Smart Health 32 (2024) 100477

Contents lists available at ScienceDirect

Smart Health

journal homepage: www.elsevier.com/locate/smhl

Multimodal speech recognition using EEG and audio signals: A novel approach for enhancing ASR systems

Anarghya Das [a,*], Puru Soni [a], Ming-Chun Huang [b], Feng Lin [c], Wenyao Xu [a]

[a] Department of Computer Science and Engineering, University at Buffalo, Buffalo, 14261, United States
[b] Department of Data and Computational Science, Duke Kunshan University, Jiangsu, 215316, China
[c] ZJU-Hangzhou Global Scientific and Technological Innovation Center, School of Cyber Science and Technology, Zhejiang University, Zhejiang, 310007, China

ARTICLE INFO

Keywords:
Multimodal speech recognition
Electroencephalography (EEG)
Automatic Speech Recognition (ASR)
Multimodal analysis
Deep learning
Overt speech
Brain–Computer Interface (BCI)
Neural signal processing

ABSTRACT

Speech recognition using EEG signals captured during covert (imagined) speech has garnered substantial interest in Brain–Computer Interface (BCI) research. While the concept holds promise, current implementations must improve performance compared to established Automatic Speech Recognition (ASR) methods using audio. An area often underestimated in previous studies is the potential of EEG utilization during overt speech. Integrating overt EEG signals with speech data by leveraging advancements in deep learning presents significant potential to enhance the efficacy of these systems. This integration proves particularly advantageous in noisy environments and for individuals with speech impairments—challenges even conventional ASR techniques struggle to address effectively. Our investigation delves into this relationship by introducing a novel multimodal model that merges EEG and speech inputs. Our model achieves a multiclass classification accuracy of 95.39%. When subjected to artificial white noise added to the input audio, our model exhibits a notable level of resilience, surpassing the capabilities of models reliant solely on single EEG or audio modalities. The validation process, leveraging the robust techniques of t-SNE and silhouette coefficient, corroborates and solidifies these advancements.

1. Introduction

Human speech involves two primary mechanisms: overt speech and covert speech, also recognized as imagined speech. Overt speech involves physically articulating words using muscular movements to create audible sounds. On the contrary, covert speech involves internally simulating speech without engaging the motor functions, relying solely on mental faculties for linguistic conceptualization and construction (Cooney et al., 2022). While decoding speech has been a longstanding focus within Brain–Computer Interfaces (BCI), the predominant emphasis in research has revolved around utilizing EEG recorded during covert speech for analysis (Lopez-Bernal et al., 2022). Despite exhibiting promise over the years, these methodologies have yet to achieve significantly higher performance than state-of-the-art Automatic Speech Recognition (ASR) techniques.

Automatic Speech Recognition (ASR) systems have undergone substantial advancements, yet they confront many challenges that hinder their optimal performance in real-world scenarios. One significant impairment lies in the struggle to maintain accuracy and reliability in background noise. Noisy environments adversely affect ASR systems by degrading crucial acoustical features for accurate speech recognition. This degradation leads to losing vital time–frequency correlations in the speech signal, resulting

[*] Corresponding author.
E-mail address: anarghya@buffalo.edu (A. Das).

https://doi.org/10.1016/j.smhl.2024.100477
Received 6 March 2024; Accepted 18 March 2024
Available online 3 April 2024

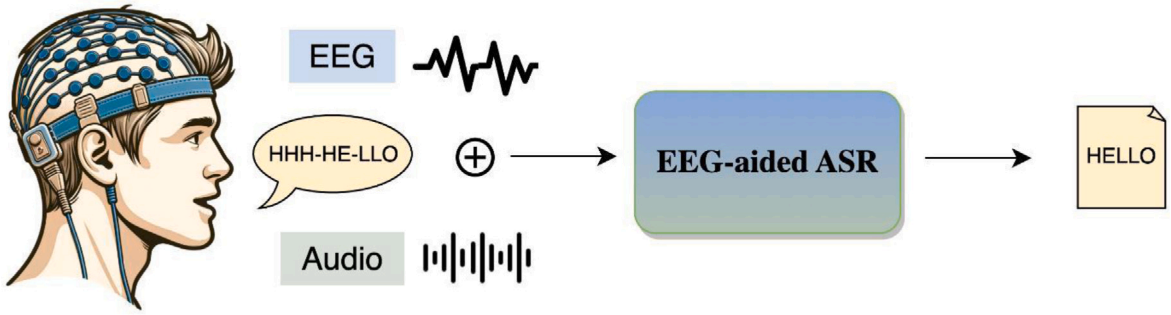

Fig. 1. Illustrates enhancing speech recognition with EEG data, resulting in the accurate text "HELLO" from a stuttered audio input.

