# Enhancing the EEG Speech Match Mismatch Tasks With Word Boundaries

*Akshara Soman, Vidhi Sinha, Sriram Ganapathy*

LEAP Lab, Indian Institute of Science Bangalore, India

aksharas@iisc.ac.in, vidhis@iisc.ac.in, sriramg@iisc.ac.in

## Abstract

Recent studies have shown that the underlying neural mechanisms of human speech comprehension can be analyzed using a match-mismatch classification of the speech stimulus and the neural response. However, such studies have been conducted for fixed-duration segments without accounting for the discrete processing of speech in the brain. In this work, we establish that word boundary information plays a significant role in sentence processing by relating EEG to its speech input. We process the speech and the EEG signals using a network of convolution layers. Then, a word boundary-based average pooling is performed on the representations, and the inter-word context is incorporated using a recurrent layer. The experiments show that the modelling accuracy can be significantly improved (match-mismatch classification accuracy) to 93% on a publicly available speech-EEG data set, while previous efforts achieved an accuracy of 65-75% for this task.

**Index Terms**: Speech-EEG match mis-match task, auditory neuroscience, word segmentation, speech comprehension.

## 1. Introduction

Humans have the unique ability to communicate through speech. While speech comprehension is mastered from a young age, many neural processes enabling this seamless activity are unknown. One of the simplest ways of furthering the understanding of speech comprehension is through the recording of neural responses using electroencephalography (EEG). The EEG is a non-invasive neural imaging technique that measures electrical activity in the brain by placing electrodes on the scalp [1]. It has been demonstrated that the EEG signal recorded during a speech listening task contains information about the stimulus [2]. One can investigate how the brain comprehends continuous speech by developing models that relate the speech with the EEG signal using machine learning techniques [3].

The early attempts explored linear models for relating continuous natural speech to EEG responses [4, 5, 6, 7]. They can be categorized into three different types - forward models, backward models, or hybrid models. The forward models predict EEG from speech stimuli, while the backward models reconstruct speech from EEG responses. In many studies, the correlation between the predicted and ground truth signal is considered as a measure of neural tracking [8]. However, linear models may be ill-equipped to capture the non-linear nature of the auditory system. Recently, deep neural networks have been employed to compare and analyze speech stimuli and EEG responses. Several studies have shown promising results with deep learning models for EEG-speech decoding [9, 10, 11, 12]. These advancements in speech decoding from the brain will also be beneficial for the development of brain-computer interfaces(BCIs).

In many of the computational approaches, the speech envelope has been the most commonly used feature [4, 5, 7]. Other features such as spectrograms [13, 14], phonemes [13], linguistic features [14, 15], and phono-tactics [16] have also been explored with linear forward/backward models. Lesenfants et al. [17] demonstrated that combining phonetic and spectrogram features improves the EEG-based speech reception threshold (SRT) prediction.

While forward/backward models and correlation tasks were previously explored, the match mismatch tasks have been recently investigated as an alternative task [18, 11]. Here, the task is to identify whether a portion of the brain response (EEG) is related to the speech stimulus that evoked it. In the previous studies using the match mismatch task, the auditory stimulus and speech of a fixed duration (5s) are processed through a series of convolutional and recurrent layers [9, 12, 19].

In this work, we argue that the prior works on speech-EEG match mismatch tasks are incomplete without considering the fragmented nature of speech comprehension. While speech and EEG signals are continuous, the neural tracking of speech signals is impacted by the linguistic markers of speech [20]. The most striking of this evidence comes from models of word surprisal [21] with N400 response evoked for unpredictable words [22, 23]. In the simplest form, we hypothesize that the task of relating continuous speech with EEG must also include word-level segmentation information.

We propose a deep learning model to perform match mismatch classification tasks on variable length inputs using word boundary information. The model consists of convolutive feature encoders of both the speech and EEG inputs. Further, the word segmentation information, obtained by force-aligning the speech with the text data using a speech recognition system, is incorporated in the feature outputs through a word-level pooling operation. The pooled representations are further modelled with recurrent long short-term memory (LSTM) layers to model the inter-word context. The final output from the LSTM network for the speech and EEG streams is used in the match mismatch classification task.

