

Decoding Spoken Syllables from EEG Data via Contrastive Embedding Alignment

T. Cortecchia¹, P. Volpato², S. Tortora^{2,3}, E. Menegatti², L. Tonin^{2,3}

¹Polytechnic University of Turin, Turin, Italy

²Department of Information Engineering University of Padova, Padova, Italy

³Padova Neuroscience Center, Padova, Italy

E-mail: tommaso.cortecchia@polito.it

ABSTRACT: We explore the potential of electroencephalography (EEG)-based brain-computer interfaces as an alternative communication system for people with speech impairments. Instead of considering single words or phonemes, we exploit the syllabic structure of the Italian language as the fundamental building block of speech. We collected data from 21 participants as they pronounced and syllabified a set of sentences. We then trained a neural network to learn a latent representation of EEG segments through contrastive alignment with embeddings extracted from a pre-trained automatic speech recognition model. We framed the evaluation as an information retrieval task, utilizing a top- k retrieval within the EEG latent space, comparing a global model to subject-specific models. We achieved an average Top-10 accuracy of 37.4% for the multi-subject model and of 38.2% for subject-specific models across a 192-class vocabulary. The proposed framework provides a scalable approach to non-invasive speech decoding, enabling greater expressiveness and speed compared to traditional solutions in assistive communication.

INTRODUCTION

Augmentative and alternative communication (AAC) devices complement or replace speech for individuals with severe speech disorders or paralysis; however, they are currently limited by slow conversational rates of around 12-18 words per minute (WPM), far below the average 125-185 WPM rate of natural speech [1]. Brain-computer interfaces (BCIs) for speech decoding are on a promising path to improve speed and accessibility of such systems. While invasive implants have recently enabled accurate, continuous decoding of attempted and silent speech [2-4], their high cost, the need for surgery and clinical support, and the risk of long-term data degradation hinder widespread adoption of these technologies [5].

BCIs based on electroencephalography (EEG) may offer a viable alternative for a low-cost, non-invasive solution suitable for everyday use. Although the research on this technology has seen a growing interest over the past few years, the past works have focused either on passive tasks

like listening or reading [6-9], lacking the active component of a BCI, or on the production of restricted sets of words or phonemes [10, 11] which do not allow for the composition of complex language. To truly enable fluent and naturalistic communication for the end-user, a system should offer a wide vocabulary (or a way to build that) and an intuitive mapping from the mental task to the decoded part of speech. A simple solution would just be collecting data during the pronunciation of sentences. Zhang et al. collected a large dataset of imagined speech for Chinese sentences [12]. Sato et al. investigated large-scale open-vocabulary decoding with 175+ hours of high-density EEG [13]. Craik et al. decoded continuous and discrete acoustic properties of the sound produced during aloud repetition of sentences [14]. With these methods, they are effectively collecting a large vocabulary, but it is difficult to align neural data to its linguistic counterpart, especially in the case of silent speech.

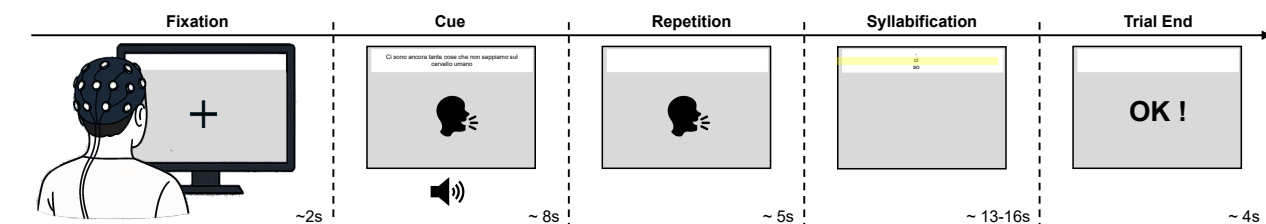
The words of the Italian language, like many other western languages, have the peculiarity of being divisible into smaller units, the syllables. Syllables are small sequences of phonemes that together form a single, distinct sound. Italian counts around 2,700 orthographical syllables (more than 3,700 if we consider phonetics) that contribute to forming more than 150,000 words [15].

In this work, we use the syllabic structure of the Italian language as a building block for a speech BCI. Specifically, we (1) introduce a new protocol that allows for the recording of a broad set of speech units during a natural task, and (2) establish a decoding performance baseline for overt speech decoding by applying state-of-the-art methodologies [9, 16] to the resulting dataset.

