



Enhancing Syllabic Recognition via Speech-EEG Phase Analysis and Non-Activity State Modeling

Rini Sharon¹, Hema Murthy¹

¹Department of Computer Science, IIT Madras, India

evanjilina@gmail.com, hema@cse.iitm.ac.in

Abstract

Decoding inner speech from brain activity holds transformative promise for brain-computer interfaces and speech rehabilitation. This study examines EEG-based neural patterns across speech production, imagination, and perception. Signal-level visualizations reveal distinct, phase-specific cortical signatures that emerge within precise temporal windows in spatially localized regions. These insights motivate the introduction of a non-activity(NA) state to account for resting and transitional EEG periods—akin to pauses in natural speech. Capturing and excluding these “cognitive pauses” de-noises the EEG signal, improving the focus on task-relevant neural activity. Comparative evaluation across continuous and isolated speech datasets demonstrates that NA modeling improves syllabic recognition accuracy by 1.8% and 1.15%, respectively. These results underscore the role and generalizability of cognitive pauses in imagined speech-EEG decoding, enabling robust brain computer interfaces.

Index Terms: Speech recognition, EEG, imagined speech, non-activity state modeling, brain-computer interface

1. Introduction

Understanding the cortical processes involved in speech production, imagined speech, and auditory processing is vital for developing brain-computer interfaces (BCIs) aimed at restoring communication in individuals with severe speech impairments, such as locked-in syndrome [1, 2, 3]. Electroencephalography (EEG), with its high temporal resolution and non-invasiveness, is a leading neuroimaging tool for capturing speech-related neural activity [4, 5, 6]. Recent work has demonstrated the feasibility of speech synthesis directly from neural signals, showing progress towards real-time BCI speech systems [7, 8].

Despite this potential, imagined speech recognition remains a significant challenge, primarily due to the low signal-to-noise ratio and inter-subject variability [9, 5] in EEG signals during covert speech processes [3, 2]. Early studies focused on classifying a limited set of imagined words, achieving modest accuracies; however, performance levels have consistently lagged behind those obtained for overt speech decoding [4, 9, 10, 11].

Recent advances in multi-phase frameworks leveraging correlations between imagined, heard, and spoken speech phases have shown promise in improving classification performance [12, 13, 14]. Additionally, deep learning techniques and phase-aware feature extraction networks, such as WMCnet, have improved the discriminability of imagined speech features, particularly at the syllable level [12, 5, 7]. Nevertheless, decoding continuous imagined speech—where boundaries between syllables are not explicitly provided—remains particularly challenging, with sentence-level decoding accuracies typically below

30% [9, 3, 15]. Approaches like hidden Markov models (HMM) and EEG-based voice activity detection (VAD) have been employed to segment continuous imagined speech streams, improving alignment with neural activity and boosting classification performance [16, 5, 17]. More recently, [18] proposed a parallel decoding framework capable of simultaneously reconstructing listened speech waveforms and predicting phoneme sequences from EEG embeddings.

A notable advancement is the multi-phase correlation approach introduced in [12], which jointly models the hearing, production, and imagination phases to exploit cross-phase dependencies. This method significantly improved syllable-level classification, achieving up to 18.2% accuracy gain on imagined speech trials by leveraging phase relationships. Similar phase-wise decoding approaches have been extended to intracranial data, demonstrating the potential to bridge non-invasive and invasive decoding strategies [19].

While the existing multi-phase framework has demonstrated effectiveness in prior work, it assumes relatively clean phase transitions and continuous activity during speech production and imagination. In practice, non-activity(NA) periods (e.g., pauses before, after, or between syllables) are common in both continuous and isolated speech paradigms, often leading to misalignment between neural activity and speech labels [16]. These NA states can significantly impact classification accuracy if not properly accounted for.

This paper introduces an extension to the multi-phase EEG decoding framework proposed in [12] by explicitly modeling NA states. We incorporate NA detection and labeling into the classification pipeline, thereby refining the temporal alignment between EEG segments and speech labels. This modification substantially improves decoding accuracy, especially in continuous speech, where pauses and transitions are more frequent.