in diminished performance and comprehension. ASR systems also grapple with the challenge of diverse dialects and vocabulary sizes (Alharbi et al., 2021). Moreover, individuals grappling with speech impairments such as dysarthria encounter challenges in articulating speech effectively, often exhibiting symptoms like slurring or pauses (Calvo et al., 2021). These impairments pose considerable difficulties for current ASR technologies, hindering their ability to decode and comprehend the speech of these individuals accurately. Pronunciation variations and the unpredictable nature of speech content also pose considerable hurdles for these systems (Young & Mihailidis, 2010).

Leveraging EEG activity during overt speech offers a promising avenue to enhance decoding capabilities. The extensive literature indicates that researchers have observed distinct neural activations in vital speech-related areas of the brain, such as Broca's and Wernicke's regions, during overt speech (Cooney et al., 2018). By fusing this neurological understanding with speech data, the potential emerges to transcend the limitations of singular EEG-based methodologies. This integration amplifies performance and supplements prevailing Automatic Speech Recognition (ASR) techniques, especially in challenging scenarios. The prospect of leveraging this amalgamation extends into a spectrum of practical applications, from improving assistive communication technologies for individuals with speech impairments to refining ASR techniques during challenging scenarios, as demonstrated in Fig. 1.

Our work delves into this uncharted relationship between overt speech and brain activity, aiming to enhance the efficacy of existing ASR systems using a publicly available dataset, especially in challenging conditions that traditional approaches struggle to address. Our exploration offers potential solutions for effective communication among individuals with speech impairments and demonstrates the synergistic potential of combining EEG signals with speech data. Below, we highlight the primary contributions of our work.

- **Enhanced Multimodal Performance**: This study demonstrates improved multiclass classification accuracy through cross-validation by integrating EEG and overt speech inputs, achieving an average accuracy of 95.39%.
- **Robustness**: Adding white noise to overt speech demonstrates that the multimodal approach outperforms individual modalities. This robustness against artificially generated white noise underscores its potential for real-world applicability in scenarios where traditional ASR systems fall short.
- **Improved Learning of Embeddings**: The study showcases how the multimodal model's efficient learning of embeddings leads to improved classification. Comparative analysis using t-distributed Stochastic Neighbor Embedding (t-SNE) and silhouette coefficient illustrate the model's superior ability to leverage learned embeddings compared to single-modality EEG or audio models.

2. Background

Brain–computer interface (BCI)-aided speech recognition involves four stages: Signal Acquisition, Pre-processing, Feature Extraction, and Classification. The integration of EEG recorded during overt speech within this framework presents inherent challenges due to muscular artifacts during signal acquisition, a complexity absent in covert speech. Nevertheless, recent strides in deep learning techniques have diminished the need for extensive EEG data pre-processing (Gong et al., 2022), enabling studies to leverage EEG collected during overt speech in various applications. Previous studies in this domain have focused on applications like emotion recognition (Wang et al., 2022) while only a few have highlighted the benefits of unifying overt EEG and speech (Krishna, Tran, Carnahan et al., 2019; Krishna, Tran, Yu et al., 2019). Therefore, there remains a paucity of literature extensively exploring this intersection.

EEG signal analysis for speech involves diverse methods of capturing essential information for precise signal classification. Time domain statistics reveal signal traits while using the Fast Fourier Transform (FFT), the frequency domain uncovers frequency-specific details crucial for identifying neural speech patterns. These methods can be applied individually or across multiple channels. Advanced techniques like channel cross-covariance matrices and Riemannian geometry combine channel features, improving EEG-based speech analysis (Lopez-Bernal et al., 2022). Our study employed FFT and covariance matrices to extract EEG speech features across domains for accurate signal classification. 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 (Davis & Mermelstein, 1980). Thus, we used MFCCs for audio feature extraction in our research.