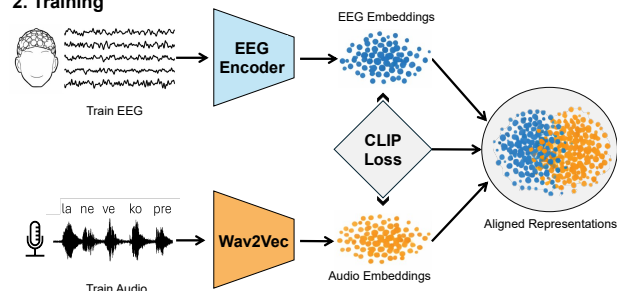
EXPERIMENTAL SETUP

This section introduces the experimental setup, including the recording protocol, the trial structure, and the method used to syllabify the sentences. Although the results we will present in the following sections relate exclusively to the decoding of overt (vocalized) speech, the protocol was designed with the requirements for recording imagined (silent) speech in mind as well. For this reason, we describe the procedures for both conditions here.

1. Data Collection



2. Training



3. Inference

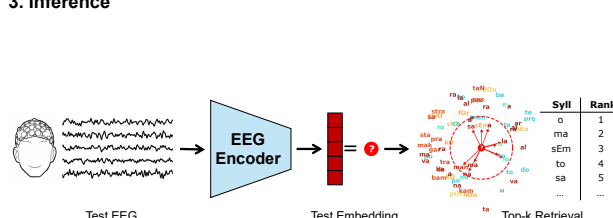


Figure 1: Overview of the proposed framework. (1) *Recording Protocol*: structure of a trial, stage duration is reported below each. Every run consists of 20 repetitions with different sentences. (2) *Cross-modal alignment*: the EEG–audio framework trained to align neural representations with acoustic embeddings using a contrastive objective. (3) *Syllable decoding*: at inference time, acoustic data is discarded, and the target syllable is predicted via a top- k retrieval across the EEG reference set.

When designing the protocol, we kept two key aspects in mind: (1) *Naturalness of the protocol*: Trial-based paradigms work well but are too slow and unnatural, making it difficult to collect a diverse set of classes in a short time. On the other hand, we needed a way to align EEG data with the expected class, especially in the case of silent speech; (2) *Target linguistic unit*: identifying a set of linguistic elements that did not have too many elements (such as words), which would make it impossible to build an effective vocabulary, but also not too elementary (such as letters), because we would come across the same speed limitations as current AAC systems. The perfect trade-off between sample size and production speed can come from dividing words into syllables. Each word allows for the collection of multiple syllables, and the same syllable appears in several different words. Furthermore, by pronouncing segmented words at a measured pace, it is possible to obtain aligned data even in the absence of stimuli. At the same time, syllabification of a sentence is a simple and natural task, and in a single trial it is possible to collect multiple syllables by, for example, syllabifying an entire sentence.

The experiment involved 21 participants (13 male, 8 female), native Italian speakers, each recruited for a single recording session. Each participant performed three runs per condition for the same task. Each run consisted of 20 trials, the same sentences were reused across runs. A total of 60 full trials per condition (20 sentences $\times$ 3 repetitions) was collected this way. A single trial consists of the repetition of the cued sentence, presented both visually on-screen and auditorily via audio playback, that must then be repeated once fluently, at a natural speed, and once performing syllabification, at a cued pace. Each pro-

cedure was repeated twice: first via *overt speech* (vocalized pronunciation), and then via *imagined speech* (mental rehearsal with no acoustic or motor output).

Below we describe in detail each phase of the trial:

- *Fixation (2s)*: A black cross is displayed at the center of the screen. The participant is required to look at the cross and remain still.
- *Sentence presentation (8s)*: The sentence appears in a center-aligned text box while its pronunciation is played using a text-to-speech synthesizer (gTTS, Google). A visual cue representing a person speaking or thinking is displayed. After playback, a 3-s pause allows the subject to finish reading and prepare for repetition.
- *Sentence repetition (5s)*: The sentence is cleared from the screen. The participant then pronounces the sentence at his own natural pace.
- *Syllabification (13-16s)*: After a 1-s pause, the syllabified sentence appears. Syllables scroll upward every 0.6 seconds, with the target highlighted in the middle row. A 3-2-1 dot countdown establishes rhythm, after which the participant articulates the syllables at the given pace.
- *Trial end (4s)*: The text box is cleared, and an on-screen cue signals the end of the trial. A 4-s interval precedes the next trial.