The primary contributions of this paper are as follows:

1. **Non-Activity (NA) State Modeling:** We introduce an NA state modeling approach for EEG-based multi-phase speech decoding to handle inactivity intervals between syllables, improving temporal alignment and enhancing decoding performance in both continuous and isolated speech tasks.
2. **Comparison of Continuous vs. Isolated Speech EEG:** We present the first systematic evaluation of NA state modeling across continuous and isolated speech EEG datasets.
3. **Phase-Specific EEG Visualization:** We analyze Event Related Potential-based EEG responses across phases, revealing phase-dependent neural patterns and pre-onset cue effects.
4. **Psycho-Acoustic Analysis:** We investigate subject-level differences in speaking-imagination rate coupling, prosodic consistency, and fatigue, providing behavioral context for EEG variability.

2. Data Collection and Databases

This study makes use of two pre-existing EEG datasets involving speech tasks: one capturing continuous speech and the other focusing on isolated speech units such as syllables and words. Dataset 1 (DS1) originates from the publicly available KARA ONE database, while Dataset 2 (DS2) was compiled at IIT Madras, consisting of continuous speech recordings. To ensure robust analysis, performance metrics were assessed both at the individual subject level and across multiple sessions.

2.1. Dataset 1: KARA ONE Database

DS1 consists of EEG recordings from 14 participants using a 64-channel Neuroscan cap (1 kHz sampling rate, 10-20 system) [11]. Participants minimized movements while responding to four word prompts ("pat," "pot," "knew," "gnaw") and seven syllables ("iy," "uw," "piy," "tiy," "diy," "m," "n"), each repeated 12 times. Data collection followed three phases: In the Articulation Phase (A1), participants were presented with a text prompt on a screen while simultaneously hearing its corresponding audio. They were given two seconds to prepare for articulation. The Imagined Speech Phase (I1) required participants to subvocalize the speech prompt for five seconds without any overt articulation. Finally, in the Overt Speech Phase (S1), participants verbally articulated the prompts while EEG signals recorded neural activity.

2.2. Dataset 2: Continuous Speech Corpus

DS2 was collected at IIT Madras [20]. Nine participants (ages 25-34, 2:1 male-to-female ratio) recorded in an acoustically isolated chamber using a 128-channel HydroCel Geodesic Sensor Net (250 Hz sampling rate, referenced to Cz). Pre-recorded conversational phrases from male and female non-native English speakers were played randomly. Each subject completed 1-2 sessions with 25 unique sentences, derived from the TIMIT database, played once per trial (3 trials per session). DS2 structured data collection into three phases. During the Listening Phase (H2), participants passively listened to the speech stimuli to focus on auditory comprehension. In the Internalization Phase (I2), they were required to mentally rehearse or subvocalize the prompt. Finally, in the Verbalization Phase (S2), participants repeated the presented sentences aloud while ensuring minimal head movement. To account for individual variation, a manual mouse-click mechanism marked completion of internalized and spoken responses, and spoken audio was recorded for segmentation. DS2 constructed phrases from DS1's syllabic components, with example sentences like "Don't rag me," "Carry your suit," and "Ask like that." Resting-state EEG was recorded before and after each session for baseline comparison.

3. General Experimental Set-up

We adopt the state-of-the-art multi-phase architecture from [12], which leverages correlations between phases, for robust syllable-level EEG classification, in both continuous and isolated speech. Our experimental setup builds upon this established architecture, with modifications introduced to improve temporal precision, and to better align EEG data with speech events. The details of the original framework, along with the novel modifications, are described below and in Algorithm 1.

3.1. Existing Framework Overview

The EEG classification pipeline builds upon prior multi-phase frameworks [12], comprising three stages: pre-processing, feature extraction, and classification. EEG data is fil-