Late fusion is a multimodal integration technique that combines information from diverse sources or modalities in the later stages of the data processing pipeline. This method enables independent processing of each modality until integrating their outputs, offering tailored processing techniques to optimize the characteristics of individual modalities. Late fusion ensures computational efficiency and combines diverse data sources into a unified framework, enhancing overall system performance in various multimodal applications (Zhang et al., 2021). In our study, this fusion method played a pivotal role in integrating EEG and audio modalities to improve the efficacy of our system.

3. Methods

In this study, we utilized a publicly available dataset, trained separate EEG and audio encoder models, and subsequently fused them to develop a robust multimodal model. Our analysis focused on evaluating the outcomes derived from this integrated approach.

3.1. Dataset information

The dataset (Coretto et al., 2017) used for this paper is publically available and encompassed 15 healthy young individuals proficient in their native Spanish language. The participant pool comprised seven females and eight males. This dataset captured EEG signals using Ag-AgCl cup electrodes with conductive paste, diverging from the traditional cap setup. Electrode placement centered on a 6-channel configuration in the central region of the head, adhering to the 10–20 system. The EEG recordings were sampled at 1024 Hz, while the accompanying audio was captured at 44.1 kHz. The experimental protocol consisted of four distinct stages: an initial 2-second rest period, a subsequent 2-second phase displaying the prompt on the screen, a 4-second duration dedicated to the mental rehearsal of speech without physical articulation, and finally, a 4-second spoken phase where participants vocalized the prompt. The prompt collection encompassed 11 items: five Spanish vowels and six command words in Spanish. Each subject repeated each prompt at least 50 times throughout the experiment. We exclusively utilized the temporally synchronized EEG and audio recordings captured explicitly during the 4-second spoken phase of the experiment.

3.2. Data pre-processing

The EEG data in the dataset had undergone pre-processing using a bandpass FIR filter, limiting frequencies to 2 Hz to 40 Hz. This process successfully isolated the EEG frequency spectrum while eliminating the disruptive 50 Hz line noise. While muscle movement or electrode placement artifacts were manually identified and removed from the data, blink artifacts were specifically annotated but not subjected to further removal. In this study, we refrained from additional cleaning of the EEG data beyond the initially provided pre-processing and consequently retained blink artifacts within the dataset for analysis. This decision was grounded in the understanding that removing these artifacts would have yielded little improvements in accuracy or overall outcomes of the study (Delorme, 2023). We utilized the MNE package (Gramfort et al., 2013) in Python to process the EEG data by segmenting it into epochs, each lasting 4 s. Each subject performed multiple repetitions of words/vowels, termed trials, resulting in a dataset comprising 1974 trials across all 15 subjects. Simultaneously, the EEG data and corresponding audio recordings were segmented into 100 ms windows with a 50% overlap to facilitate model training and testing. This segmentation produced 79 overlapping windows for the EEG data, each containing 4096 samples per channel over 4 s. Similarly, the audio recordings for each trial were divided into overlapping windows of the same duration and overlap percentage to enable parallel processing and analysis.

3.3. Feature extraction

Our study utilized a power spectrum cross-covariance matrix for EEG feature extraction, leveraging its demonstrated efficacy and potential in prior studies focused on imagined speech applications (Rusnac & Grigore, 2022). We initially computed each EEG channel's power spectrum in decibels using the discrete Fast Fourier Transform (FFT), as depicted in Eq. (1). Only the first half of the frequencies were considered in the power spectrum calculation to account for the symmetry of real-valued signals in the frequency domain:

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

Here, x represents the EEG data from each channel, and y denotes the corresponding power spectrum. We then computed the covariance across all EEG channels to generate the final windowed EEG features using a $channel{\times}channel$ approach, resulting in 6×6 feature matrices in our specific case. We extracted 13-MFCC features from the audio data by leveraging the librosa library (McFee, 2023). 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. By aligning the MFCC sequence with the EEG data, we ensured synchronized time points, enabling meaningful comparisons and integrative insights, thereby enhancing the coherence and interpretability of cross-modal analyses between auditory and EEG data.

