Journal of Neural Engineering

PAPER • OPEN ACCESS

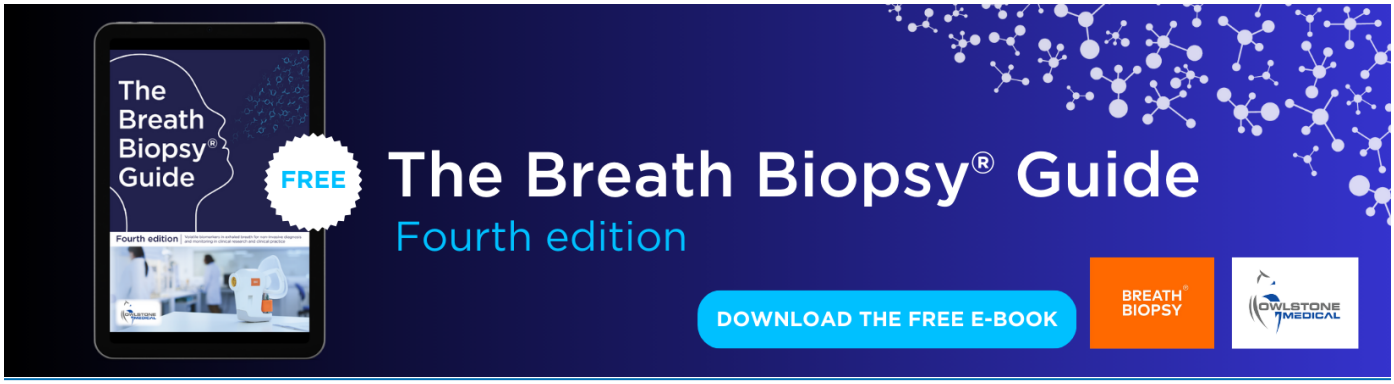
The role of vowel and consonant onsets in neural tracking of natural speech

To cite this article: Mohammad Jalilpour Monesi *et al* 2024 *J. Neural Eng.* **21** 016002

View the article online for updates and enhancements.

You may also like

- Data-driven machine learning models for decoding speech categorization from evoked brain responses
Md Sultan Mahmud, Mohammed Yeasin and Gavin M Bidelman
- Inferring the drivers of language change using spatial models
James Burridge and Tamsin Blaxter
- Direct classification of all American English phonemes using signals from functional speech motor cortex
Emily M Mugler, James L Patton, Robert D Flint et al.



This content was downloaded from IP address 87.67.244.63 on 29/01/2024 at 09:47

IOP Publishing J. Neural Eng. 21 (2024) 016002 https://doi.org/10.1088/1741-2552/ad1784

Journal of Neural Engineering

CrossMark

OPEN ACCESS

RECEIVED
22 June 2023

REVISED
31 October 2023

ACCEPTED FOR PUBLICATION
20 December 2023

PUBLISHED
11 January 2024

Original Content from this work may be used under the terms of the Creative Commons Attribution 4.0 licence.

Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.

PAPER

The role of vowel and consonant onsets in neural tracking of natural speech

Mohammad Jalilpour Monesi[1,2,*], **Jonas Vanthornhout**[2], **Tom Francart**[2] **and Hugo Van hamme**[1]

[1] Department of Electrical Engineering (ESAT), PSI, KU Leuven, Leuven, Belgium
[2] Department Neurosciences, ExpORL, KU Leuven, Leuven, Belgium
* Author to whom any correspondence should be addressed.

E-mail: mohammad.jalilpourmonesi@esat.kuleuven.be

Keywords: EEG, LSTM, vowel, consonant, match–mismatch, broad phonetic classes

Abstract

Objective. To investigate how the auditory system processes natural speech, models have been created to relate the electroencephalography (EEG) signal of a person listening to speech to various representations of the speech. Mainly the speech envelope has been used, but also phonetic representations. We investigated to which degree of granularity phonetic representations can be related to the EEG signal. *Approach*. We used recorded EEG signals from 105 subjects while they listened to fairy tale stories. We utilized speech representations, including onset of any phone, vowel–consonant onsets, broad phonetic class (BPC) onsets, and narrow phonetic class onsets, and related them to EEG using forward modeling and match–mismatch tasks. In forward modeling, we used a linear model to predict EEG from speech representations. In the match–mismatch task, we trained a long short term memory based model to determine which of two candidate speech segments matches with a given EEG segment. *Main results*. Our results show that vowel–consonant onsets outperform onsets of any phone in both tasks, which suggests that neural tracking of the vowel vs. consonant exists in the EEG to some degree. We also observed that vowel (syllable nucleus) onsets exhibit a more consistent representation in EEG compared to syllable onsets. *Significance*. Finally, our findings suggest that neural tracking previously thought to be associated with BPCs might actually originate from vowel–consonant onsets rather than the differentiation between different phonetic classes.

1. Introduction

The ability to decode speech from electroencephalography (EEG) signals has the potential to revolutionize diagnostic tests for hearing loss and other speech-related disorders. EEG allows for a non-invasive and cost-effective method of measuring brain activity related to speech perception and production. Despite numerous studies on auditory EEG, our understanding of neural mechanisms of underlying speech perception is still limited, and there is a need to develop new techniques to determine the relationship between EEG and speech perception or hearing.

A common approach involves recording EEG signals from individuals as they listen to natural running speech, and then attempting to relate the recorded EEG data to the presented speech (e.g. Crosse *et al* 2016, Vanthornhout *et al* 2018, Lesenfants *et al* 2019b). There are various methods for relating EEG to speech, but the most common approaches are backward, forward, and hybrid modeling. In backward modeling, usually, a model is used to reconstruct speech representations, such as the speech envelope or spectrogram, from EEG signals (e.g. Crosse *et al* 2016, Vanthornhout *et al* 2018, Verschueren *et al* 2019). In forward modeling, EEG responses are predicted from speech representations (e.g. Di Liberto *et al* 2015, Crosse *et al* 2016, Lesenfants *et al* 2019b). In both approaches, the correlation between the reconstructed or predicted and original signal is used as a proxy for the model's ability to relate EEG to the speech stimulus. In hybrid modeling, such as canonical correlation analysis, both speech and EEG signals are transformed to a common space where they have the maximum correlation with each other (de Cheveigne *et al* 2018, 2019). Similarly,

© 2024 The Author(s). Published by IOP Publishing Ltd

the correlation between the transformed EEG and the transformed speech is used as an evaluation metric for the hybrid model's ability to relate EEG to speech. It has been shown that these approaches can be used to assess speech understanding from EEG (Vanthornhout *et al* 2018, Lesenfants *et al* 2019a, Di Liberto *et al* 2021).

de Cheveigne *et al* (2018) introduced a novel paradigm called the match–mismatch task to relate EEG to speech. The task involves determining whether a segment of EEG matches a segment of speech. In other words, if the EEG segment was recorded while the subject was listening to the speech segment, then the (EEG, speech) pair is considered matched. Conversely, if the EEG was not recorded during the presentation of the speech segment, the (EEG, speech) pair is considered mismatched. Recently, models based on artificial neural networks (ANNs) have shown promising results on this task, outperforming the linear methods by a large margin. Monesi *et al* (2020) introduced a long short term memory (LSTM) based model to determine whether an (EEG, speech) pair is matched or mismatched. Similarly, Accou *et al* (2021a) introduced a model based on dilated convolutional layers with similar results as those of the LSTM-based model. In a follow-up study, the authors showed that it is possible to predict speech intelligibility based on match–mismatch classification accuracy of the dilated model (Accou *et al* 2021b). In another study, which uses the same match–mismatch task and the same dilated model from Accou *et al* (2021a), Puffay *et al* (2022) showed that it is possible to relate the fundamental frequency of the voice (f0) to EEG.

