal direction focuses on Electromyography (EMG) signals due to ease of data collection and diversity (EMG2Speech).
- **Brain-to-Speech (BCI):**
  - *Non-Invasive BCI (EEG):* Prioritized for data availability (EEG2Speech). Current research is limited to closed vocabularies (50 words) due to high noise and hardware constraints. Significant gaps remain between output quality and reality; hence, this is not the primary focus for Year 1.
  - *Invasive BCI (Utah Array):* Utilizes high-resolution signals from the motor cortex via existing datasets.

With recent improvements in Word Error Rates (WER) — approximately 5.8% for invasive brain signals and 12.2% for EMG — techniques from traditional ASR/TTS/NLP are increasingly being integrated into this field [3].

2 Silent Speech Interfaces (SSI) with EMG

Silent Communication Processes (SCP) investigate how facial and neck muscle activities, captured via Electromyography (EMG), can decode speech and emotional states even in the absence of acoustic output. This field holds transformative potential for speech-assistive technologies, secure silent communication, and emotion-aware human-computer interaction. Some SOTA research paper:

2.1 Digital Voicing of Silent Speech: Addressing the Ground-Truth Paradox

A fundamental challenge in SSI is the **signal discrepancy** between training and inference phases, as identified in the works of Gaddy et al. [5–7]:

- **The Paradox:** Models are typically trained on *vocalized facial EMG* (collected while speaking aloud) to utilize the synchronized audio as ground truth. However, the target application is *silent facial EMG* (muscle movements without sound).
- **Data Contribution:** To bridge this gap, a novel 20-hour dataset was introduced, featuring parallel recordings of both silent and vocalized EMG signals [8].
- **Methodological Evolution:**
  - *Baseline Approach:* Directly applying a model trained on Vocalized EMG to Silent EMG results in a severe **domain shift**, with Word Error Rates (WER) reaching 66%.
  - *Proposed Transfer Strategy:* Instead of direct decoding, the method maps *Silent EMG* $\rightarrow$ *Vocalized EMG* $\rightarrow$ *Audio*. This intermediate alignment significantly improves performance, achieving $\sim$3.6% WER for closed vocabularies and a 20% relative reduction in open vocabulary tasks.

2.2 A Cross-Modal Approach with LLM-Enhanced Recognition (2024)

Recent SOTA advancements have further pushed the boundaries of non-invasive SSI, achieving a breakthrough WER of below 15% for open vocabulary for the first time. The framework introduces two core innovations:

1. **MONA (Multimodal Orofacial Neural Audio):** Learns a shared cross-modal latent representation between silent neural signals and audio. This allows the system to inherit linguistic and phonetic richness from large-scale pre-trained speech models.
2. **LISA (LLM Integrated Scoring Adjustment):** A post-processing refinement stage that utilizes Large Language Models (LLMs) for rescoring. LISA leverages high-level semantic context to correct misalignments and OOV (Out-of-Vocabulary) errors.

Benchmark Performance: The MONA+LISA architecture demonstrates significant improvements over previous SOTA on the Gaddy (2020) benchmark:

Test Set	**Previous SOTA (WER)**	**MONA + LISA (WER)**
Silent Speech (Open Vocab)	28.8%	**12.2%**
Vocalized EMG	23.3%	**3.7%**

Beyond EMG, this approach was fine-tuned on the **Utah Array** dataset, securing a top-2 position in the *Brain-to-Text Benchmark 24*, proving its robustness across both peripheral and central nervous system signals.

3 Brain–Computer Interface (BCI) with Invasive Utah Array

Invasive BCI research focuses on recording direct action potentials (spikes) from neurons to achieve high-fidelity speech reconstruction. This direction primarily utilizes data from the **Brain-to-Text Benchmark 24** challenge.

3.1 Signal Characteristics and Decoding Framework

The core signal input is derived from **intracortical microelectrode array (MEA)** recordings using Utah arrays. These invasive intracortical spike recordings provide significantly higher spatial and temporal resolution compared to EEG, making them ideal for high-precision speech decoding. The decoding task is formulated as mapping a variable-length time series of neural activity into corresponding text sequences.