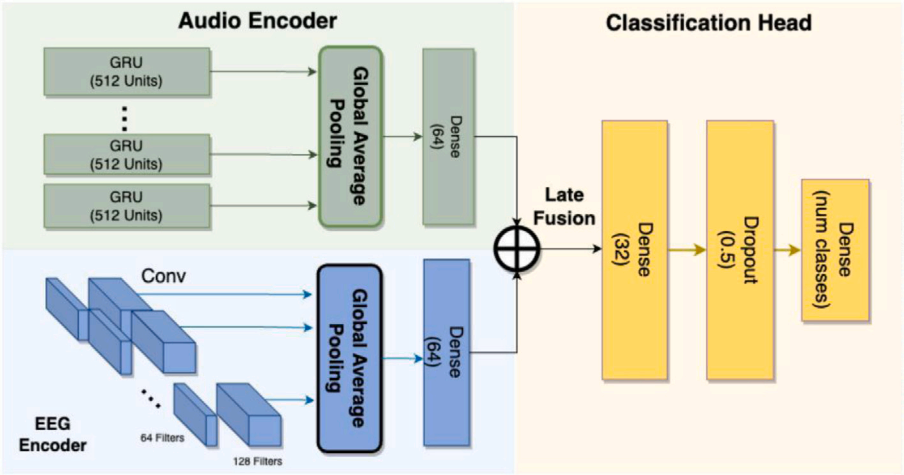

Fig. 2. Architecture of the combined multimodal model integrating Audio and EEG Data.

3.4. Multimodal model

We introduce a novel multimodal neural network architecture for classifying spoken phrases, as illustrated in Fig. 2. The model integrates information from two modalities: Mel-frequency cepstral coefficients (MFCCs) extracted from audio signals and cross-covariance matrix features of EEG data. It comprises distinct encoders for each modality. For the EEG data, the model incorporates a Time-Distributed CNN consisting of two convolution layers with 64 and 128 filters, followed by ReLU activation and flattening operations. This architecture tailors explicitly to capture the inherent spatial information in EEG signals. Inspired by insights from Krishna, Tran, Yu et al. (2019), our audio encoder includes a GRU layer with 512 units adept at capturing temporal dependencies present in audio signals. Following the GRU layer, global average pooling condenses the temporal information into a fixed-size representation, further processed through a dense layer with ReLU activation. We intricately design these encoders to extract and encode relevant features. The model utilizes concatenation layers to fuse information extracted from both modalities before channeling it through dense layers for classification. These concatenated features undergo dense layers with ReLU activation and dropout regularization, preventing overfitting.

Consequently, an output layer utilizing softmax activation classifies the spoken phrases. This approach notably integrates spatial and temporal characteristics from EEG and MFCC data, enhancing spoken phrase classification. The inclusion of dropout layers ensures the model's robustness and generalization.

3.5. Experiments

Our experiments aimed to compare audio data's performance alone and combine audio and EEG signals for classification tasks. Additionally, we explored the impact of introducing white noise to the audio recordings and its influence on the classification results of both the audio-only and multimodal models. Gaussian white noise, drawn from a normal distribution, was added to the audio data. To simulate a specific Signal-to-Noise Ratio (SNR), the noise power was computed using Eq. (2):

$$\text{Noise Power} = \text{audio}^2 \times 10^{-\frac{\text{SNR}}{10}}. \tag{2}$$

Afterward, we scaled the noise power with randomly sampled noise from the standard normal distribution and added the resulting noise to the audio. This process generated noisy audio with the desired signal-to-noise ratio (SNR).

The experimental setup involved a 5-fold stratified cross-validation procedure, ensuring an equal distribution of class labels in each fold. This approach aimed to maintain label balance across the dataset partitions during training and evaluation. The model was trained for 100 epochs using Categorical Cross Entropy as the loss function. We employed the Adam optimizer with a learning rate set to 0.003 to optimize the model parameters during training.

4. Results

4.1. Cross-validation

The cross-validation analysis comparing the multimodal model, integrating both EEG and audio data, with the audio-only model, specifically utilizing noise-free audio data, revealed notable insights. Across the 5-fold cross-validation, the multimodal model

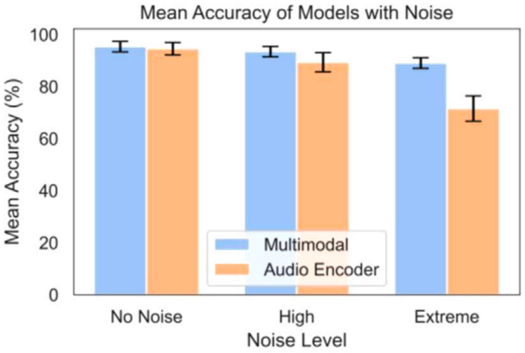

Fig. 3. Illustrates the average accuracy of the Multimodal and Audio encoder models derived from averaging the 5-fold cross-validation accuracies, considering diverse noise levels incorporated into the audio dataset. Error bars denoting standard deviation accompany the presented results.