Algorithm 1 EEG-Based Multi-Phase Speech Classification

- 1: **Input:** Raw EEG data $X \in \mathbb{R}^{T \times C}$, speech phases (Rest, Hear, Imagine, Speak), channel subset C
 - 2: **Output:** Predicted class label y
 - 3: **Pre-processing:**
 - 4: Band-pass filter X (0.3–60 Hz), apply 50 Hz notch filter
 - 5: Common average referencing, DC mean removal
 - 6: Ocular and muscle artifact suppression
 - 7: Select speech-relevant channels (10 for DS1, 36 for DS2)
 - 8: Segment X into task phases based on trial markers
 - 9: **Primary Feature Extraction:**
 - 10: **for** each $c \in S$ **do**
 - 11: Window EEG phase data into overlapping segments (50% overlap, window size W)
 - 12: Extract 63-dimensional handcrafted feature vector per window (spectral, temporal, statistical)
 - 13: Concatenate window features to form channel vector
 - 14: **end for**
 - 15: **Secondary Feature Extraction (WMCnet):**
 - 16: Input per-channel feature for phases into WMCnet
 - 17: Project features onto a common latent representation (CL)
 - 18: Reconstruct phase-specific features from the CL
 - 19: Extract discriminative secondary features for classification
 - 20: **Label Refinement and Non-Activity Modeling:**
 - 21: Introduce additional NA label for inactivity intervals
 - 22: Propagate NA labels across phases (Hear, Imagine, Speak)
 - 23: Refine temporal boundaries for label alignment
 - 24: **Classification:**
 - 25: Train GMM-HMM and DNN hybrid model
 - 26: During inference, obtain per-channel predicted labels
 - 27: Aggregate per-channel predictions using plurality voting
 - 28: Assign final class label=0
-

tered, artifact-suppressed, channel-selected, and segmented into speech phases [11, 21]. Primary features are handcrafted, while secondary features are derived using WMCnet, exploiting phase relationships. Final syllable-level classification is performed using a GMM-HMM and DNN hybrid system in Kaldi [22], with plurality voting across channels. During training, WMCnet models the correlation between imagined and support phases; during testing, it projects imagined speech features onto a shared latent space for improved decoding.

3.2. Novel Modifications to the Framework

The following modifications were made to the existing architecture to support experiments in this work:

Non-Activity State Integration: We introduced NA states into the labeling scheme to represent periods of silence before, after, and between syllables. These states were treated as additional classes during classification, enabling the system to differentiate between active speech processing and background neural activity. This approach accounts for natural pauses and prevents the classifier from assigning incorrect syllables to silent intervals, thus reducing misclassification errors due to misalignment between EEG activity and speech prompts.

EEG-Based Voice Activity Detection (VAD) in continuous speech data: To support NA state detection in continuous speech data DS2, we incorporated an EEG-based VAD module, inspired by [16]. This module identifies regions of silence in the EEG signal corresponding to pauses in speech. Sentence

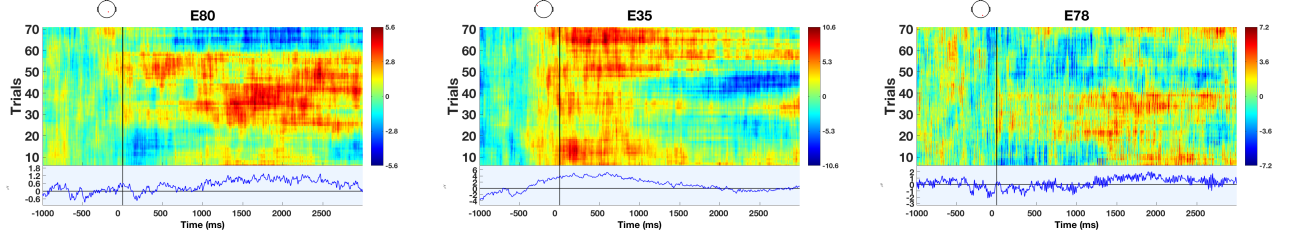


Figure (1): Signal-level visualization of EEG responses across the three phases: Left to Right: (a) Production, (b) Imagination, and (c) Audition. Each subplot displays the trial-level ERPI and the raw EEG time-domain average for multiple trials.

boundary markers were combined with EEG-derived silence intervals to refine the segmentation process, ensuring that NA periods were explicitly labeled and excluded from syllable decoding.

Improved Label Alignment in isolated speech data: For isolated speech data in DS1, precise alignment between EEG and syllables was achieved by comparing the recorded speech audios with EEG data for production. For articulation and imagination phases, a flat-start alignment approach was first employed, treating the entire segment as a single block. This was followed by common word template matching to refine the alignment between EEG time segments and syllable-level speech activity[21]. This two-step alignment process reduced timing mismatches between EEG data and phoneme boundaries.

Label Propagation Across Phases: The NA state labels were propagated across all three phases, ensuring that non-activity intervals were consistently represented in each phase. This ensured that the WMCnet received cleaner phase-wise inputs and labels during training, leading to improved cross-phase correlation learning.