3.2 Benchmark Dataset: High-Performance Speech Neuroprosthesis

A pivotal resource for this research is the dataset provided by Willett et al. [10]:

- **Scale:** 80GB of high-resolution neural recordings.
- **Content:** 12,100 sentences of intended speech from a participant with late-stage Amyotrophic Lateral Sclerosis (ALS).

- **Hardware:** Neural activity captured via 256 electrodes implanted in the speech-related motor cortex areas.
- **Evaluation:** Performance is measured by Word Error Rate (WER) on 1,200 held-out sentences. The competition timeline spans from June 2, 2023, to January 14, 2027.

3.3 SOTA Approaches and Leaderboard Analysis

Current methodologies on the 50-word vocabulary leaderboard [11] demonstrate rapid progress:

- **Baseline (Willett et al.):** Decodes neural activity into phoneme probabilities, followed by an n-gram language model (RNN + 5-gram + OPT).
  - 9.1% WER (50-word vocabulary); 23.8% WER (125,000-word vocabulary).
  - Decoding speed: 62 words per minute.
- **Innerspeech:** RNN-transformer with 3-gram rescoring (WER 10.08).
- **Stanford Silent Speech:** Fine-tuned LISA (LLM Integrated Scoring Adjustment) on an ensemble of 10x LSTM with 5-gram beam search (WER 8.93).
- **Okubo Lab (CIBR, Beijing):** RNN Decoder with bidirectional GRU (WER 8.26).
- **UCLA NECL:** Causal Transformer combined with Llama 8B (WER 5.68).
- **BraIn-to-Text (BIT):** Currently leading with a **WER of 5.10**.

4 Brain–Computer Interface (BCI) with Non-Invasive EEG

Non-invasive Brain-Computer Interfaces (BCI) leveraging Electroencephalography (EEG) represent a primary research frontier due to their safety and accessibility. This section explores the modalities, inherent challenges, and the evolution of decoding strategies aimed at high-fidelity speech reconstruction.

4.1 Neural Speech Modalities

EEG-based speech research categorizes neural activity into three primary paradigms:

- **Overt/Spoken Speech:** Decoding signals during active vocalization. While providing clear ground truth, it is prone to muscle movement artifacts.
- **Auditory Perception:** Analyzing neural responses to heard speech for reconstruction. Current research often utilizes the *Brennan dataset* for this purpose.
- **Covert/Imagined/Mimed Speech:** The most complex modality where subjects "speak in their mind." This is the core focus for assistive communication but suffers from the lowest Signal-to-Noise Ratio (SNR).

4.2 Technical Challenges and Bottlenecks

The transition from neural spikes to audible speech via EEG is hindered by several critical factors:

1. **Signal-to-Noise Ratio (SNR):** EEG signals are inherently noisy, attenuated by the skull, and heavily contaminated by non-neural artifacts (EOG, EMG).
2. **Generalization Deficit:** Data scarcity in this domain often leads to overfitting, making cross-subject and cross-session generalization difficult.
3. **Temporal Alignment Paradox:** In imagined speech, there is no physical onset or acoustic ground truth to mark timing, creating significant hurdles for end-to-end supervised training.

4.3 Dataset Landscape

Current datasets primarily support classification (e.g., discrete word recognition) rather than full-waveform reconstruction.

- **General Benchmarks:** Repositories such as *speech_decoding* aggregate various speech-related EEG tasks.
- **Imagined Speech Specific: Chisco** is currently the largest open-source dataset for imagined speech. Other emerging resources include **MSEEG** and **VocalMind**.

4.4 Evolution of Neural Decoding Approaches

4.4.1 Traditional Mel-Spectrogram Reconstruction

Conventional methods utilize a multi-stage pipeline: *Raw EEG* $\rightarrow$ *Feature Extraction* $\rightarrow$ *Neural Decoder* $\rightarrow$ *Mel-spectrogram* $\rightarrow$ *Vocoder* $\rightarrow$ *Waveform*.

- **Brain-to-Speech (BTS):** A notable implementation demonstrating synthesis from Korean brain signals, though formal peer-reviewed documentation and public datasets remain pending.
- **NeuroTalk (AAAI 2023):** Addresses the imagined-to-spoken gap by training on spoken EEG (with audio labels) and employing transfer learning to align imagined speech features, allowing reconstruction in the user's own voice [12].