achieved an average classification accuracy of 95.39%, slightly surpassing the 94.58% accuracy achieved by the audio-only model. Although the margin of improvement may not be substantial, these results underscore the potential synergy achieved by integrating information from brain activity alongside audio data.

While the observed improvement in performance was modest, a more promising scenario unfolded while assessing the models' efficacy when subjected to various levels of white noise. Our study introduced high and extreme noise levels into the audio recordings, with a signal-to-noise ratio (SNR) of 1 for high noise and 0.5 for extreme noise. It is important to note that SNR values below 12 dB are typically considered unfavorable for audio analysis. In this rigorous assessment, post-cross-validation analysis revealed a substantial divergence in performance between the two models. Notably, the classification accuracy of the audio-encoder model exhibited a downward trend, whereas the performance difference compared to the multimodal model further widened, as depicted in Fig. 3. Under extreme noise conditions, the multimodal model demonstrated an average accuracy of 89.06%, significantly surpassing the audio-only model, which experienced a notable decline to 71.58%. These findings underscore the heightened robustness of EEG-augmented classification in adverse noise environments compared to relying solely on audio signals.

4.2. *Exploration of learning representations*

We sought to visualize learning representations before classification to delve deeper into the learning processes of the Audio encoder, EEG encoder, and multimodal models shown in Fig. 4. This exploration aimed to elucidate the distinctive learning patterns inherent in each model when processing audio and EEG data. To accomplish this, t-Distributed Stochastic Neighbor Embedding (t-SNE) (Maaten & Hinton, 2008) was employed to visualize each model's final layers preceding the classification head. The t-SNE plot was constructed by extracting the outputs from the dense layer preceding the classification stage and transforming these representations into a two-dimensional space. The labels corresponding to each class were also prominently highlighted within the plot. This visualization technique facilitated a comparative analysis of the learned representations, shedding light on the nuanced learning dynamics inherent in the audio-only, EEG-only, and multimodal models when confronted with audio and EEG data.

To evaluate the quality of the t-SNE visualizations for classification prediction, we employed the Silhouette coefficient (Rousseeuw, 1987) metric. Each sample's mean intra-cluster and nearest-cluster distance define the silhouette coefficient. It provides insights into clustering efficacy by indicating the appropriate grouping of data points within their respective clusters. Ranging from −1 to 1, a value closer to 1 represents well-clustered data, 0 indicates overlapping clusters and negative values suggest potential incorrect sample assignments due to similarities with other clusters. Upon evaluation, the Audio-only encoder model demonstrated a silhouette coefficient of 0.44, indicating moderate intra-cluster coherence.

In contrast, the multimodal model exhibited a higher silhouette coefficient of 0.69, signifying significantly enhanced performance by leveraging both EEG and audio inputs. This improvement demonstrates the model's heightened capacity to delineate distinct classes more effectively. Conversely, the silhouette coefficient for the EEG-only model recorded at −0.12 suggests a need for more well-defined clusters. This negative score indicates that relying solely on EEG features with the current model architecture lacks sufficient discriminatory power for autonomous classification tasks.

5. Conclusion

This study introduces a pioneering multimodal model designed to classify spoken phrases by leveraging EEG and audio signals. By fusing both modalities, the model achieves an impressive average accuracy of 95.39% across 11 classes of Spanish vowels and words. Notably, it exhibits robustness when white noise is added to overt speech, surpassing the performance of the audio encoder model by a significant margin. Moreover, the multimodal model displays enhanced learning of embeddings, as evidenced by t-SNE visualization and silhouette coefficient analysis. These findings underscore the robust potential of integrating brain activity and speech signals, which is particularly beneficial for improving communication and comprehension in diverse conditions, especially