Figure 1 summarizes the trial structure described above. Each run lasted approximately 12 minutes, totaling more than 30 minutes of recording per condition. In addition to EEG data, audio was recorded via microphone during the overt speech block, and we used it to align EEG segments with speech onsets during post-processing. Initial alignment was conducted automatically using Silero

VAD [17] and the Montreal Forced Aligner [18], followed by a manual refinement step using the Praat toolbox [19]. This manual inspection also ensured that any syllables containing pronunciation or timing errors could be accurately corrected or discarded.

All sentences contain a recurrent syntactical structure (subject-verb-complement plus adjectives and articles), such that the effect of sentence length and structure on the recorded brain activity is negligible. Each sentence contains between 11 and 15 words, between 20 and 25 syllables. Syllabification has been automatically processed using phonetic transcriptions from PhonItalia [15], which is also providing the categorization of phonemes according to the International Phonetic Alphabet. The resulting phonetical syllables may differ from the orthographical ones as they reflect the characteristics of articulation and sound, rather than following the typographic rules used in written language (e.g., the orthographical syllabification of the word "pasta" is "pa-sta" while the phonological one is "pas-ta"). Each run allowed for the collection of 462 syllables (192 unique), which enable the construction of up to 22,000 distinct words based on PhonItalia metrics.

In this work, we employed only data related to the syllabification phase, full sentence repetitions were not taken into account. Moreover, the following sections will focus exclusively on the overt speech condition; decoding analyses of the imagined speech data are reserved for future study.

MODEL TRAINING

The high number of syllabic classes, combined with limited training data and the inherently low signal-to-noise ratio of EEG signals, poses a significant challenge for training a deep neural network from scratch. On the other hand, existing models for automatic speech recognition (ASR) have demonstrated high performance in extracting meaningful representations from acoustic data. Therefore, rather than employing a traditional end-to-end classification paradigm, we frame this as a representation learning problem. In this approach, the model learns a mapping from raw EEG data to a structured latent space, specifically aiming to align the embeddings generated by the EEG encoder with those produced by a pre-trained acoustic model. The procedures explained below are an adaptation of the work of [16].

Data processing: Data were recorded with a 64-channel AntNeuro eego system (standard 10-20 montage) at a sampling rate of 2,048 Hz. Channel CPz was used as reference. Frontal channels (AFx, FPx), occipital channels (Ox), and channels M1 and M2 were removed to reduce ocular artifacts and/or due to bad signal quality. The remaining 51 channels cover both the auditory cortex—involved in audio and speech processing—and the primary motor cortex—involved in speech articulation [20, 21]. Preprocessing was kept minimal to comply with real-time closed-loop requirements. The signal was notch

filtered at 50 Hz and band-pass filtered between 1 and 70 Hz. We chose to keep frequencies in the gamma band as they were shown to be effective for phonetic decoding in previous works [22, 23]. Finally, we downsampled the signal to 256 Hz to reduce the computational load.

Both EEG and audio data corresponding to the pronounced syllables were segmented according to the manual annotations. The audio signals were cropped to the exact syllable duration, while the EEG signals were epoched around speech onsets using a fixed window of [0.0, 0.88] s relative to the syllable onset. To extract acoustic features, the audio segments were processed through a pre-trained ASR model. Specifically, we utilized facebook/wav2vec2-large-xlsr-53-italian [24] as a feature extractor, capturing the embeddings from the feature projection layer prior to the transformer blocks. The resulting tensors were averaged across the temporal dimension to obtain a single vector representation for each syllable. For the experimental evaluation, the third run of each subject was reserved as a hold-out test set. The first two runs were split sequentially into training and validation sets using a 90%/10% ratio. To accelerate model convergence, all data underwent channel-wise robust scaling using the training set's median and 10-90% quantiles. We evaluated the architecture on both a multi-subject and a single-subject basis, maintaining consistent data splits across both comparisons to ensure a robust performance baseline.