Most of the abovementioned studies use the envelope or the spectrogram as the representation for speech. The latter is a very informative representation since automatic speech recognition systems are capable of accurately transcribing from spectral representations (e.g. Hannun *et al* 2014, Hrinchuk *et al* 2019, Park *et al* 2019). The former is, however, less informative since it merely represents the instantaneous broadband energy. Yet, this representation of speech is quite effective in the match–mismatch task (Monesi *et al* 2020). However, it is unclear what level of detail in the speech representation is sufficient to be successful at the match–mismatch task. Rather than modifying the spectral resolution to control the informativeness of the speech representation, we opt for a linguistic representation of speech, which is more interpretable toward the end goal of diagnosing speech and language-related disorders. A first choice is that speech will be represented by phonetic classes, ranging from 37 narrow phonetic classes (NPCs) over five broad phonetic classes (BPCs) to the binary vowel versus consonant distinction. Notice that our stimulus data is (automatically) annotated with 37 NPCs that collapse allophonic variation in a single class, though pronunciation variants among the 37 narrow classes are automatically identified. Our data hence does not allow us to perform an accurate study about the perception of articulatory features such as aspiration. We restrict phonetic detail to a hierarchy of classes as mentioned above. A second choice regards the temporal encoding of phonetic representations. Apart from the speech versus silence property, phonetic representations are encoded at the phone onset. With this choice, repeated occurrences within the same phonetic class are still encoded. By contrast, an encoding of the class identity for the whole phone duration would merge consecutive phones of the same class into a single segment and onset information would be lost. The onset-based encoding hence allows to quantify the impact of class width definition. A third choice regards temporal resolution. We compare the high-rate phone-level with the low-rate syllable-level encoding of onsets. We apply linear models and ANNs to relate the above speech representations to the EEG signal. In the present study, it is unclear whether these models relate language-specific phonemic responses or acoustic phonetic responses from the brain with the discrete speech representations described above. It will also be clear from our analysis below that this distinction is hard to make, given the small differences in match–mismatch task performance observed between fine and broad phonetic classes. For this reason, we will refer to the above representations as *phonetic information*.

There are previous studies relating phonetic information to EEG (or electrocorticography) in which subjects listen to natural running speech. Mesgarani *et al* (2014) found that different locations in the superior temporal gyrus had different responses to different BPCs (referred to as phonetic features in the original study). Khalighinejad *et al* (2017) found that phones (called phonemes in the original study) of various phonetic categories (such as vowels, plosives, nasals, and fricatives) encode different phonetic distinctions across different time intervals with regard to phone onsets. Di Liberto *et al* (2015) reported that adding BPCs (called phonetic features in the original study) on top of the spectrogram yields better EEG predictions in linear forward modeling. In multiple follow-up studies (e.g. Di Liberto and Lalor 2017, Di Liberto *et al* 2018a, 2018b, Liberto *et al* 2023), authors reported that BPCs together with spectrograms yield better EEG predictions (higher correlations) than only using the spectrogram. Similar results were reported by Lesenfants *et al* (2019b). Prinsloo and Lalor (2022) reported that NPCs (phonemes) representations contribute to EEG prediction when speech is intelligible for listeners. In these studies that have investigated neural tracking of phonetic information, simpler speech representations such as phone onsets or vowel vs. consonant onsets have not been utilized. As a result, it is

unclear whether the reported increase in EEG prediction accuracy is due to the extra information provided in these phonetic classes or simply the presence of vowel vs. consonant or phone onsets. Daube *et al* (2019) reported that adding articulatory features on top of the spectrogram improves EEG predictions of the linear model in forward modeling. Interestingly, they observed a comparable increase in performance when using the phone (phoneme) onsets together with spectrograms. As a result, the authors concluded that the improved prediction accuracy is primarily attributed to the temporal information provided by phone onsets rather than the phonetic information present in different phonetic classes.

While there are other studies that relate EEG to phonetic representations, such as distinguishing between vowels and consonants, these studies have either used unnatural speech stimuli or measurement paradigms that involve articulating or imagining speech rather than simply listening to speech stimuli. For example, Parhi and Tewfik (2021) developed two models based on convolutional neural network (CNN) and LSTM neural networks to classify imagined vowels (i, a, u) from EEG. They showed that using only frontal lobe EEG channels, classification accuracy exceeded 85% for all of the subjects. In another study, Banerjee *et al* (2022) showed that applying principal component analysis to a CNN-based feature vector increases the classification of imagined vowels. In (Mahapatra and Bhuyan 2022), authors combined temporal and spatial features using CNNs in a deep learning model to classify imagined speech to five vowels and six words. They obtained an average accuracy of 96.49%, which is 11% higher than the state of the art model (Sarmiento *et al* 2021). In a different paradigm where subjects listened to separate vowels or syllables in the form of consonant–vowel (CV), Wang *et al* (2012) classified 8 consonants and 4 vowels from the recorded EEG. More specifically, they showed that it is possible to classify distinctive features such as voicing, continuant, place, height, and backness from the recorded EEG. In a similar study, Kovács *et al* (2017) studied the event-related potentials (ERPs) in response to different consonant classes. They reported that different ERP responses were elicited in response to different phonetic categories (fricatives, plosives, liquids, nasals, and affricates).

In this study, we aim to understand to which degree of granularity phonetic information can be related to the EEG signals recorded from subjects while they listened to a natural running speech. To this end, we tried to relate EEG to speech representations such as the onset of any phone, of vowels, of consonants, of BPCs and of NPCs, using two different tasks. In the forward modeling task, we used a linear model to predict EEG from these speech representations, while in the match–mismatch task, we trained an LSTM-based model to classify matched and mismatched (EEG, speech) pairs.

2. Methods

In this section, we will first provide details about the EEG dataset used in this study. Then, we will explain how we processed the EEG data and give an explanation of the speech representations that we extracted from the speech stimulus. Next, we will define the two tasks that were used as an evaluation metric to measure how well the speech stimulus can be related to EEG. The first task is the common approach of predicting EEG from the stimulus, also known as forward modeling. The second task is a recent approach, inspired by the auditory attention decoding literature, to relate EEG to the stimulus using a classification task called match–mismatch. Lastly, we will introduce the LSTM-based model (Monesi *et al* 2020, 2021) that was used in the match–mismatch task.

2.1. Dataset

EEG signals were recorded from subjects while they listened to a natural running speech in the form of a fairy tale story. Subjects were normal-hearing Flemish speaking in the [18–30] age group. Subjects gave their informed consent, which was reviewed and approved by the Medical Ethics Committee at KU Leuven in Belgium, under reference number S57102. First, participants were screened for a normal hearing with pure tone audiometry and Flemish matrix-test (Luts *et al* 2014). Only participants with normal hearing (hearing threshold <25 dBHL) went through EEG measurements.