These modifications enable more precise handling of EEG variability during natural speech production and imagination tasks. This approach improves the robustness of classification, particularly for continuous speech, and reduces the impact of irrelevant intervals and phase misalignment, which are often overlooked in conventional EEG-based speech decoding studies.

4. Results and Discussion

4.1. Temporal Analysis of EEG Responses Across Phases

Understanding the temporal evolution of cortical activity across speech phases is crucial for decoding EEG patterns associated with speech production, imagination, and auditory perception. EEG data from DS2 recorded using the 10-20 electrode system was analyzed across three key phases: (i) Production (overt articulation of prompts), (ii) Imagination (imagining speaking the prompts), and (iii) Audition (passive listening of prompts). Data was epoched from -1000 ms to +3000 ms relative to task onset, with ERPI (Event-Related Potential Image) visualizations generated by trial-sorting and smoothing (Figure 1). Analysis focused on three representative electrodes corresponding to phase-specific cortical regions: E80 (vertex/Cz – sensorimotor cortex, relevant to articulation), E35 (frontal-temporal – associated with Broca’s area, relevant to imagination), and E78 (temporal – auditory cortex, relevant to listening).

Production Phase: EEG patterns during production exhibit high-amplitude, sustained activity from 0 to 1500 ms post-onset, reflecting motor cortex activation as jaw, tongue, and vocal tract movements are executed. Raw EEG signals show a pronounced positive shift, though dominated by low-frequency

muscle artifacts, particularly over Cz-region electrodes, aligning with the sensorimotor cortex’s role in controlling speech articulators during overt speech execution.

Imagination Phase: The ERPI reveals moderate sustained activity post-onset, similar in temporal profile to production but with lower amplitude. The raw EEG signal shows a positive shift post-onset, devoid of muscle artifacts suggesting that imagination engages overlapping motor and auditory networks, reflecting inner articulation and auditory imagery/simulation.

Audition Phase: Passive listening results in low-amplitude auditory evoked responses, with N1 (≈ 100 ms) and P2 (≈ 200 ms) peaks, characteristic of primary auditory cortex activation over temporal regions[23, 24]. The raw EEG displays these transient deflections, reflecting primary auditory cortex activation.

Pre-onset ERP similarities across production and imagination phases reflect auditory cue effects, while gaps before, after, and between speech units suggest cognitive pauses. These NA states, if ignored, cause misalignment between EEG activity and speech labels, reducing decoding accuracy. Signal-level visualization thus reveals distinct neural patterns across phases and highlights the need to model NA intervals for accurate speech EEG decoding.

4.2. Non-activity state Modelling

ERPI analysis reveals that speech-related EEG involves transient bursts of task-specific cortical activity interspersed with non-active periods, likely corresponding to resting, preparation, or planning intervals. This aligns with speech production studies, where pauses between phonetic units support motor planning and linguistic formulation. Analogously, NA states in EEG reflect cognitive pauses between speech units. Ignoring such NA states, risks contaminating task-related feature extraction with resting background noise, which deteriorates model performance and class separability. Sharp ERP transitions at task onset further underscore the need to model task boundaries and rest intervals.

To address this, NA states were explicitly modeled as separate labels(Section 3.2), distinguishing activity from non-activity portions in EEG data. Figure 3 shows that NA modeling improved decoding performance across both datasets, with an absolute average accuracy gain of 1.8% over the state-of-the-art model. DS1 exhibited larger but more variable improvements (mean 2.13%, SD 0.7%), suggesting alignment refinements benefited certain trials more. In contrast, continuous speech DS2 showed smaller but consistent gains (mean 1.47%, SD 0.25%), reflecting its greater dependence on pause detection for accurate segmentation. These results reinforce the need to accommodate cognitive pauses (NA states) in EEG-based speech decoding, helping separate task-relevant neural patterns from NA regions and enhancing task discrimination accuracy.

4.3. Psycho-acoustic Analysis

The primary objective of this study has been to investigate various phases of speech-induced EEG signals to find the associations between cognition of phonological segments, and the acoustic events themselves. A more nuanced understanding of these associations requires an examination of subject-specific psycho-acoustic cues. These offer insights into how individual differences in perception, imagination, and production of speech might affect neural processing patterns. For this purpose, we analyzed the following subject-dependent measures in DS2 (capturing temporal variations specific to each subject).