The major contributions of this paper are:

- Proposing a match mismatch classification model that can incorporate word boundary information.
- Proposing a loss function based on Manhattan distance for the match mismatch task.
- Experimental illustration of the effectiveness of the model, where the classification performance is significantly improved over the prior works.
- A detailed set of ablation experiments to elicit the impact of word boundary information in speech EEG matching task.
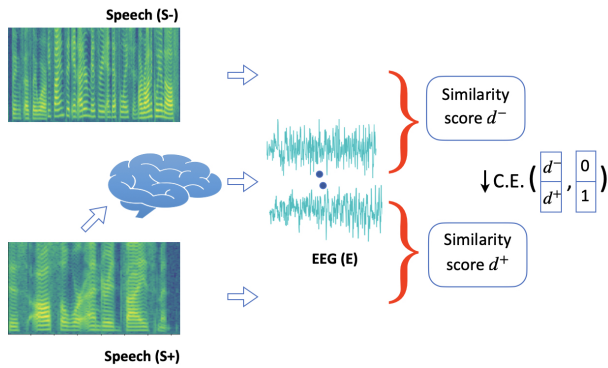
Figure 1: **Match-mismatch classification task:** *It is a binary classification paradigm associating the EEG and speech segments. The EEG segment (**E**) and corresponding stimulus sentence (**S**+) form the positive pair while the same EEG and another (unrelated) sentence (**S**−) form the negative pair. The similarity score computation is achieved using the model depicted in Figure 2. Here, C.E. denotes the cross entropy loss.*

# 2. Methods

## 2.1. Dataset

We experiment with a publicly available speech-EEG data set[1] released by Broderick et al. [24]. It contains electroencephalographic (EEG) data recorded from 19 subjects as they listened to the narrative speech. The subjects listened to a professional audio-book narration of a well-known work of fiction read by a single male speaker. The data consists of 20 trials of roughly the same length, with each trial containing 180s of audio. The trials preserved the chronology of the storyline without repetitions or breaks. The sentence start and end time, and the word-level segmentation of the speech recordings are provided. The word segmentation is obtained using a speech recognition-based aligner [25]. The EEG data were acquired using the 128-channel BioSemi system at a sampling rate of 512Hz, while the audio data was played at 16kHz. Overall, the speech-EEG data amounted to a duration of 19 hours.

## 2.2. EEG Preprocessing

The CNSP Workshop 2021 guidelines[2] served as the basis for the EEG pre-processing pipeline. It is implemented using the EEGLAB software [26].The EEG signal is band-pass filtered between 0.5-32 Hz. Then it is down-sampled to 64Hz. After removing noisy channels (determined using the channel level statistics), the EEG channels are re-referenced to the mastoids. The data from each channel is also normalized by computing the z-score. The EEG pre-processing code and the codes used for further analysis discussed in this paper publicly available[3].

## 2.3. Acoustic Feature - Mel Spectrogram

The mel spectrogram of the speech signal is used as the stimulus feature. The mel spectrogram is computed for each sentence. A mel filter bank with 28 filters distributed in the mel-scale ranging from 0-8kHz frequency is used. The input audio is pre-emphasized with a factor of 0.97 before windowing. In order

---

[1]https://doi.org/10.5061/dryad.070jc
[2]https://cnspworkshop.net/resources.html
[3]https://github.com/iiscleap/EEGspeech-MatchMismatch

---

to obtain speech features at a sampling frequency of 64Hz, the spectrogram computation uses a Hamming window function of the width 31.25ms with half overlap.