4.4.2 Modern Architectural Breakthroughs (2024–2025)

Recent trends move toward direct synthesis and large-scale generative modeling:

- **EEG-to-Text:** Converting neural signals directly to text prior to synthesis. While effective, current SOTA is largely limited to closed vocabularies, hindering real-world scalability [13].
- **Diffusion-based Generation:** *Brain2Speech Diffusion* (2023) utilizes diffusion models pre-trained on massive audio corpora, fine-tuned on ECoG/EEG signals (e.g., VariaNTS corpus) to generate high-quality, naturalistic speech.
- **Hybrid Fusion (2025):** To solve the alignment challenge, researchers are integrating EEG with sEEG, ECoG, or EMG signals, utilizing *Dynamic Time Warping (DTW)* to synchronize imagined neural sequences with ASR-TTS pipelines [14].

- **Augmentation and Backbone SOTA (2025):** Approaches using *Transformer + VAE* architectures are being deployed to "expand" latent representations and improve noise robustness (e.g., on the Brennan dataset) [15].

4.4.3 Clinical Frontiers in Neuroprosthetics (2025)

Two landmark studies published in 2025 represent the current pinnacle of the field:

- **Streaming Brain-to-Voice (Card et al., Nature Neuroscience):** A real-time system for patients with severe paralysis. It maps neural activity to intended sentences with a latency reduction from 8s to 1s, incorporating voice cloning to restore the patient's pre-injury vocal identity [16].
- **Instantaneous Synthesis (Nature, 2025):** A paradigm shift that bypasses the text-intermediary. By utilizing high-density microelectrodes (256 channels), this system enables *instantaneous* speech synthesis, allowing users to sing short melodies and speak with expressive intonation [17].

5 Method Proposal

Based on the current state-of-the-art and the availability of high-quality datasets, the first year of my PhD will focus on two tracks. The primary objective is to develop robust neural-to-speech architectures that leverage the linguistic power of Large Language Models (LLMs) to achieve benchmark-breaking performance.

5.1 Track 1: Advancing Silent Speech Interfaces (SSI) via EMG

The immediate goal is to outperform the *MONA-LISA* framework on both Vocalized and Silent EMG benchmarks. This track capitalizes on established baselines and publicly available datasets (Gaddy, 2020).

The research will implement an **End-to-End LLM-based Signal Processing** pipeline. A key innovation involves the development of a specialized **Neural-to-Speech Representation Mapper** designed to project raw EMG features directly into the latent space of a pre-trained speech model. This approach is an evolution of the methodology successfully demonstrated in my recent work, *ChatterInstruct*, which utilized similar cross-modal alignment techniques. By refining the mapper to handle the specific stochastic nature of facial EMG:

5.2 Track 2: Invasive BCI Decoding via the Utah Array Challenge

The second track involves participation in the **Brain-to-Text Benchmark 24** leaderboard which provides a rigorous environment to track research progress and validate models against global competitors.

My target is to surpass the current SOTA held by *UCLA NECL* (WER 5.68) which utilizes a Causal Transformer combined with Llama 8B

6 Research Community

- **Leading Scholars in Biosignal Processing:** Prof. Tanja Schultz and Kevin Scheck (University of Bremen); Prof. Gopala Anumanchipalli and Peter Wu (UC Berkeley); Prof. Satoshi

Nakamura (NAIST); Prof. Alan W Black (CMU).

- **Neural and Clinical Engineering:** Prof. Haizhou Li and Dr. Siqi Cai (National University of Singapore); Prof. Prasanta Kumar Ghosh (IISc Bangalore); Prof. Carol Espy-Wilson and Dr. Yashish Siriwardena (University of Maryland).
- **Clinical and Behavioral Integration:** Dr. Melinda Chang (Ophthalmology, USC Keck); Dr. Adam Frank (Psychiatry, USC Keck); Dr. Adela Timmons (Clinical Science, UT Austin); Prof. Theodora Chaspari (Computer Science, CU Boulder).
- **Institutional Centers:** The **Brain and Creativity Institute (BCI)** directed by Profs. Hanna and Antonio Damasio (USC Dornsife), focusing on the intersection of neural activity and human expression.