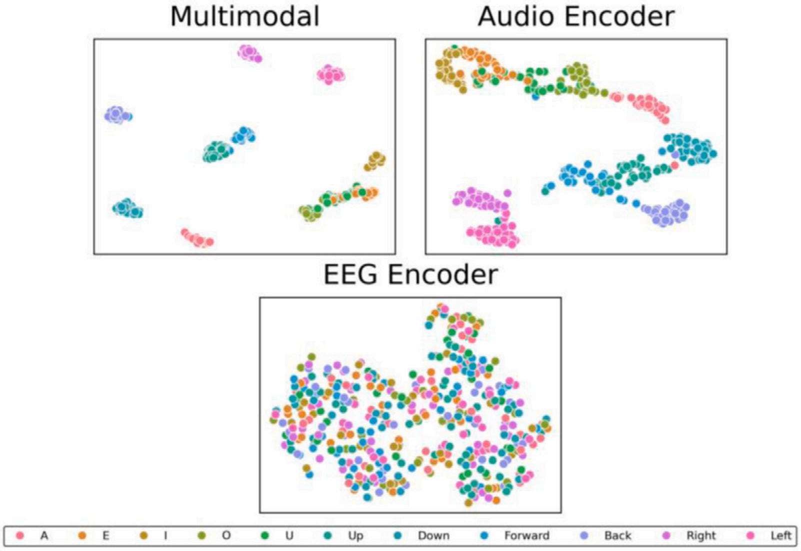

Fig. 4. t-SNE visualization comparing the clustering of phoneme representations by Multimodal, Audio encoder, and EEG encoder models.

among individuals with speech impairments. However, it is crucial to recognize the constraints imposed by the limited availability of data that captures both EEG and speech. To improve the evaluation methodology, expanding the dataset to encompass participants from diverse linguistic backgrounds and incorporating more complex speech tasks could significantly enhance the generalizability and robustness of the study's findings. Despite this limitation, our research addresses critical gaps by demonstrating the unique advantages of the fusion of overt speech and EEG data, emphasizing its crucial role in advancing ASR systems across various contexts and populations. Future research endeavors should explore alternative multimodal architectures, integrating a range of EEG encoders to enrich signal interpretation further. An intriguing direction for future research entails integrating overt EEG signals with established ASR models like Whisper (Radford et al., 2022), followed by a thorough comparative analysis. This approach promises to unveil exciting possibilities for advancement within this domain. Collecting data from individuals with speech impairments could further validate the model's real-world performance, offering insights into its practicality and reliability. Exploring these avenues holds promise for refining multimodal architectures and expanding their applicability across a broader spectrum of contexts and users.

CRediT authorship contribution statement

Anarghya Das: Formal analysis, Methodology, Software, Visualization, Writing – original draft, Writing – review & editing, Data curation, Investigation, Project administration, Validation. **Puru Soni:** Software, Validation, Writing – original draft, Writing – review & editing. **Ming-Chun Huang:** Funding acquisition. **Feng Lin:** Writing – review & editing. **Wenyao Xu:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Ming-Chun Huang reports financial support was provided by Kunshan Municipal Government. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgment

The research outcomes presented in this article were in part supported by funding from the Kunshan Municipal Government.

References

Alharbi, S., Alrazgan, M., Alrashed, A., Alnomasi, T., Almojel, R., Alharbi, R., Alharbi, S., Alturki, S., Alshehri, F., & Almojil, M. (2021). Automatic Speech Recognition: Systematic Literature Review. *IEEE Access*, *9*, 131858–131876. http://dx.doi.org/10.1109/ACCESS.2021.3112535, URL: https://ieeexplore.ieee.org/document/9536732, Conference Name: IEEE Access.

Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A. B., Grajzer, M., Gilardone, M., & Corbo, M. (2021). Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech. *Folia Phoniatrica Et Logopaedica: Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, *73*(5), 432–441. http://dx.doi.org/10.1159/000511042.

Cooney, C., Folli, R., & Coyle, D. (2018). Neurolinguistics Research Advancing Development of a Direct-Speech Brain-Computer Interface. *iScience*, *8*, 103–125. http://dx.doi.org/10.1016/j.isci.2018.09.016, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6174918/.

Cooney, C., Folli, R., & Coyle, D. (2022). Opportunities, pitfalls and trade-offs in designing protocols for measuring the neural correlates of speech. *Neuroscience & Biobehavioral Reviews*, *140*, Article 104783. http://dx.doi.org/10.1016/j.neubiorev.2022.104783, URL: https://www.sciencedirect.com/science/article/pii/S014976342200272X.