## 2.4. Match-mismatch classification task

The accuracy of a match-mismatch classification task is employed in this study as a measure of the neural tracking of speech. Figure 1 illustrates this paradigm in detail. The classification model is trained to relate the speech segment to its corresponding EEG response. In this study, the segment is chosen to be a sentence. We also compare with prior works [9, 27], which perform this task at the sentence level. The time-synchronized stimulus of the EEG response segment is the matched speech. Another sentence from the same trial of data collection is chosen as the mismatched speech. Selecting mismatched samples from the same trial makes the classification task challenging enough to encourage the model to learn the stimulus-response relationships. This sampling approach also avoids the chances of memorizing the speech features along with its label. The matched EEG response for these speech sentences is also included in the mini-batch training to ensure that memorisation is disallowed.

## 2.5. Model architecture

We employed different modelling paradigms to analyze the encoding of acoustic and semantic features in EEG signals.

### 2.5.1. Baseline Model

Recently, Monesi et al. [9] showed that convolutional neural network (CNN) and long short-term memory (LSTM) based architectures outperform linear models for modelling the relationship between EEG and speech. This work employed a match mismatch classification task on fixed duration windows of speech and their corresponding EEG data. The work also demonstrated that mel spectrogram features of the speech stimulus provide the best neural tracking performance compared to other representations like speech envelope, word embedding, voice activity and phoneme identity [27]. They have performed the match mismatch task of 5s duration segments with 90% overlap between successive frames. The prior works [9, 27] use an angular distance between EEG and speech representations, average pooling over time, and a sigmoid operation. The model is trained with binary cross entropy loss [27]. We use this approach as the baseline setup for the proposed framework.

### 2.5.2. Proposed match mismatch Model

The speech signal representation $\mathbf{S}$ is the mel-spectrogram of dimension $28 \times T$, where $T$ denotes the duration of a speech sentence at 64Hz. Similarly, the EEG data for the same sentence is denoted as $\mathbf{E}$, and it is of dimension $128 \times T$.

Both the speech and the EEG features are processed through a parallel neural pipeline, as depicted in Figure 2, without any weight sharing. This sub-network consists of a series of convolutional layers and LSTM layers. The convolutional layers implement 1-D and 2-D convolutions with $1 \times 8$ and $16 \times 9$ kernel sizes, respectively. The 1-D and 2-D layers have 8 and 16 kernels, respectively. Further, the 2-D CNN layers also introduce a stride of $(1, 3)$ to further down-sample the feature maps.

The word boundary information available in the dataset is converted to the equivalent sampling rate (both EEG and audio representations at $\frac{64}{3}$ Hz). The audio and EEG feature maps
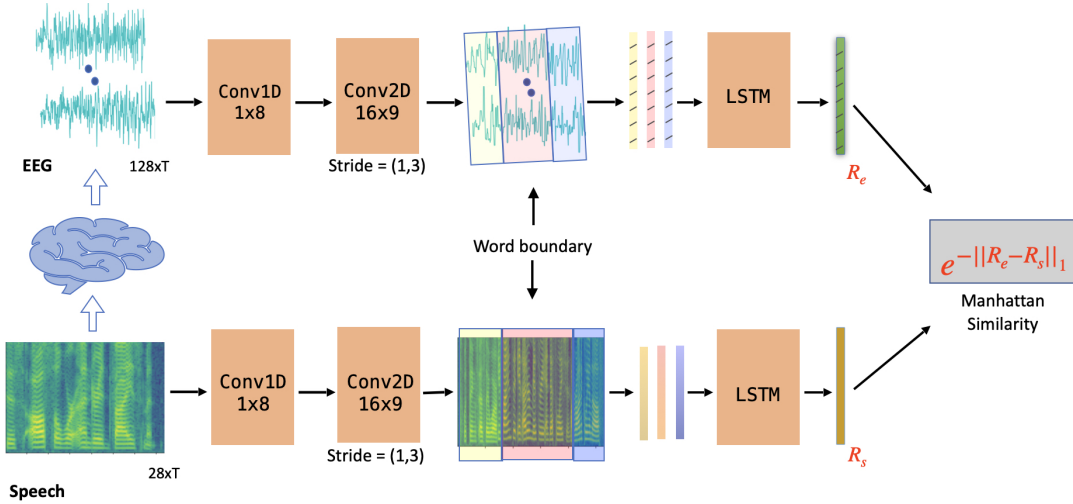
Figure 2: *Proposed model for match mismatch task on speech EEG data. The model training paradigm is outlined in Figure 1.*