Architecture: The model is composed of four primary blocks. Data first passes through a *Spatial Attention Layer*, which recombines information across the channel dimension using an attention mechanism; subsequently, a *Subject Layer* applies a participant-specific mapping. These two blocks are described in detail in [16]. In contrast to their approach, we maintain the original channel count during the spatial attention stage rather than enlarging the set. The resulting signals are then processed by an EEGNet-like encoder [25] to extract features from the projected data. The classification head of the standard EEGNet is replaced by a *Temporal Projection* layer (two Conv1d layers interleaved with a BatchNormalization and a GELU activation function). This layer yields a tensor of shape $B \times T \times H$, where B represents the batch size, and T and H match the temporal and hidden dimensions of the Wav2Vec embeddings, respectively. Finally, mirroring the audio embedding process, we average the EEG embeddings across the temporal dimension to obtain a single vector representation per syllable.

Training: The EEG embeddings were aligned to their acoustic counterparts using a contrastive objective inspired by the CLIP loss [26]. This applies a cross-entropy loss over the batch, effectively pulling corresponding EEG and acoustic embeddings together in the shared latent space while pushing non-matching samples apart. Formally, let X be an EEG segment and Y be its corresponding acoustic latent representation (the positive sample). Within a batch, we define $N - 1$ negative acoustic samples, grouping them such that the positive sample

is $\tilde{Y}_N = Y$. The EEG encoder maps X to a latent representation Z . By computing the inner product $\langle \cdot, \cdot \rangle$ followed by a softmax, we train the model using a simplified cross-entropy loss to maximize the probability of selecting the true positive sample:

$$L_{\text{CLIP}} = -\langle Z, Y \rangle + \log \left(\sum_{j'=1}^N e^{\langle Z, \tilde{Y}_{j'} \rangle} \right)$$

The experiments were conducted on a workstation equipped with an AMD Ryzen 7 7700 processor, 16 GB of RAM, and an NVIDIA RTX 3060 GPU with 12 GB of VRAM. The models were trained for a maximum of 1,000 epochs, utilizing an early stopping criterion if the validation loss failed to improve for 75 consecutive epochs. Optimization was performed using the AdamW optimizer with a learning rate of 5×10^{-4} , weight decay of 10^{-4} , and a CosineAnnealing scheduler. Batch size was set to 128. Data and code will be provided in a future release.

RESULTS

We approached the syllable decoding task as an information retrieval problem, effectively implementing a top- k retrieval in the learned latent space. During the evaluation phase, the acoustic modality was completely discarded. Instead, we constructed a reference set exclusively from the EEG embeddings of the training data. For every unseen test sample, we computed its cosine similarity against all individual training embeddings in this reference set, selecting the k nearest candidates to compute our retrieval accuracies. Let $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\}$ be the set of reference embeddings corresponding to our syllable vocabulary, and $\hat{\mathbf{y}}$ be the predicted EEG embedding. The predicted syllable S is given by:

$$S = \operatorname{argmax}_{i \in \{1, \dots, N\}} \frac{\mathbf{v}_i \cdot \hat{\mathbf{y}}}{\|\mathbf{v}_i\| \|\hat{\mathbf{y}}\|}$$

A prediction is considered correct if the target class is present between the first k predictions of the model.

Table 1 reports the per-subject and average decoding accuracies (Top-1 and Top-10) for both multi-subject and single-subject variants. Overall, the single-subject approach generally outperforms the multi-subject model, achieving an average Top-1 accuracy of 9.5% and a Top-10 accuracy of 38.2%, compared to Top-1 of 9.1%, Top-10 of 37.4% of the multi-subject model. Despite the significantly smaller volume of training data available per model in the single-subject configuration, its superior performance underscores the inherent difficulties of integrating multi-subject EEG data due to high inter-subject variability. When evaluated against a ZeroR baseline, performances exceed chance levels (3,9% Top-1, 22,9% Top-10) in nearly all cases.

In the ablation study (Table 2) we evaluated the individual and combined contributions of the *Spatial Attention Layer* and *Subject Layer* against the base architecture. Results highlight a critical role of the *Subject Layer*,

Table 1: Comparison of retrieval accuracy between single- and multi-subject models