Imagination and Speaking Rate: In DS2, participants indicated the time taken to speak or imagine a given phrase by manually registering the completion of their task via a mouse click. Analyzing the duration data on a per-subject basis across all trials revealed a strong within-subject consistency in speaking and imagination rates. Specifically, subjects with faster production rates tended to exhibit shorter imagination durations (Figure 2). This observation suggests a tightly coupled temporal mapping between covert imagination and overt production, aligning with existing evidence on sensorimotor integration in speech preparation [25, 26, 27]. Further stratification by gender revealed that female participants, on average, exhibited higher speaking and imagination rates compared to their male counterparts. Specifically, the mean production rate for females was 5.4 syllables per second (SD = 0.7), while for males, it was 4.9 syllables per second (SD = 0.6). Similarly, the average imagination duration was shorter for females (M = 2.1s, SD = 0.4) compared to males (M = 2.5s, SD = 0.5). These findings align with prior work on gender differences in speech rates [28, 29].

Repeatable Prosody: The audio stimuli presented to participants were recorded by different speakers, resulting in variations in prosody (intonation, rhythm, and stress patterns) and timing characteristics. Since subjects were instructed to memorize and repeat the auditory prompts verbatim, it was hypothesized that their produced speech would align closely with the prosodic features of the input speech. This aligns with "echoic memory" and auditory-motor coupling, where short-term memory retains prosodic contours for accurate speech reproduction [30, 31, 32].

To quantitatively assess alignment, Dynamic Time Warping (DTW) distance was used as a similarity metric between input speech and subjects' spoken outputs, accounting for non-linear temporal variations [33]. Analysis revealed that in 73% of trials, the spoken speech following the auditory prompt exhibited the minimum DTW distance to the input prompt compared to other available input recordings for the same text content.

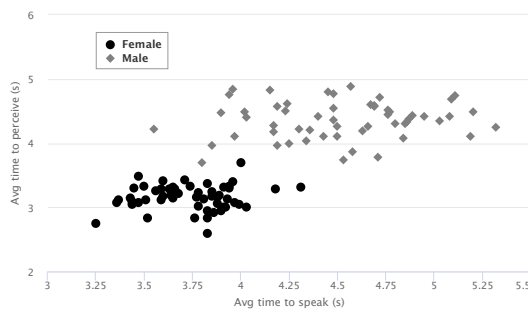


Figure (2): Plot of Production Vs Imagination duration for a subset of 2 male and 2 female subjects in DS2 across trials

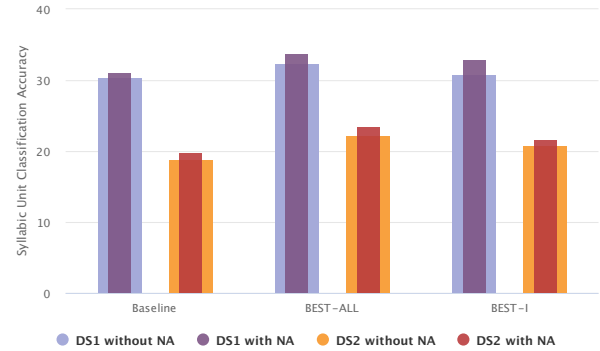


Figure (3): Effect of introducing NA modeling. Baseline: Primary features only. BEST-ALL: All support phases used for decoding. BEST-I: Imagined phase used alone. This terminology is borrowed from [12]

This result suggests that subjects were not merely reconstructing the linguistic content but were retaining and reproducing prosodic patterns from the preceding auditory stimulus. Trials with high DTW distances (>1.5 normalized units) occurred predominantly later in the session, likely reflecting cognitive fatigue or lapses in auditory attention, consistent with vigilance decline in prolonged speech tasks [34, 35].

Performance Degradation Over Time: Subject feedback collected post-experiment indicated a noticeable decline in concentration as the session progressed, particularly in the latter half with repeated prompts. This self-reported fatigue aligned with both behavioral and neural performance measures.

Task accuracy decreased at a rate of -2.13% per trial block, reflecting a gradual performance decline, consistent with cognitive fatigue literature showing prolonged tasks impair speech consistency and working memory capacity [35, 36]. Completion times increased by 9.4% in the final quarter compared to the initial quarter, with speech production rates slowing from 5.2 syllables per second to 4.8 syllables per second.