For this study, we used EEG recordings of 105 subjects who listened to fairy tale stories. For most subjects, we had 7 or 8 recordings. For 9 subjects, we had less than 7 recordings due to measurement errors. Each story was approximately 14 min and 30 s in length, and the order of presentation was randomized. Participants were given breaks between recordings. The stimuli were presented using the APEX 4 software developed at ExpORL (Francart *et al* 2008) and delivered binaurally through Etymotic ER-3A insert phones at 62 dBA. To encourage attentive listening, participants were informed that they would be asked a question about the content of the story. Their responses were not used in this study. EEG signals were recorded using the 64 channel ActiveTwo BioSemi system with a 8192 Hz sampling rate, within an electromagnetically shielded and soundproofed cabin. The EEG recordings were synchronized with the stimuli after each session. For more detailed information about the dataset, please refer to Bollens *et al* (2023).

2.2. Preprocessing

EEG: the EEG signals were first high-pass filtered at 0.5 Hz for drift correction using a fourth-order Butterworth filter. Matlab's filtfilt function was used to have zero-phase filtering. The recordings were then downsampled from 8192 Hz to 1024 Hz using

Table 1. Summary of speech representations used in this study.

Speech representation	Dimension	Explanation
VAD	1	One for speech activity, zero otherwise
NPC	37	Narrow phonetic classes onsets
BPC	5	Contains onsets of five categories: short vowels, long vowels, plosives, fricatives, nasals and approximants
Vowel–consonant	2	Vowel and consonant onsets
Phone	1	Phone onsets, it is one for the onsets of any phone and zero otherwise
Vowel	1	Vowel onsets
Consonant	1	Consonant onsets
Syllable	1	Syllable onsets

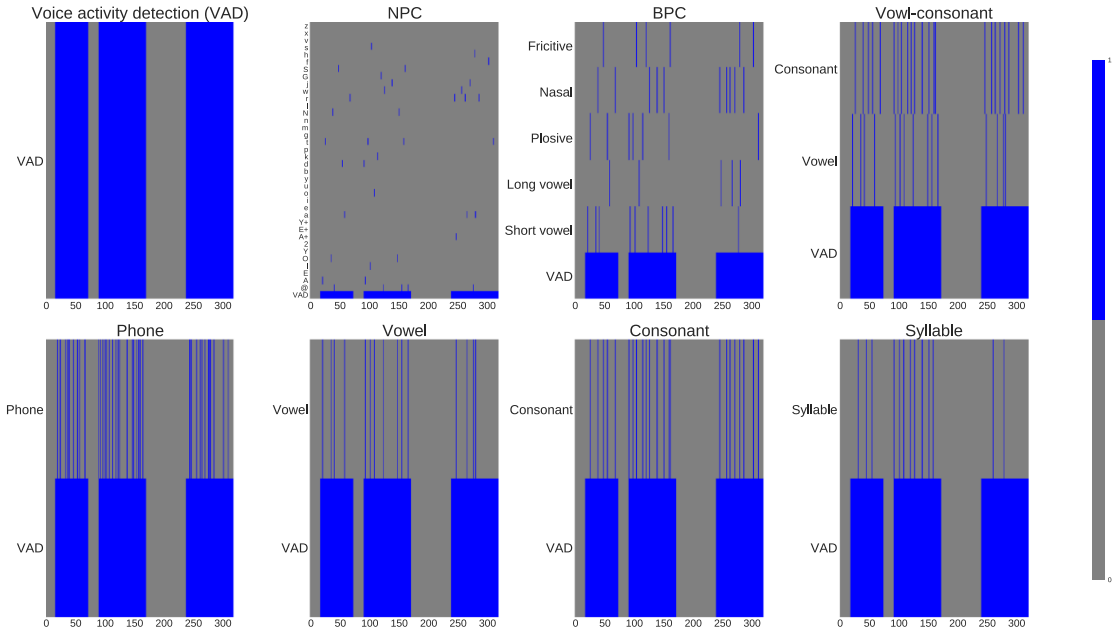


Figure 1. The speech representations employed in this study are displayed over a 5 s interval. Note that VAD is added to each of the other representations as baseline speech information. With the exception of the syllable onsets, all of the speech representations were calculated based on narrow phonetic classes (NPCs) onsets.

Matlab's resample function, which includes an anti-aliasing low-pass filter prior to resampling. After removing eyeblink artifacts from the EEG signals using the multi-channel Wiener filter (Somers *et al* 2018), all EEG channels were re-referenced to the common average of the channels. The EEG signals were then downsampled again to 64 Hz using Matlab's resample function. The resulting EEG signals had a bandwidth between 0.5 and 32 Hz. It is worth mentioning that in the preprocessing step, no EEG sections were removed. Matlab version 2022a was used in this study.

Speech representations: a summary and visualization of each speech representation is provided in table 1 and figure 1, respectively. The following speech representations were used in this study.

(1) Voice activity detection (VAD): this one-dimensional feature distinguishes speech from silence, with a value of 1 indicating speech and 0 indicating silence. This distinction was shown by Monesi *et al* (2021) to be the main contributor to match–mismatch accuracy. As a result, we used this feature as baseline information and thus added it as a first dimension to all of the following speech representations. It is worth noting that in this study, VAD was calculated based on phone alignment rather than speech energy thresholding.

(2) NPCs onsets: using forced alignment developed especially for reading tasks (Duchateau *et al* 2009), we segmented each story into a sequence of phone units. The story text is mapped to a hidden Markov model that allows for disfluencies by the narrator, including word restarts and repetition of and skipping over several words. The HMM also chooses the most likely among several pronunciations. The pronunciation model accounts for coarticulation effects such as degemination. The model, however, does not distinguish in fine allophonic detail (e.g. realization of /r/ as uvular trill, alveolar trill

or alveolar flap) and hence we term the units as *NPCs*. Based on the International Phonetic Alphabet, we had 39 NPC (phones) for the Dutch language but we removed two phones as they did not appear in our data. Then for each of the 37 NPC, we converted the symbol to a 37 dimensional one-hot vector for each time sample. More specifically, the encoding was done at phone onsets to preserve the duration between phones. The following speech representations (except for the syllable onsets) are all derived from NPC onsets by decreasing the granularity level of phonetic information.

(3) BPC onsets: following Lesenfants *et al* (2019b), we started from NPC and grouped vowels into two categories (short vowels and long vowels) and consonants into three categories (plosives, fricatives, and nasals/approximants) based on the manner of articulation. As a result, BPC is a five-dimensional one-hot vector for each time sample. Note that compared to NPC, BPC contains less phonetic information.

(4) Vowel–consonant onsets: we created a two-dimensional one-hot vector from BPC onsets by grouping vowels and consonants. Note that this representation is derived from BPC onsets and contains less phonetic information compared to BPC onsets.

(5) Phone onsets: we combined vowels and consonants in the vowel–consonant onsets representation to one category called phone onsets. This one-dimensional representation does not contain the discrimination between vowels and consonants.

(6) Vowel onsets: this representation is directly derived from vowel–consonant onsets by excluding consonant onsets. As a result, this one-dimensional feature contains only vowel (syllable nucleus) onsets.

(7) Consonant onsets: this representation is directly derived from vowel–consonant onsets by excluding vowel onsets. As a result, this one-dimensional feature contains only consonant onsets.

(8) Syllable onsets: this representation is one at syllable onsets and zero otherwise.

The objective behind developing simplified phonetic representations is to examine the level of granularity at which the model can effectively leverage phonetic information to relate EEG to speech. All the speech representations were calculated at a 64 Hz sampling rate so they are synchronous with the EEG.