Coretto, G. A. P., Gareis, I. E., & Rufiner, H. L. (2017). Open access database of EEG signals recorded during imagined speech. *Vol. 10160*, In *12th international symposium on medical information processing and analysis*. SPIE, Article 1016002. http://dx.doi.org/10.1117/12.2255697.

Davis, S., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech and Signal Processing*, *28*(4), 357–366. http://dx.doi.org/10.1109/TASSP.1980.1163420, URL: https://ieeexplore.ieee.org/document/1163420, Conference Name: IEEE Transactions on Acoustics, Speech, and Signal Processing.

Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, *13*(1), 2372. http://dx.doi.org/10.1038/s41598-023-27528-0, URL: https://www.nature.com/articles/s41598-023-27528-0, Number: 1 Publisher: Nature Publishing Group.

Gong, S., Xing, K., Cichocki, A., & Li, J. (2022). Deep Learning in EEG: Advance of the last ten-year critical period. *IEEE Transactions on Cognitive and Developmental Systems*, *14*(2), 348–365. http://dx.doi.org/10.1109/TCDS.2021.3079712, URL: https://ieeexplore.ieee.org/document/9430619/.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*, URL: https://www.frontiersin.org/articles/10.3389/fnins.2013.00267.

Krishna, G., Tran, C., Carnahan, M., & Tewfik, A. (2019). Advancing Speech Recognition With No Speech Or With Noisy Speech. In *2019 27th European signal processing conference (EUSiPCo)* (pp. 1–5). http://dx.doi.org/10.23919/EUSIPCO.2019.8902943, URL: https://ieeexplore.ieee.org/document/8902943, ISSN: 2076-1465.

Krishna, G., Tran, C., Yu, J., & Tewfik, A. H. (2019). Speech Recognition with No Speech or with Noisy Speech. In *ICASSP 2019 - 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1090–1094). http://dx.doi.org/10.1109/ICASSP.2019.8683453, URL: https://ieeexplore.ieee.org/abstract/document/8683453, ISSN: 2379-190X.

Lopez-Bernal, D., Balderas, D., Ponce, P., & Molina, A. (2022). A State-of-the-Art Review of EEG-Based Imagined Speech Decoding. *Frontiers in Human Neuroscience*, *16*, Article 867281. http://dx.doi.org/10.3389/fnhum.2022.867281, URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9086783/.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, *9*(86), 2579–2605, URL: http://jmlr.org/papers/v9/vandermaaten08a.html.

McFee, B. (2023). Librosa/librosa: 0.10.1. http://dx.doi.org/10.5281/zenodo.8252662.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision. http://dx.doi.org/10.48550/arXiv.2212.04356, URL: http://arxiv.org/abs/2212.04356 arXiv:2212.04356 [cs, eess].

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*, 53–65. http://dx.doi.org/10.1016/0377-0427(87)90125-7.

Rusnac, A.-L., & Grigore, O. (2022). CNN Architectures and Feature Extraction Methods for EEG Imaginary Speech Recognition. *Sensors*, *22*(13), 4679. http://dx.doi.org/10.3390/s22134679, URL: https://www.mdpi.com/1424-8220/22/13/4679.

Wang, Q., Wang, M., Yang, Y., & Zhang, X. (2022). Multi-modal emotion recognition using EEG and speech signals. *Computers in Biology and Medicine*, *149*, Article 105907. http://dx.doi.org/10.1016/j.compbiomed.2022.105907, URL: https://www.sciencedirect.com/science/article/pii/S0010482522006503.

Young, V., & Mihailidis, A. (2010). Difficulties in Automatic Speech Recognition of Dysarthric Speakers and Implications for Speech-Based Applications Used by the Elderly: A Literature Review. *Assistive Technology*, *22*(2), 99–112. http://dx.doi.org/10.1080/10400435.2010.483646, Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/10400435.2010.483646.

Zhang, Y., Sidibé, D., Morel, O., & Mériaudeau, F. (2021). Deep multimodal fusion for semantic image segmentation: A survey. *Image and Vision Computing*, *105*, Article 104042. http://dx.doi.org/10.1016/j.imavis.2020.104042, URL: https://www.sciencedirect.com/science/article/pii/S0262885620301748.