Subj	Multi-subject		Single-subject	
	Top-1	Top-10	Top-1	Top-10
<i>k7</i>	14,3%	44,8%	16,5%	47,0%
<i>i5</i>	15,4%	56,4%	15,2%	57,0%
<i>k3</i>	18,3%	54,8%	15,1%	53,9%
<i>i1</i>	13,4%	45,7%	14,1%	45,0%
<i>k4</i>	10,5%	40,2%	14,0%	50,7%
<i>j8</i>	8,1%	35,7%	13,3%	46,2%
<i>c9</i>	9,5%	39,3%	11,9%	45,3%
<i>j4</i>	9,6%	40,8%	11,6%	40,2%
<i>j5</i>	11,6%	43,7%	10,7%	42,8%
<i>i8</i>	10,2%	41,5%	9,6%	39,1%
<i>f9</i>	11,5%	41,2%	9,4%	37,7%
<i>d7</i>	6,5%	37,4%	8,2%	37,0%
<i>h9</i>	7,2%	34,6%	7,8%	36,7%
<i>i7</i>	6,3%	35,4%	7,8%	37,4%
<i>k2</i>	9,5%	39,3%	6,7%	31,5%
<i>k5</i>	5,0%	30,3%	6,1%	22,4%
<i>j3</i>	5,9%	31,7%	5,7%	33,7%
<i>k6</i>	9,5%	28,4%	4,6%	29,9%
<i>j6</i>	2,6%	21,0%	4,3%	25,8%
<i>k1</i>	2,0%	18,0%	4,1%	23,9%
<i>j7</i>	3,8%	24,2%	3,6%	18,0%
Average	9,1%	37,4%	9,5%	38,2%

which delivers a substantial performance boost across both paradigms. In the multi-subject setting, appending the *Subject Layer* to the base model improves Top-1 accuracy by 6.3% and Top-10 accuracy by 22.7%. This shows how adding the layer enables subject-specific processing that facilitates cross-subject data integration. Interestingly, the *Subject Layer* also improves the single-subject model, improving Top-1 accuracy by 1.5% and Top-10 by 5.67%. Although subject-specific processing is already inherent to the single-subject approach, we hypothesize that this layer acts as an additional linear transformation that effectively enhances the model’s capacity for generalization.

The addition of the *Spatial Attention Layer* by itself fails to provide a performance boost over the base model, instead causing a drop in accuracy for both multi-subject (-0.2% Δ Top-1, Top-10 unchanged) and single-subject (-5.3% Δ Top-1, -15.2% Δ Top-10) configurations. However, when coupled with the *Subject Layer*, spatial attention further boosts multi-subject classification performance, allowing it to reach the full model accuracies reported above. While it introduces a slight degradation in the single-subject context—dropping from 9.6% to 9.5% in Top-1 and from 39.5% to 38.1% in Top-10, this difference is marginal. Consequently, we chose to retain it in the final architecture. Its inclusion maintains architectural alignment for direct comparison with previous methodologies [9, 16], while simultaneously providing a mechanism for improved interpretability—such as the inspection of spatial attention maps—in future work.

To obtain a qualitative assessment of the learned latent

Table 2: Model ablations

model	Multi-subject		Single-subject	
	Top 1	Top 10	Top 1	Top 10
<i>Full model</i>	9,1%	37,4%	9,5%	38,2%
+ <i>Subject Layer</i>	8,2%	36,0%	9,6%	39,5%
+ <i>Spatial Attention</i>	1,6%	13,3%	2,9%	18,6%
<i>Base Model</i>	1,8%	13,3%	8,1%	33,8%

space, UMAP was applied to the test embeddings to reduce their dimensionality for 2D visualization. Data points were colored according to a gradient reflecting sonority and manner of articulation. The emerging topology, resulting from the cross-modal alignment, confirms the efficacy of the contrastive training objective. Additionally, it establishes a robust theoretical foundation for zero-shot or few-shot vocabulary expansion without explicit retraining, although this hypothesis remains to be verified. Figure 2 illustrates the learned space of the top four subjects, ranked according to their Top-10 accuracy in the multi-subject model. Notably, while the underlying topological sorting of phonetic features remains robustly consistent across all speakers, the specific geometric shape of the latent manifold exhibits subject-specific variations, highlighting the model’s capacity to capture both global linguistic structures and individual speaker variability.