Split and normalization: each individual EEG recording (for most subjects, we had 7 or 8 recordings of approximately 14 min and 30 s each) was split into a training set (80%), a validation set (10%), and a test set (10%). More specifically, the data was split into (40%, 10%, 10%, 40%) portions. The train set contained 40% of the recording at the start and 40% of the data at the end of the recording. Validation and test sets were selected from the middle of the recording in order to prevent any artifacts that might exist at the beginning or end of the recording. Furthermore, for some stimuli, there are brief periods of silence before the main story starts. However, this issue is less significant as it only impacts the initial few seconds. Lastly, each recording was mean–variance normalized. The mean and variance of the training set were used to normalize training, validation, and test sets.

2.3. Linear forward modeling

We used forward modeling (e.g. Crosse *et al* 2016, Lesenfants *et al* 2019b) as one of the evaluation methods to relate EEG to speech. In this approach, a linear model was used to predict the recorded EEG from the speech stimulus. After predicting the EEG, the correlation between the recorded EEG and the predicted EEG was used as a measure of performance. We used ridge regression as a regularization method. The analytical solution of the linear forward model with ridge regression regularization is as follows:

$$W = \left(S^T S + \lambda I\right)^{-1} S^T R \tag{1}$$

S is the lagged time series of the speech representation and hence takes the delay of the brain responses with regard to the stimulus into account. The integration window was set to 400 ms, meaning that speech samples up to 400 ms prior to the EEG time sample were used for prediction. R represents the recorded EEG signal and I is the identity matrix . Finally, W is the linear model, which is also known as the temporal response function (TRF). The value of λ is usually determined based on a validation set or k-fold cross-validation. In this study, the value of the λ was chosen based on the performance of the linear model on the validation set. The Spearman correlation between the predicted EEG channel and the actual EEG channel was used as a performance metric for this task.

2.4. Deep-learning model in a match–mismatch task

In this section, we will explain the second task used in this study to relate EEG to the speech stimulus. This binary classification task, referred to as the match–mismatch task, has been used in recent studies to relate EEG to the presented speech stimulus (e.g. de Cheveigné *et al* 2020, Monesi *et al* 2020, Accou *et al* 2021b). The task involves determining whether a segment of a recorded EEG is aligned with a presented speech stimulus. The EEG and speech segments are considered a match if the EEG was recorded while the speech was being presented, meaning they are temporally aligned. Conversely, the EEG and speech segments are considered a mismatch if there is no temporal alignment between the two. The procedure for extracting matched and mismatched (EEG, speech)

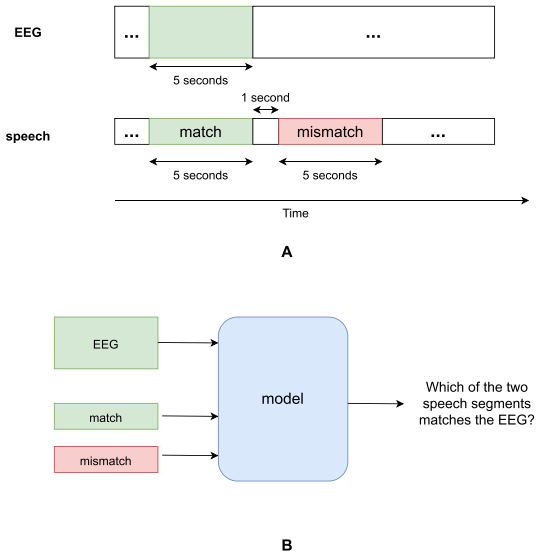


Figure 2. (A) Extracting matched and mismatched speech segments with respect to an EEG segment of 5 s (decision window). (B) Match–mismatch classification task where a model is trained to determine which of the two given speech candidates matches the provided EEG.

pairs is illustrated in figure 2(A). It depicts (EEG, speech) pairs of 5 s length (referred to as the decision window), extracted with 80% overlap. For the mismatched (EEG, speech) pair, the speech is taken 1 s after the end of the EEG segment (in the future).

After defining matched and mismatched pairs, we can now define the final match–mismatch classification task. As illustrated in figure 2(B), the task is a binary classification where the model is trained to determine which of two candidate speech inputs matches the input EEG segment. It should be noted that always one of the speech inputs is a matched pair with the EEG input, and the other one is mismatched. Moreover, the position of the matched and mismatched speech inputs is alternated to prevent the model from consistently identifying one of the inputs as matched.

Note that one could use other match–mismatch classification tasks than the one defined here. For example, Monesi *et al* (2020) used a match–mismatch task in which the task of the model was to determine whether a given (EEG, speech) pair is matched or mismatched. This choice mainly affects the level of difficulty of the task. We opted for the match–mismatch task defined here as it is relatively simple to implement and it allows comparison with earlier studies (Monesi *et al* 2021, Accou *et al* 2021a, Puffay *et al* 2023).

In order to train a generalizable model, the mismatched speech (similar to hard negative sampling in the contrastive learning literature; Oord *et al* (2019)) should be difficult enough. Therefore, the mismatched speech segment is taken from the same stimulus and also temporally close to the matched speech. In this way, the mismatched speech has some general similarities to the matched speech, such as originating from the same stimulus speaker or similar listener fatigue. Thus, the model needs to rely on the content of the two candidate speech segments to determine which one is matched.

Model: we have used an LSTM-based model introduced by Monesi *et al* (2020) and slightly adjusted by Monesi *et al* (2021) to do the match–mismatch task. The architecture of this model is shown in figure 3. The model includes two networks to project each of the inputs into an embedding space. The idea is that in the embedding space, EEG and matched speech have similar representations for each time frame, while EEG and mismatched speech have dissimilar representations in the embedding space. Here, the similarity is measured as cosine similarity, which is calculated for each time frame with the assumption that there is enough information to align these short segments of EEG and speech. We opt for an LSTM in the speech-related network path because apart from being able to model non-linearities, it can also model the delayed brain response observed in the EEG signals. Ideally, the LSTM layer has the capability to introduce a delay to the speech input and therefore synchronize it with the EEG stream (Monesi *et al* 2020). The layers in the speech path are shared between the two speech candidates.

The hyperparameters of the model are shown in figure 3. These hyperparameters were tuned for mel spectrogram representation in a previous study (Monesi *et al* 2021). In this study, only the kernel size of the conv2D layer was tuned to have an appropriate receptive field for each speech representation. Kernel sizes in the range [9–29], which equals a receptive field in the range [140 ms, 450 ms], were used in this study. The model has about 94 000 learnable parameters in total. During the training, we used 30 epochs with early stopping. The code of the model is provided at https://github.com/exporl/match-mismatch-LSTM.