Table 1: *Match mismatch classification accuracy of baseline model [9] for fixed duration sequences. The step size between adjacent frames is* 0.5s.

| Frame Width (sec.) | Test Accuracy (%) |
|---|---|
| 1 | 62.21 |
| 3 | 72.41 |
| 5 | 76.12 |

Table 2: *The match-mismatch classification accuracy of speech stimulus and its EEG responses in sentence level for baseline [9] and the proposed model. Here, a random speech sentence was chosen as the mismatch sample for each EEG sentence.*

| Test Set | Baseline Model | Proposed Model | | |
|---|---|---|---|---|
| | | Cos. | Euclidean | Manhattan |
| Fold 1 | 65.39 | 88.22 | 93.49 | 94.02 |
| Fold 2 | 65.32 | 88.73 | 93.68 | 94.00 |
| Fold 3 | 64.98 | 86.54 | 93.72 | 93.91 |
| **Average** | 65.23 | 87.83 | 93.63 | **93.97** |

are average pooled at the word level using the word boundary information. As a result, for a given sentence, the EEG and speech branches generate vector representations sampled at the word level. An LSTM layer models the inter-word context from these representations. This layer is included in both the stimulus (speech) and response (EEG) pathways. The last hidden state of the LSTM layer, of dimension 32, is used as the embedding for the stimulus/response, denoted as $R_s/R_e$ respectively.

We propose the Manhattan distance between the stimulus and response embeddings. The similarity score is computed as,

$$d(\mathbf{E}, \mathbf{S}) = \exp(-||R_e - R_s||_1) \quad (1)$$

The similarity score for the matched pair $(\mathbf{E}, \mathbf{S}^+)$ and mismatched pair $(\mathbf{E}, \mathbf{S}^-)$ are computed. The model, with a dropout factor of 0.2, is trained using a binary cross-entropy loss, with $[d(\mathbf{E}, \mathbf{S}^+), d(\mathbf{E}, \mathbf{S}^-)]$ mapped to $[1, 0]$ targets.

*2.5.3. Training and Evaluation Setup*

The dataset contained recordings from 19 subjects. All the experiments reported in this work perform subject-independent evaluation (the subjects used in training are not part of the evaluation). Further, we report the average results of 3-fold validation, with classification accuracy as the metric. The experiments are run with a batch size of 32. The models are trained using Adam optimizer with a learning rate of 0.001 and weight decay parameter of 0.0001. The models are learned with a binary cross-entropy loss.

## 3. Results and Discussion

### 3.1. Baseline model on fixed duration segments.

The baseline implementation for comparison is the work reported in Monesi et al. [9]. This architecture is an LSTM model that operates on fixed-duration audio EEG data. All experiments are run for 20 epochs of training. The result of the model with fixed duration frames is given in Table **??**. In order to increase the amount of training data, we also use 90% overlap between segments.

### 3.2. Baseline model at sentence level

The baseline model architecture is implemented for fixed-duration segments in training and testing. In order to operate at the sentence level, we have modified the dot product operation as element-wise multiplication followed by an average pooling. This score is passed through the sigmoid function, and the model is learned on sentence-level audio-EEG pairs. For the mismatch condition, a random speech spectrogram is paired with the EEG to generate the score. These results are reported in Table 2.