7 Research Networks

- **Cognitive Systems Lab (CSL) and ABUR Lab:** Pioneer centers in Electromyography (EMG) and biosignal-enabled spoken communication, led by Prof. Tanja Schultz and teams at University of Groningen (`https://aburlab.web.rug.nl/`).
- **USC SAIL Lab (BaCaTeC):** A collaborative hub between USC and the Technical University of Munich, focusing on EMG-to-speech projects and precision psychiatry. This research links biosignals to health-relevant biological pathways for detecting disorders like depression and schizophrenia.
- **Stanford Neurosciences PhD Program:** Focused on high-fidelity brain-to-text systems, recently demonstrated in the Nature (2023) high-performance neuroprosthesis studies.
- **Shlizerman Lab (University of Washington):** Developers of the DCoND-LIFT framework, currently holding top positions in the Brain-to-Text Benchmark 2024.
- **CIBR-Okubo Lab (Beijing):** A leading center in neural sequence decoding, placing Top 2 in the Brain-to-Text Challenge 2024 via advanced RNN-based architectures (`https://github.com/CIBR-Okubo-Lab`).
- **Neuroprosthetics Lab:** Open-source hub for clinical-grade neural decoding repositories and hardware-software integration (`https://github.com/Neuroprosthetics-Lab`).

References

[1] Anonymous Authors, "ChatterInstruct: Leveraging LLMs for Neural Signal Instruction Tuning," 2025. [Online]. Available: `https://anonymous-research-oss.github.io/chatterinstruct-demo/`

[2] UCLA NECL, "Causal Transformer and Llama 8B for Neural Decoding," Brain-to-Text Benchmark 2024.

[3] "Innerspeech Toolkit: Open-source tools for Biosignal-to-Speech," 2024. [Online Source referenced in text].

[4] T. Schultz, "Biosignal-Enabled Spoken Communication," in *Proc. IbPRIA/IberSPEECH*, 2018. [Online]. Available: `https://www.isca-archive.org/iberspeech_2018/schultz18_iberspeech.html`

[5] D. Gaddy and D. Klein, "Digital Voicing of Silent Speech," in *Proc. EMNLP*, 2020. https://aclanthology.org/2020.emnlp-main.445.pdf

[6] D. Gaddy, "Improved Transfer Learning for EMG-based Speech Reconstruction," in *Proc. ACL-IJCNLP*, 2021. https://aclanthology.org/2021.acl-short.23.pdf

[7] D. Gaddy, "Communication Interfaces using Electromyography," Ph.D. dissertation, UC Berkeley, 2022.

[8] D. Gaddy and D. Klein, "EMG-to-Speech Dataset," Zenodo, 2020. https://zenodo.org/records/4064409

[9] X. Chen et al., "A Cross-Modal Approach to Silent Speech with LLM-Enhanced Recognition," *arXiv preprint*, 2024.

[10] F. R. Willett et al., "A high-performance speech neuroprosthesis," *Nature*, vol. 620, pp. 1031-1036, 2023. https://www.nature.com/articles/s41586-023-06377-x

[11] "Brain-to-Text Benchmark 2024," EvalAI Challenge. https://eval.ai/web/challenges/challenge-page/2099/leaderboard/4944

[12] Y. Lee et al., "NeuroTalk: Towards Voice Reconstruction from EEG during Imagined Speech," in *Proc. AAAI*, 2023. https://arxiv.org/abs/2301.07173

[13] "EEG-To-Text Decoding with Closed Vocabulary," *arXiv preprint*, 2024. https://arxiv.org/abs/2405.06459

[14] "Fusion Approach for sEEG and EMG in Speech Decoding," *arXiv preprint*, 2025. https://arxiv.org/abs/2512.22146

[15] "Decoding EEG Speech Perception with Transformers and VAE," *arXiv preprint*, 2025.

[16] S. L. Card et al., "A streaming brain-to-voice neuroprosthesis to restore naturalistic communication," *Nature Neuroscience*, March 2025. https://www.nature.com/articles/s41593-025-01905-6

[17] "An instantaneous voice-synthesis neuroprosthesis," *Nature*, vol. 638, 2025. https://www.nature.com/articles/s41586-025-09127-3