3. Results

We aim to identify which phonetic information in the presented speech stimulus can be related to the EEG response. We have defined eight speech representations with varying levels of phonetic information, as described in the methods section. In the first part of this section, we will show the results of relating these speech representations to the recorded EEG using linear forward modeling. To do so, correlation scores between predicted EEG channels and actual EEG channels will be reported. TRFs as well as correlation and TRF topoplots will be shown for further interpretation. In the second part of this section, the results of relating EEG to these speech representations in the match–mismatch task will be discussed. In order to determine whether there is a significant difference between the means of two groups,

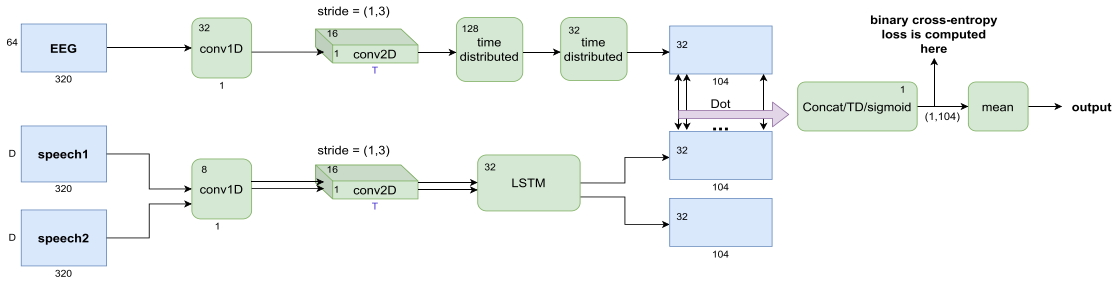


Figure 3. The LSTM-based model (Monesi *et al* 2021) that is used in match–mismatch classification. The *Dot* layer applies a normalized dot product (cosine similarity) between EEG and speech representations for each time step. The green color indicates a model layer, while the blue color indicates the input or output of a layer. The selected hyperparameters, such as kernel size or number of units, are shown for each layer. The number of frames at the output (number 104 in the figure) corresponds to a conv2D kernel size of 9 ($T = 9$).

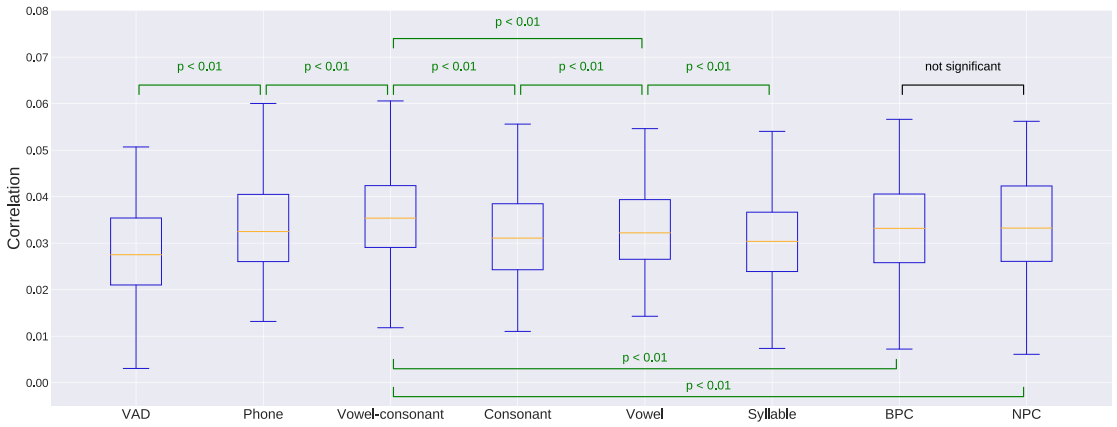


Figure 4. The Spearman correlations of the linear model for phonetic representations. Box plots are shown over 105 subjects. The model is trained separately for each subject (subject-specific model). Statistical significance tests are only shown for important pairs to simplify the figure. *P* values are corrected for multiple comparisons using the Holm–Bonferroni method.

we used a signed-rank Wilcoxon test throughout this section. All the p-values are corrected for multiple comparisons using the Holm–Bonferroni method (Holm 1979).

3.1. Relating phonetic information to EEG using linear forward modeling

With the linear forward modeling approach, EEG signals are predicted from speech representations (regression task) using a linear model. Then, the correlation between the actual EEG and the predicted EEG is used as a proxy for neural tracking of speech. Similar to Lesenfants *et al* (2019b), we included 27 fronto-temporal channels.

The phonetic information within each speech representation was systematically controlled, varying from the lowest level of granularity (phone onsets) to the highest level (NPC onsets). In terms of phonetic information, we have the relation: phone $<$ vowel–consonant $<$ BPC $<$ NPC. Also, each speech representation contains the previous representation's information. For example, BPC contains all the information available in the vowel–consonant representation plus some extra phonetic information, related to manner of articulation for consonants and duration for vowels.

Correlation scores between predicted EEG and actual recorded EEG signals are shown in figure 4 for each speech representation. We see that all the speech representations outperform the baseline VAD representation (all $p < 0.01$). This shows that the model is able to use the extra phonetic information on top of speech vs. silence to make better EEG predictions. Interestingly, vowel–consonant onsets performs better than phone onsets ($p < 0.01$). This means that providing vowel and consonant discrimination helps the model to have a better EEG prediction. As shown in figure 4, vowel–consonant onsets outperforms vowel onsets and consonant onsets ($p < 0.01$). This suggests that both vowel and consonant onsets contribute to EEG predictions. Also, notice that vowel onsets are better related to EEG than consonant onsets ($p < 0.01$).

We also included syllable onsets in our experiments to compare it with vowel (syllable nucleus) onsets. As seen in the figure, vowel onsets outperform syllable onsets (p-value < 0.01). Finally, we observe that adding more phonetic information on top of vowel–consonant onsets does not result

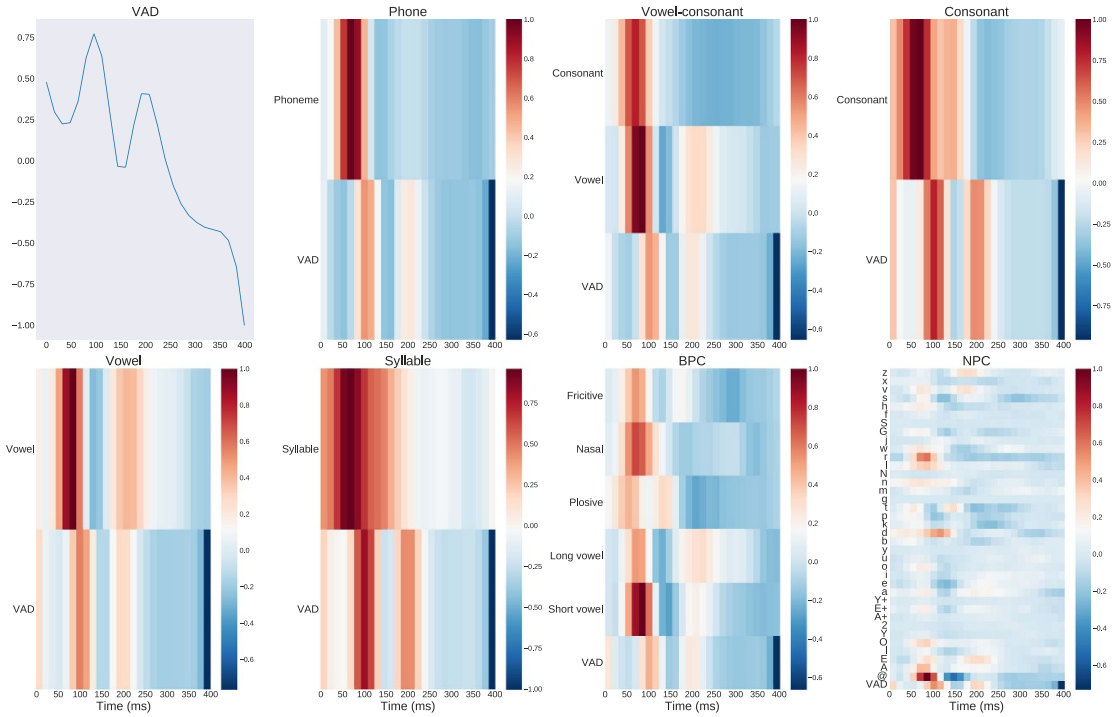


Figure 5. Average TRFs over 105 subjects are displayed for different speech representations.

in an improvement in EEG predictions. BPC and NPC representations perform worse ($p < 0.01$) than vowel–consonant onsets despite having more phonetic information. In theory, the linear models using the BPC and NPC representations should be able to achieve a fit to the data that is as good as, if not better than, linear models using the vowel–consonant representations. However, this is not the case due to three causes: (i) we work with finite data sets, which causes variance in estimation and evaluation. The filters in the NPC and BPC models are estimated on fewer data points (due to more classes in these representations) than in the CV model, leading to higher variance. (ii) Regularization—required due to finite data sets—results in models that do not minimize mean squared error (MSE) (see formula (1)) (iii) The models are trained for minimal MSE, but evaluated on correlation. In other words, a lower MSE loss does not necessarily mean a higher correlation.