### 3.3. Proposed model with sentence level processing

The results with the proposed model are also reported in Table 2. We compare three different similarity scoring approaches, i) Angular (Cosine) similarity, ii) Negative L2 distance (Euclidean) and iii) proposed Manhattan similarity (Eq. 1). As seen in the results, the Euclidean and Manhattan similarity improves over the cosine similarity. The proposed EEG-speech match-mismatch classifier model reports an average accuracy of 93.97%, which is statistically significantly higher than the baseline model's sentence-level performance (Wilcoxon signed-
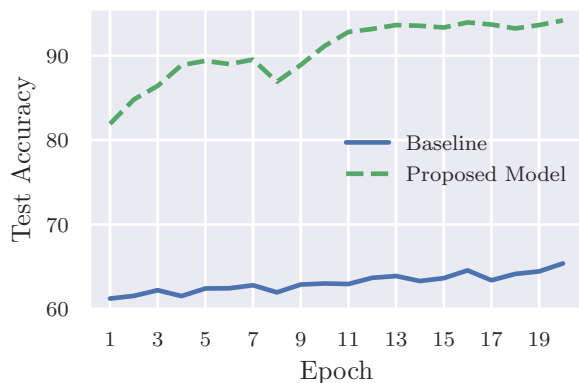
Figure 3: *This figure shows the match-mismatch classification accuracy of the proposed model for test fold-1 as a function of the training epoch for the baseline model and the proposed approach.*

Table 3: *Impact of mismatch sample selection strategy on classification accuracy.*

| Mismatch Selection Strategy | Test Accuracy (%) |
|---|---|
| Random Sentence | 93.97 |
| Next sentence | 91.56 |

rank test, $p < 1e - 4$). The epoch-wise accuracy for test fold-1 is also illustrated in Figure 3.

### 3.4. Mismatch sample selection for sentence processing

Previous match-mismatch EEG-speech studies [11, 9] dealt with fixed-duration speech and EEG segments. Cheveigne et al. [11] used an unrelated random segment as a mismatched sample, while studies like [9, 12, 27] employ a neighbouring segment as the mismatched sample. The sampling of the mismatched segments from the same trial ensures that the distribution of the matched and mismatched segments is similar. We explore a similar strategy for sentence-level analysis by selecting the neighbouring sentence in the same trial as the mismatched sample. Table 3 shows how the mismatch selection strategy affects the classification accuracy. The average accuracy has a slight degradation when the next sentence is used as the mismatch sample.

### 3.5. Importance of accurate word boundaries

We conducted several ablation tests to understand the impact of the word boundary information. The model is fed with random word boundaries in the first set of experiments. Each sentence is assumed to contain a fixed number of words and their boundaries are chosen at random. The results are reported in Figure 4. The accuracy improves gradually when the number of word boundaries is increased, even though they are random. The accuracy of the experiment using 8 words in a sentence is 64%, which is significantly lower than the model's performance with accurate boundary information (Wilcoxon signed-rank test, $p < 0.0001$). The final experiment shown in Figure 4 assumes a random number of words in each sentence with random boundaries, and it provided an accuracy of 60%.

In the second set of experiments, we provide accurate word boundary information but skip the word boundary information
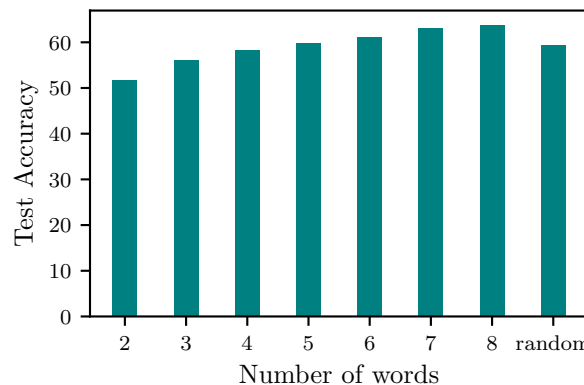


Figure 4: *Average match-mismatch classification accuracy of the proposed model for random word boundaries.*