These psycho-acoustic patterns contextualize EEG variability across subjects, suggesting that future decoding models could integrate such behavioral indices to improve classification accuracy and reduce subject-specific noise, as proposed in hybrid speech-BCI approaches [37, 38].

5. Conclusion

This paper advanced EEG-based speech decoding by introducing non-activity state modeling to handle transient intervals before, between, and after syllables, improving temporal alignment between EEG activity and speech events. Evaluation across continuous and isolated speech tasks demonstrated absolute accuracy gains of 1.8% and 1.15%, respectively, highlighting the wide applicability of NA modeling. Phase-specific EEG visualizations revealed distinct temporal patterns in neural activity across production, imagination, and auditory phases, while subject-level psycho-acoustic insights on speech-imagination rate coupling, prosodic consistency, and fatigue emphasized the need to integrate behavioral variability into decoding pipelines. Future work could incorporate behavioral modeling into decoding pipelines to further personalize and improve classification performance. Together, these findings enhance the robustness of imagined speech recognition systems, contributing towards more reliable BCI applications for speech restoration and communication.

6. References

- [1] G. Schalk and J. R. Wolpaw, "Brain-computer interfaces for communication and control," *Handbook of Clinical Neurology*, vol. 86, pp. 67–85, 2008.
- [2] J. S. Brumberg, K. M. Pitt, J. D. Burnison, and M. Fried-Oken, "Brain-computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2010.
- [3] E. Cooney, L. V. Hiscox, I. M. Wiggins, and D. Sharma, "Imagined speech decoding: a review of current progress and future directions," *Frontiers in Neuroscience*, vol. 16, p. 807235, 2022.
- [4] F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, C. Jeunet, and A. Lécuyer, "A review of classification algorithms for eeg-based brain-computer interfaces: a 10-year update," *Journal of Neural Engineering*, vol. 15, no. 3, p. 031005, 2018.
- [5] C. Herff, L. Diener, E. Mugler, M. C. Tate, M. Goldrick, D. J. Krusienski, and T. Schultz, "Eeg-based speech recognition: From single spoken words to continuous speech," *Frontiers in Human Neuroscience*, vol. 13, p. 126, 2019.
- [6] D. Dash, A. Wisler, M. Sharma, D. Chander, A. Dutta, S. Sharma, R. Mishra, and U. Rajendra Acharya, "A review on eeg-based speech classification with insights on deep learning models," *Biomedical Signal Processing and Control*, vol. 59, p. 101920, 2020.
- [7] G. K. Anumanchipalli, J. Chartier, and E. F. Chang, "Speech synthesis from neural decoding of spoken sentences," *Nature*, vol. 568, pp. 493–498, 2019.
- [8] D. A. Moses, S. L. Metzger, J. R. Liu, G. K. Anumanchipalli, J. Chartier, R. S. Rikhye, U. Krishna, M. E. Dougherty, P. M. Liu, and E. F. Chang, "Neuroprosthesis for speech decoding from brain signals," *Nature*, vol. 620, no. 7973, pp. 120–127, 2023.
- [9] Y. Deng, Q. Gao, and Z. Zhou, "Eeg-based decoding of imagined speech and its application to bcis: A systematic review," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 482–493, 2021.
- [10] J. da Silva Souza, C. Fernandes, A. Tomé, C. E. Castro, M. Santana, and G. Pires, "Recognition of imagined speech using eeg and deep learning," in *2017 International Conference on Intelligent Systems (IS)*. IEEE, 2017, pp. 863–868.
- [11] S. Zhao and F. Rudzicz, "Classifying phonological categories in imagined and articulated speech using eeg," *Proceedings of Inter-speech*, pp. 2430–2434, 2015.
- [12] S. Rini *et al.*, "Multi-phase neural representation learning for imagined speech classification," *Journal of Signal Processing*, vol. 180, p. 107910, 2021.
- [13] B. Krishna, R. K. Reddy, and S. Jayaraman, "Speech imagery recognition using eeg with capsule networks," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2020, pp. 3598–3601.
- [14] R. Gonzalez, D. Nguyen, H. Gu, and D. Wu, "Decoding inner speech from eeg: A deep learning approach to human-computer interaction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 31, pp. 1392–1401, 2023.