To conclude, we investigated which phonetic information can be used by the linear model to predict the EEG. The results indicate that all speech representations outperform the baseline VAD representation, with vowel–consonant onsets representation performing the best. The main finding was that vowel–consonant onsets contribute to the predictions of the EEG on top of phone onsets and suggest that neural markers of vowel vs. consonant exist in the EEG signals. The study also finds that vowel onsets are better related to EEG than syllable onsets. Furthermore, increasing the granularity level of phonetic information beyond vowel–consonant does not improve EEG predictions.

3.1.1. TRFs

TRFs are often used in forward models to analyze the latency of brain responses to speech. In figure 5, average TRFs over all 105 subjects are shown for each of the speech representations. We see that for the baseline VAD representation, there are two positive peaks at 100 ms and 200 ms (VAD subplot in figure 5).

The primary positive peak of the VAD representation occurs at approximately 100 ms, while it occurs earlier for other representations, such as consonants and vowels. For instance, the main and only positive peak for phone onsets occurs around 75 ms (phone subplot in figure 5). Additionally, we observe that the vowel onsets representation peaks later than the consonant onsets representation (vowel–consonant subplot in figure 5). Unlike vowel onsets, consonant onsets exhibit only one positive peak in the TRF. Furthermore, the first positive peak of the vowel TRF appears to be stronger than that of the consonant in the TRF of vowel–consonant onsets. This indicates that the linear model assigns greater importance to vowel onsets than to consonant onsets. This observation might provide additional evidence to support the notion that EEG tracking of vowel onsets is easier than that of consonant onsets, as demonstrated not only by correlations (as shown in figure 4) but also by the weights assigned by the model (vowel–consonant subplot in figure 5).

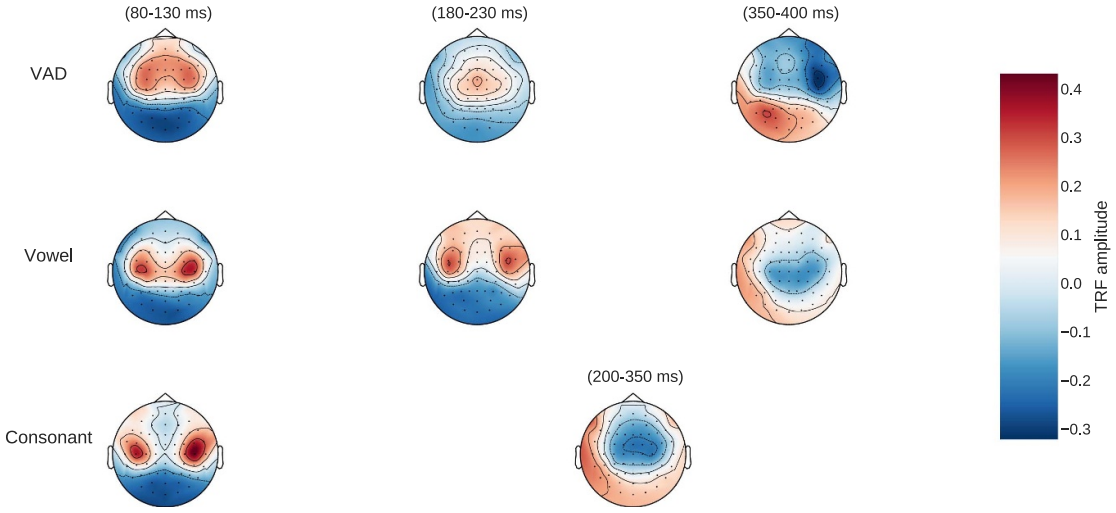


Figure 6. TRF topoplots over 64 channels are shown for VAD, vowel onsets, and consonant onsets (vowel–consonant onset representation is used here). For VAD and vowel onsets the average TRF values are presented over two positive peaks, [80–130 ms] and [180–230 ms], and a negative peak [350–400 ms]. For consonant onsets, TRF values are shown for one positive peak [80–130 ms] and one negative peak [200–350 ms].

The TRF for syllable onsets is comparable to that of consonant onsets, which is not surprising considering that most (around 60%) syllable onsets are consonant onsets rather than vowel onsets. The positive peak in the TRF of both consonant and syllable onsets appears to be broad. Among the BPC categories, short vowels and nasals seem to elicit the strongest TRF responses. Overall, the TRF peaks of BPC are in line with those reported in a similar study by Lesenfants *et al* (2019b). It is quite difficult to interpret the TRF of a sparse representation such as NPC since some of the classes (phones) occur rarely in the training set. Nevertheless, it seems that vowels yield two positive peaks, and most consonants yield one positive peak, which is in line with the TRF of other representations in this study.

3.1.2. EEG topoplots

In order to know which EEG channels have stronger TRF amplitudes during positive and negative peaks, usually TRF topoplots are used. In figure 6, TRF topoplots are shown for sub-components of the vowel–consonant representation, namely VAD, vowel onsets, and consonant onsets. VAD and vowel onset TRFs have two positive and one negative peak, while consonant onsets have one positive and one negative peak. As a result, time intervals [80–130 ms], [180–230 ms], and [350–400 ms] are chosen for VAD and vowel onsets, while [80–130 ms] and [200–350 ms] are chosen for consonant onsets.

During the first positive peak interval of [80–130 ms], high positive weights are seen in both temporal and central channels and high negative weights are seen in occipital channels for all three representations. For the [180–230 ms] time interval, higher model weights are concentrated in central channels for VAD and in temporal channels for vowel onsets. The TRF weight pattern observed during the negative peak interval of vowel onsets ([350–400]) is found to be the same as that observed during consonant onsets ([200–350]). Specifically, high negative weights are noted in the central channels, while high positive weights are noted in the left temporal and occipital channels. For the VAD TRF, the right temporal and frontal channels have negative values, while the left temporal and occipital channels have positive values.

Finally, we also provide EEG prediction topoplots in figure 7 to see which EEG channels are predicted more correctly than others. Figure 7 presents the correlation values per channel, which have been averaged over all subjects. The highest correlations are found in the left and right temporal channels, as well as in some central channels, while occipital and frontal channels have lower correlations. This indicates that the temporal channels are easier to predict, which is plausible as the primary auditory cortex is located in the temporal lobe of the brain.