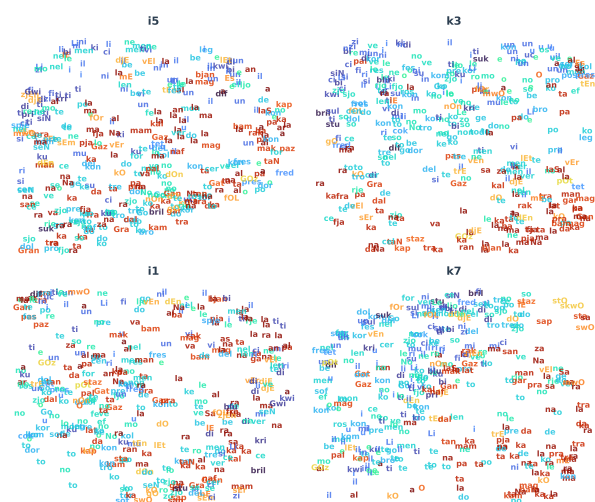


Figure 2: UMAP projections of the multi-subject model’s latent space across four distinct subjects. Data points represent individual syllables, color-coded according to a phonetic sonority gradient, ranging from low-sonority syllables often characterized by occlusive consonants (blue/purple) to high-sonority syllables dominated by open vowels (red/orange). The visualizations give interesting qualitative insights about the model’s learned representation. Syllables with similar acoustic and articulatory properties naturally form cohesive clusters along a continuous phonetic spectrum.

CONCLUSIONS

In this study, we introduced a novel framework for an EEG-based speech BCI leveraging the syllabic structure

of the Italian language. By conceptualizing syllable decoding as a cross-modal representation learning problem, we aligned high-dimensional EEG signals with pre-trained acoustic embeddings via a contrastive objective. We established a robust baseline for overt speech decoding, utilizing a retrieval strategy operating exclusively within the learned EEG latent space during inference. Our evaluations demonstrated that inter-subject variability remains a primary bottleneck for generalized BCI systems, as single-subject models outperformed multi-subject architectures despite relying on significantly less training data. Crucially, our ablation studies highlighted the instrumental role of the *Subject Layer*, which provided substantial accuracy boosts across both paradigms by facilitating subject-specific feature integration. While the *Spatial Attention Layer* did not independently improve decoding accuracy, its inclusion alongside the *Subject Layer* stabilized multi-subject performance and was retained to maintain architectural alignment with current state-of-the-art methodologies.

Despite the promising baseline established by this framework, we acknowledge some critical methodological limitations. First, the current decoding evaluation is restricted to the overt speech condition. While this provides a necessary ground truth for cross-modal alignment, vocalized speech inevitably introduces electromyographic (EMG) artifacts that most likely have inflated the decoding performance. However, for end-users unable to produce intelligible speech, residual articulatory movements may still provide a valuable complementary EMG signal. Capturing this via a non-invasive EEG cap represents a highly accessible and elegant alternative to surgical implants or dense arrays of facial electrodes. Furthermore, it is important to note that even state-of-the-art invasive neuroprostheses rely on attempted speech paradigms, in which the participant actually attempts the articulation, rather than purely imagined speech [3, 4]. This said, it still holds that the true cortical decoding capacity of our non-invasive model must be validated on the imagined speech data in future iterations. A second critical limitation of the current study is the reliance on the manual temporal alignment of EEG data to speech onsets, guided by the synchronized audio recordings. Preliminary tests conducted without this manual alignment resulted in an approximate 12% drop in decoding performance. This discrepancy highlights the necessity of developing a robust, brain-based Voice-Activity Detector (VAD) capable of accurately triggering the decoder in real time without acoustic ground truth.

FUTURE DIRECTIONS

Looking forward, our immediate next steps are to extend this decoding framework to the imagined speech data collected within our protocol, and to implement an automated VAD to ensure accurate, real-time onset detection. We estimate that our syllable-based BCI could achieve a theoretical maximum communication rate of approx-

imately 50 WPM prior to any predictive text enhancements. Although this remains well below the average rate of natural speech, this is more than double the speed of current AAC systems. To achieve full sentence decoding, we will focus on integrating syllable probabilities with N-gram language models and/or Large Language Models. By combining neural decoding with predictive autocompleteness and context-aware language modeling, this framework has the potential to scale into a fast, non-invasive, and highly expressive communication device.

ACKNOWLEDGEMENTS

This publication is part of the project PNRR-NGEU which has received funding from the MUR-DM 118/2023.