Table 4: *Accuracy (%) in match mismatch task for varying levels of word-boundary information. Here, $Skip - n$ denotes removing every nth word boundary information in the model.*

| Test set | Skip-2 | Skip-3 | Skip-4 | Skip-5 |
|---|---|---|---|---|
| Fold 1 | 82.45 | 88.96 | 90.43 | 90.64 |
| Fold 2 | 81.86 | 88.77 | 90.32 | 90.28 |
| Fold 3 | 82.60 | 88.79 | 90.30 | 90.01 |
| **Average** | 82.30 | 88.84 | 90.35 | 90.31 |

at every $n$-th word. These results are reported in Table 4. For example, Skip-3 in this table corresponds to removing the word boundary inputs at every 3-rd entry. The pooling is done with the rest of the available word boundaries for these experiments. As seen in Table 4, the results with a higher value of $n$ (of skip-$n$ experiments), approach the setting without any removal (accuracy of 93.97%). It is also noteworthy that, even with the Skip-2 setting (word boundary information available for every alternate word), the performance is 82.3%, significantly better than the baseline model. This study also demonstrates that accurate word boundary information significantly impacts the match mismatch classification, which further illustrates that the EEG signal encodes the word level tracking of speech.

## 4. Conclusions

In this paper, we have attempted to validate the hypothesis that speech comprehension in the brain is segmented at the word-level in the EEG responses to continuous speech. For this task, we developed a deep neural network model consisting of convolutional encoders, word-level aggregators and recurrent layers. A novel loss function for this task based on Manhattan similarity is also proposed. The proposed model validated the hypothesis by improving the accuracy of match-mismatch classification of speech and EEG responses at the sentence level. The incorporation of word boundary information yields statistically significant improvements compared to the baseline model, demonstrating the importance of this information in the neural tracking of speech. Moreover, the proposed model handles variable length inputs. Overall, this model can have potential applications in various domains, including speech recognition, brain-computer interfaces, and cognitive neuroscience. Future research could explore this model's extension to incorporate multi-modal inputs in the form of textual data in addition to the speech spectrogram.

# 5. References

[1] S. Sanei and J. A. Chambers, *EEG Signal Processing*. John Wiley & Sons, 2013.

[2] M. Gillis, J. Vanthornhout, J. Z. Simon, T. Francart, and C. Brodbeck, "Neural markers of speech comprehension: measuring EEG tracking of linguistic speech representations, controlling the speech acoustics," *Journal of Neuroscience*, vol. 41, no. 50, pp. 10 316–10 329, 2021.

[3] S. Sanei and J. A. Chambers, *EEG Signal Processing and Machine Learning*. John Wiley & Sons, 2021.

[4] N. Ding and J. Z. Simon, "Neural coding of continuous speech in auditory cortex during monaural and dichotic listening," *Journal of Neurophysiology*, vol. 107, no. 1, pp. 78–89, 2012.

[5] M. J. Crosse, G. M. Di Liberto, A. Bednar, and E. C. Lalor, "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli," *Frontiers in Human Neuroscience*, vol. 10, p. 604, 2016.

[6] D. D. Wong, S. A. Fuglsang, J. Hjortkjær, E. Ceolini, M. Slaney, and A. De Cheveigne, "A comparison of regularization methods in forward and backward models for auditory attention decoding," *Frontiers in Neuroscience*, vol. 12, p. 531, 2018.

[7] J. Vanthornhout, L. Decruy, J. Wouters, J. Z. Simon, and T. Francart, "Speech intelligibility predicted from neural entrainment of the speech envelope," *Journal of the Association for Research in Otolaryngology*, vol. 19, pp. 181–191, 2018.

[8] A. de Cheveigné, D. D. Wong, G. M. Di Liberto, J. Hjortkjær, M. Slaney, and E. Lalor, "Decoding the auditory brain with canonical component analysis," *NeuroImage*, vol. 172, pp. 206–216, 2018.

[9] M. J. Monesi, B. Accou, J. Montoya-Martinez, T. Francart, and H. Van Hamme, "An LSTM based architecture to relate speech stimulus to EEG," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 941–945.