3.2. Relating phonetic information to EEG using a deep learning model in the match–mismatch task

In this section, we relate the recorded EEG signals to the presented speech in the match–mismatch task using the LSTM-based model shown in figure 3. First, we trained a subject-independent (SI) model using data from all the subjects. The data split was similar to the previous section (train: 80%, validation: 10%, and test: 10%). Then, we fine-tuned the SI model for each subject using the subject's training and validation sets. It has been shown in Monesi *et al* (2020) that further fine-tuning on each subject improves the model's performance. We trained the model using the gradient descent method with the ADAM optimizer. The model was trained for 30 epochs with early stopping. If the validation loss increased for five consecutive

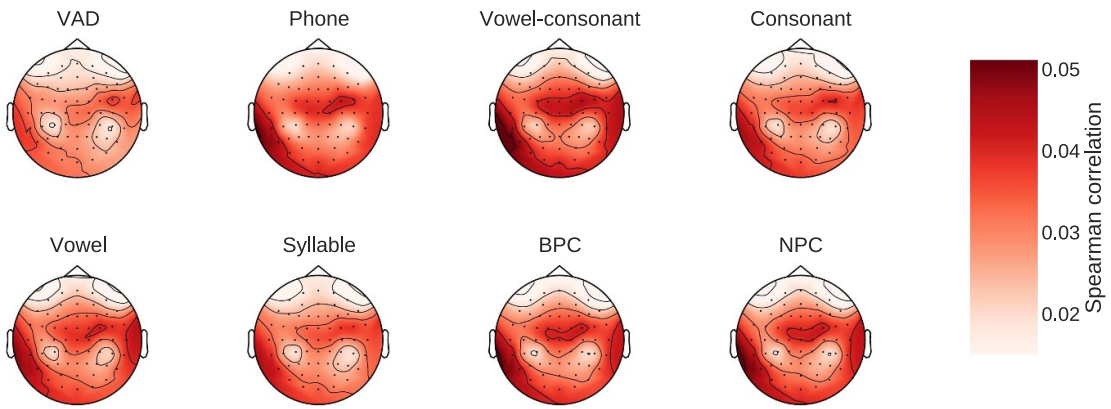


Figure 7. Topoplot representation of EEG channel prediction. The predictions are averaged over all the subjects.

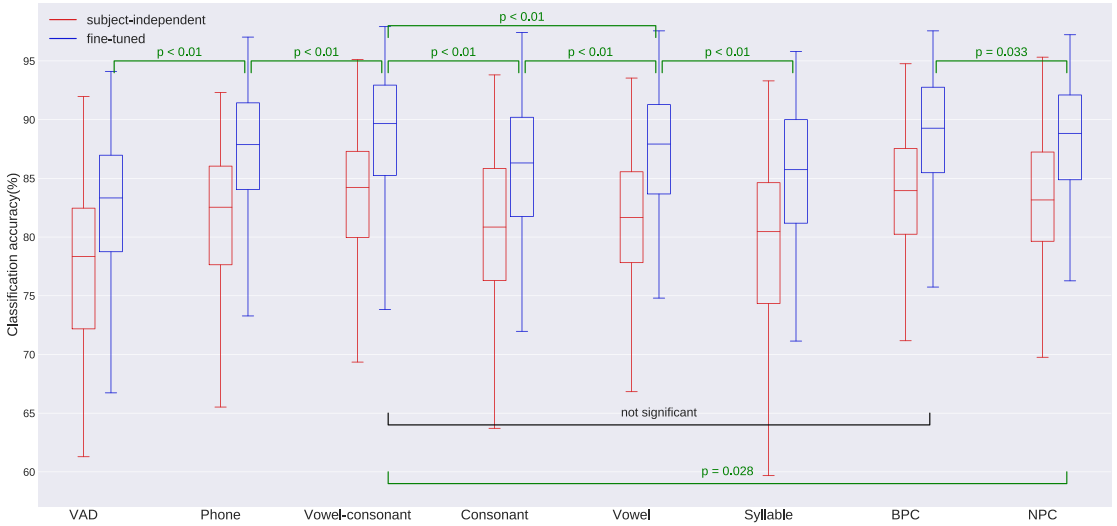


Figure 8. Match–mismatch accuracy of the subject independent and fine-tuned models for each of the speech representations using decision windows of 5 s. Box plots are shown over 105 subjects. Statistical significance tests are only presented between some of the pairs (important pairs such as phone onsets vs vowel–consonant onsets) to maintain clarity in the figure. Some of the other differences are also statistically significant.

epochs, the training would stop. The speech representations used in this section were identical to those utilized in linear forward modeling. Note that both the SI model and the fine-tuned models were evaluated on the same test set.

The match–mismatch classification accuracy for each speech representation is shown in figure 8. Similar to the linear forward modeling results, we see that all the speech representations outperform the baseline VAD representation ($p < 0.01$). This implies that the model is able to find neural tracking of this extra phonetic information in the EEG. Note that for all speech representations, training subject-specific models by fine-tuning increases the classification accuracy by around 5%. Since we have the same relative accuracy difference between representations in both the SI and fine-tuned scenarios, we will only consider the fine-tuned model's results when comparing speech representations.

We observe that vowel–consonant onsets reach 84% median accuracy compared to 82.5% median accuracy of phone onsets ($p < 0.01$). This supports the notion that discriminating between vowels and consonants helps the model extract useful information from the EEG to do the match–mismatch task. As with linear models, removing either vowel or consonant onsets decreases performance ($p < 0.01$). Once again, we observe that vowel onsets are easier related to EEG than consonant onsets ($p < 0.01$).

The results further show that the more detailed phonetic representations, namely BPC and NPC onsets, do not surpass the performance of the vowel–consonant onsets representation. This suggests that the LSTM-based model is unable to extract this extra phonetic information from the EEG.

Another question we aimed to answer was whether vowel (syllable nucleus) onsets are related better to EEG compared to syllable onsets. The results,

as shown in figure 8, indicate that vowel onsets outperform syllable onsets ($p < 0.01$).

Overall, the results are consistent with those of the forward modeling section. Vowel–consonant onsets outperforms phone onsets suggesting that the model is able to extract information about vowels and consonants from the EEG. Additionally, vowel onsets were better related to EEG signals than syllable onsets. Furthermore, fine-tuning the model for each individual showed a 5% improvement in accuracy.

4. Discussion

The purpose of this study was to investigate to which granularity level phonetic information of the presented stimulus can be related to the recorded EEG. To achieve this goal, we utilized various phonetic representations, including phone onsets, vowel–consonant onsets, syllable onsets, BPC onsets, and NPC onsets. We started from the most detailed phonetic representation, which was the NPC onsets, and progressively created representations with decreasing granularity of phonetic information. In other words, phone onsets contains the least information while NPC onsets has the most phonetic information. We included speech vs. silence information (VAD) to all representations, following the findings of Monesi *et al* (2021) that silences play a key role in the match–mismatch task. By doing so, we ensured that each speech representation contains speech vs. silence information, allowing us to determine that any improvements in accuracy or correlations are a result of the phonetic representation and not just VAD information present in the representation. We employed forward modeling (EEG prediction) and match–mismatch tasks to relate EEG recordings to speech representations.

We found that the vowel vs. consonant information is represented in the recorded EEG signals. Both the forward modeling and match–mismatch tasks indicated that including vowel vs. consonant information over phone onsets results in improved performance. This suggests that the difference between vowels and consonants, to some degree, is encoded in the recorded EEG signal. Previous research by Khalighinejad *et al* (2017) demonstrated that EEG responses of vowels differ from those of the consonant categories such as nasals, plosives, and fricatives at various time intervals relative to the phone onset. However, they compared the average time-locked EEG responses of each phonetic group, while our study utilized forward modeling and match–mismatch tasks to determine if phonetic representations can be related to EEG recordings.