- [15] B. Krishna, R. K. Reddy, and S. Jayaraman, "Deep learning approaches for decoding imagined speech eeg signals: a review," *Biomedical Engineering Letters*, vol. 12, no. 2, pp. 149–178, 2022.
- [16] A. Sharon, G. Balaji, M. Sharma, N. C. Lal, and A. Ramakrishnan, "Sound production from eeg: a methodology for mapping imagined speech to speech production," *Biomedical Signal Processing and Control*, vol. 57, p. 101770, 2020.
- [17] A. Kapur and P. M. Kapur, "Alterego: A personalized wearable silent speech interface," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2018.
- [18] J. Lee, J. Han, S. Poria, D. Chakraborty, and A. Das, "Eeg-based parallel decoding of listened speech into speech waveforms and phoneme sequences," *arXiv preprint arXiv:2501.04844*, 2025.
- [19] M. Angrick, C. Herff, G. Johnson, J. J. Shih, D. Krusienski, and T. Schultz, "Interpretable deep learning for decoding and reconstructing speech from the human brain," *Frontiers in Human Neuroscience*, vol. 15, p. 633811, 2021.
- [20] R. A. Sharon, S. Narayanan, M. Sur., and H. A. Murthy, "Neural speech decoding during audition, imagination and production," *IEEE Access*, pp. 1–1, 2020.
- [21] R. A. Sharon, S. Narayanan, M. Sur, and H. A. Murthy, "An empirical study of speech processing in the brain by analyzing the temporal syllable structure in speech-input induced eeg," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 4090–4094.
- [22] I. Kipyatkova and A. Karpov, "Dnn-based acoustic modeling for russian speech recognition using kaldi," in *International Conference on Speech and Computer*. Springer, 2016, pp. 246–253.
- [23] R. Näätänen and T. Picton, "The n1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure," *Psychophysiology*, vol. 24, no. 4, pp. 375–425, 1987.
- [24] T. W. Picton, *Human auditory evoked potentials*. Plural Publishing, 2010.
- [25] R. Vargas, F. Pineda, J. Ruiz, and G. Castellanos-Dominguez, "Eeg-based classification of speech imagery using long short-term memory neural network," *Sensors*, vol. 20, no. 19, p. 5794, 2020.
- [26] J. Derix, O. Iljina, A. Schulze-Bonhage, A. Aertsen, and T. Ball, "Electrocorticographic evidence that humans can adopt speech synthesis strategies," *Journal of Neuroscience*, vol. 32, no. 16, pp. 5658–5668, 2012.
- [27] P. Bashivan, I. Rish, S. Heisig, and C. Boulay, "Single trial analysis of eeg during perception of known and unknown objects," *Frontiers in Neuroscience*, vol. 9, p. 100, 2015.
- [28] D. Byrd, "Sex differences in the average number of syllables produced per breath group: adult versus child speakers," *Journal of Phonetics*, vol. 23, no. 1-2, pp. 1–14, 1995.
- [29] S. P. Whiteside, "Biological sex and speech processing," *Language and Linguistics Compass*, vol. 1, no. 2, pp. 122–138, 2004.
- [30] A. Baddeley, "Working memory: theories, models, and controversies," *Annual Review of Psychology*, vol. 63, pp. 1–29, 2012.
- [31] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.
- [32] W. J. Levelt, *Producing speech: The processing of speech production*. MIT Press, 1999.
- [33] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 26, no. 1, pp. 43–49, 1978.
- [34] R. Patel and K. W. Schell, "Effects of fatigue on speech production," *Human Factors*, vol. 53, no. 5, pp. 466–477, 2011.
- [35] M. A. Boksem and M. Tops, "Mental fatigue: Costs and benefits," *Brain Research Reviews*, vol. 59, no. 1, pp. 125–139, 2008.
- [36] M. M. Lorist, M. Klein, S. Nieuwenhuis, R. De Jong, and G. Mulder, "Mental fatigue and task control: planning and preparation," *Psychophysiology*, vol. 42, no. 3, pp. 273–282, 2005.
- [37] J. S. Brumberg, K. M. Pitt, J. D. Burnison, and J. E. Drinkwater, "Brain-computer interfaces for speech communication," *Speech Communication*, vol. 52, no. 4, pp. 367–379, 2011.
- [38] M. D'Zmura, S. Deng, and K. Liao, "A review on speech imagery decoding from eeg signals: Methods, applications and challenges," *Speech Communication*, vol. 106, pp. 49–64, 2018.