[10] J. R. Katthi and S. Ganapathy, "Deep correlation analysis for audio-EEG decoding," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 29, pp. 2742–2753, 2021.

[11] A. De Cheveigné, M. Slaney, S. A. Fuglsang, and J. Hjortkjaer, "Auditory stimulus-response modeling with a match-mismatch task," *Journal of Neural Engineering*, vol. 18, no. 4, p. 046040, 2021.

[12] B. Accou, M. J. Monesi, J. Montoya, T. Francart *et al.*, "Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 1175–1179.

[13] G. M. Di Liberto, J. A. O'sullivan, and E. C. Lalor, "Low-frequency cortical entrainment to speech reflects phoneme-level processing," *Current Biology*, vol. 25, no. 19, pp. 2457–2465, 2015.

[14] D. Lesenfants, J. Vanthornhout, E. Verschueren, and T. Francart, "Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech," *Journal of Neural Engineering*, vol. 16, no. 6, p. 066017, 2019.

[15] E. S. Teoh, F. Ahmed, and E. C. Lalor, "Attention differentially affects acoustic and phonetic feature encoding in a multispeaker environment," *Journal of Neuroscience*, vol. 42, no. 4, pp. 682–691, 2022.

[16] G. M. Di Liberto, D. Wong, G. A. Melnik, and A. de Cheveigné, "Low-frequency cortical responses to natural speech reflect probabilistic phonotactics," *Neuroimage*, vol. 196, pp. 237–247, 2019.

[17] D. Lesenfants, J. Vanthornhout, E. Verschueren, L. Decruy, and T. Francart, "Predicting individual speech intelligibility from the cortical tracking of acoustic-and phonetic-level speech representations," *Hearing Research*, vol. 380, pp. 1–9, 2019.

[18] C. Puffay, J. Van Canneyt, J. Vanthornhout, T. Francart *et al.*, "Relating the fundamental frequency of speech with EEG using a dilated convolutional network," *arXiv preprint arXiv:2207.01963*, 2022.

[19] C. Puffay, B. Accou, L. Bollens, M. J. Monesi, J. Vanthornhout, T. Francart *et al.*, "Relating EEG to continuous speech using deep neural networks: a review," *arXiv preprint arXiv:2302.01736*, 2023.

[20] C. Brodbeck, L. E. Hong, and J. Z. Simon, "Rapid transformation from auditory to linguistic representations of continuous speech," *Current Biology*, vol. 28, no. 24, pp. 3976–3983, 2018.

[21] M. Koskinen, M. Kurimo, J. Gross, A. Hyvärinen, and R. Hari, "Brain activity reflects the predictability of word sequences in listened continuous speech," *NeuroImage*, vol. 219, p. 116936, 2020.

[22] M. Kutas and K. D. Federmeier, "Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP)," *Annual Review of Psychology*, vol. 62, pp. 621–647, 2011.

[23] A. Soman, P. Ramachandran, and S. Ganapathy, "ERP evidences of rapid semantic learning in foreign language word comprehension," *Frontiers in Neuroscience*, p. 178, 2022.

[24] M. P. Broderick, A. J. Anderson, G. M. Di Liberto, M. J. Crosse, and E. C. Lalor, "Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech," *Current Biology*, vol. 28, no. 5, pp. 803–809, 2018.

[25] K. Gorman, J. Howell, and M. Wagner, "Prosodylab-aligner: A tool for forced alignment of laboratory speech," *Canadian Acoustics*, vol. 39, no. 3, pp. 192–193, 2011.

[26] A. Delorme and S. Makeig, "EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *Journal of Neuroscience Methods*, vol. 134, no. 1, pp. 9–21, 2004.

[27] M. J. Monesi, B. Accou, T. Francart, and H. Van hamme, "Extracting Different Levels of Speech Information from EEG Using an LSTM-Based Model," in *Proc. Interspeech 2021*, pp. 526–530.