REFERENCES

- [1] Elshahar Y, Hu S, Bouazza-Marouf K, Kerr D, Mansor A. Augmentative and alternative communication (aac) advances: A review of configurations for individuals with a speech disability. *Sensors*. 2019;19(8):1911.
- [2] Silva AB, Littlejohn KT, Liu JR, Moses DA, Chang EF. The speech neuroprosthesis. *Nature Reviews Neuroscience*. 2024;25(7):473–492.
- [3] Littlejohn KT et al. A streaming brain-to-voice neuroprosthesis to restore naturalistic communication. *Nature neuroscience*. 2025;28(4):902–912.
- [4] Qian Y et al. Real-time decoding of full-spectrum chinese using brain-computer interface. *Science Advances*. 2025;11(45):eadz9968.
- [5] Wyse-Sookoo K et al. Stability of ecog high gamma signals during speech and implications for a speech bci system in an individual with als: A year-long longitudinal study. *Journal of neural engineering*. 2024;21(4):046016.
- [6] Nieuwland MS et al. Large-scale replication study reveals a limit on probabilistic prediction in language comprehension. *ELife*. 2018;7:e33468.
- [7] Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC. Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*. 2018;28(5):803–809.
- [8] Accou B, Bollens L, Gillis M, Verheijen W, Van Hamme H, Francart T. Sparrkulee: A speech-evoked auditory response repository from ku leuven, containing the eeg of 85 participants. *Data*. 2024;9(8):94.
- [9] d’Ascoli S et al. Towards decoding individual words from non-invasive brain recordings. *Nature Communications*. 2025;16(1):10521.
- [10] Lopez-Bernal D, Balderas D, Ponce P, Molina A. A state-of-the-art review of eeg-based imagined speech decoding. *Frontiers in human neuroscience*. 2022;16:867281.
- [11] Bisla M, Anand R. A comprehensive review on state-of-the-art imagined speech decoding techniques using electroencephalography. *Artificial Intelligence in Biomedical and Modern Healthcare Informatics*. 2025;101–126.
- [12] Zhang Z et al. Chisco: An eeg-based bci dataset for decoding of imagined speech. *Scientific Data*. 2024;11(1):1265.
- [13] Sato M, Tomeoka K, Horiguchi I, Arulkumaran K, Kanai R, Sasai S. Scaling law in neural data: Non-invasive speech decoding with 175 hours of eeg data. *arXiv preprint arXiv:2407.07595*. 2024.
- [14] Craik A, Dial H, L Contreras-Vidal J. Continuous and discrete decoding of overt speech with scalp electroencephalography (eeg). *Journal of Neural Engineering*. 2025;22(2):026017.
- [15] Goslin J, Galluzzi C, Romani C. Phonitalia: A phonological lexicon for italian. *Behavior research methods*. 2014;46(3):872–886.
- [16] Défossez A, Caucheteux C, Rapin J, Kбели O, King JR. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*. 2023;5(10):1097–1107.
- [17] Team S. Silero vad: Pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. <https://github.com/snakers4/silero-vad>. 2021.
- [18] McAuliffe M, Socolof M, Mihuc S, Wagner M, Sonderegger M. Montreal forced aligner: Trainable text-speech alignment using kald. In: *Interspeech*. 2017, 498–502.
- [19] Boersma P. Praat: Doing phonetics by computer (version 5.1.05). <http://www.praat.org/>. 2009.
- [20] Hickok G, Poeppel D. The cortical organization of speech processing. *Nature reviews neuroscience*. 2007;8(5):393–402.
- [21] Mantegna F, Orpella J, Poeppel D. Time-resolved hemispheric lateralization of audiomotor functional connectivity during covert speech production. *Cell Reports*. 2025;44(1).
- [22] Jahangiri A, Sepulveda F. The relative contribution of high-gamma linguistic processing stages of word production, and motor imagery of articulation in class separability of covert speech tasks in eeg data. *Journal of medical systems*. 2019;43(2):20.
- [23] Bhadra K, Giraud AL, Marchesotti S. Learning to operate an imagined speech brain-computer interface involves the spatial and frequency tuning of neural activity. *Communications Biology*. 2025;8(1):271.
- [24] Baeviski A, Zhou H, Mohamed A, Auli M. Wav2vec 2.0: A framework for self-supervised learning of speech representations. *CoRR*. 2020;abs/2006.11477.
- [25] Lawhern VJ, Solon AJ, Waytowich NR, Gordon SM, Hung CP, Lance BJ. Eegnet: A compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of neural engineering*. 2018;15(5):056013.
- [26] Radford A et al. Learning transferable visual models from natural language supervision. *CoRR*. 2021;abs/2103.00020.