Previous studies have reported that incorporating phonetic information, such as BPC (sometimes called phonetic features) on top of the spectrogram or envelope can lead to better EEG predictions (Di Liberto *et al* 2015, Lesenfants *et al* 2019b) or match–mismatch accuracy (Monesi *et al* 2021), respectively. However, in these studies, a vowel vs. consonant onsets representation was not used. The key point is to use speech representations that contain the onset of each phone such that the duration of each phone is available, which was not the case in these studies. For example, in our prior work (Monesi *et al* 2021), we did use a vowel–consonant representation, but the speech representations, including vowel–consonant, were one-hot for the duration of the vowel or consonant. The problem with this sort of 'duration' based encoding, in contrast to 'onset' based encoding, is that sometimes the phone duration information is lost. For example, if there are two consecutive consonants, the model cannot know the duration of each consonant since it sees a one-hot vector for the whole duration of the two consonants. However, if one uses a phonetic representation such as BPC where consonants (and vowels) belong to different categories, the duration information will be available most of the time unless the consecutive consonants are from the same phonetic class. Therefore, it is not clear whether the reported neural tracking of BPC representation comes from relating actual distinction between phonetic classes to EEG or it comes from relating vowel–consonant onsets, which contains vowel vs. consonant distinction plus the duration of each phone, to EEG.

To investigate the question, we made sure in our speech representations the duration of phones is available by using a temporal encoding based on phone onsets (i.e. one-hot vectors only at the onsets, see figure 1) as well as using speech representations such as vowel–consonant onsets and phone onsets. Our results in this study suggest that the actual contribution does not come from the distinction of BPCs but rather from the vowel vs. consonant onsets. More specifically, in both the forward modeling and match–mismatch tasks, vowel–consonant onsets performed equal or better than BPC and NPC representations indicating that the extra distinction between phonetic classes could not be related to EEG beyond vowel–consonant distinction. With that said, future research is required to investigate the impact of both model selection and the choice of phonetic categories to arrive at a more precise conclusion. Furthermore, we strongly advocate for researchers to incorporate simpler speech representations such as vowel–consonant onsets as controls in their experimental designs. Moreover, we encourage researchers to use onset-based encoding for phonetic representations (in contrast to duration-based encoding) to preserve the time between phone onsets. Furthermore, it seems that both the linear model in forward modeling and the LSTM-based model in the match–mismatch task struggle to work with more sparse speech representations such as NPC. One explanation might be that we need even more training data such that there are enough samples for some of the classes (phones) that occur rarely. This being said, this study already

used a substantial EEG dataset and obtaining even a bigger dataset is quite challenging.

We also compared syllable onsets to vowel (syllable nucleus) onsets. We found that in both forward modeling and match–mismatch tasks, vowel onsets were better related to the recorded EEG than syllable onsets. Our findings are consistent with those of Oganian and Chang (2019), who used a speech representation based on a half-rectified speech envelope derivative (envelope peak rates) and reported that it outperformed the envelope in the forward modeling task. They showed that envelope peak rates are more aligned to vowel onsets than syllable onsets.

In our analysis of the linear forward modeling task, we delved further into the TRF of speech representations. Our observations indicated that the majority of the representations displayed a positive peak before 100 ms. Our results were consistent with the TRFs of NPC (called phonemes) and BPC (phonetic features) reported by Lesenfants *et al* (2019b) and Di Liberto *et al* (2015). However, our baseline VAD representation was found to have two positive peaks at around 100 and 200 ms and a negative peak at 400 ms. The timing and polarity of the peaks differ from those reported in studies by Di Liberto *et al* (2015) and Lesenfants *et al* (2019b), who found a negative peak at 80 ms and a positive peak at 150 ms for the speech envelope. The differences in the results of our study and the mentioned studies may stem from the difference in the speech representations used (VAD versus envelope) as well as the variation in the regularization matrices. We used an identity matrix in equation (1), while the studies mentioned above used a matrix that penalizes the difference between neighboring terms (Crosse *et al* 2016). Additionally, we observed that the first peak of vowel onsets occurs after that of consonant onsets, which we hypothesize could be due to the fact that the energy of many consonants is concentrated closer to the onset, whereas vowel energy is more widely spread.

5. Conclusion

This study aimed to determine which granularity of phonetic information could be effectively related to EEG signals. We employed speech representations based on NPC onsets, preserving the duration between phones. Notably, our results revealed that the vowel–consonant onsets representation outperformed phone onsets in both forward modeling and match–mismatch tasks. This finding indicates that vowel vs. consonant information is represented in EEG.

Furthermore, we explored whether the extra phonetic information present in BPCs could be related to EEG. Interestingly, our results indicated that providing additional phonetic details beyond vowel vs. consonant did not yield improvements in the models' performance for either task. This suggests that the previously reported neural tracking associated with BPCs in studies such as (Di Liberto *et al* 2015, Lesenfants *et al* 2019b, Monesi *et al* 2021) may not be attributed to the full phonetic information present in these phonetic classes, but rather to the information present in vowel vs. consonant onsets.

Data availability statement

The study uses EEG recording from subjects whose data we cannot share publically. The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgment

Funding for this project is provided by the KU Leuven Special Research Fund C24/18/099 (C2 project to Tom Francart and Hugo Van hamme) and the Flemish Government under 'Onderzoeksprogramma AI Vlaanderen'. This project has also received funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation programme (Grant Agreement No 637424, ERC starting Grant to Tom Francart).

ORCID iDs

Mohammad Jalilpour Monesi https://orcid.org/0000-0002-9217-6745
Jonas Vanthornhout https://orcid.org/0000-0002-1503-599X

References

Accou B, Jalilpour Monesi M, Montoya J, Van hamme H and Francart T 2021a Modeling the relationship between acoustic stimulus and EEG with a dilated convolutional neural network *2020 28th European Signal Processing Conf. (EUSIPCO)* pp 1175–9

Accou B, Jalilpour Monesi M, Van hamme H and Francart T 2021b Predicting speech intelligibility from EEG in a non-linear classification paradigm *J. Neural Eng.* **18** 066008

Banerjee O, Govind D, Kumar Dubey A and Gangashetty S V 2022 Significance of Dimensionality Reduction in CNN-Based Vowel Classification from Imagined Speech Using Electroencephalogram Signals *Speech and Computer* (*Lecture Notes in Computer Science*) ed S R M Prasanna, A Karpov, K Samudravijaya and S S Agrawal (Springer International Publishing) pp 44–55

Bollens L, Accou B, Van hamme H and Francart T 2023 A large auditory EEG decoding dataset (https://doi.org/10.48804/K3VSND)

Crosse M J, Di Liberto G M, Bednar A and Lalor E C 2016 The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli *Front. Hum. Neurosci.* **10** 604

Daube C, Ince R A A and Gross J 2019 Simple acoustic features can explain phoneme-based predictions of cortical responses to speech *Curr. Biol.* **29** 1924–37.e9

de Cheveigne A, Di Liberto G M, Arzounian D, Wong D D E, Hjortkjaer J, Fuglsang S and Parra L C 2019 Multiway canonical correlation analysis of brain data *NeuroImage* **186** 728–40

de Cheveigné A, Slaney M, Fuglsang S A, and Hjortkjaer J 2020 Auditory stimulus-response modeling with a

match-mismatch task *bioRxiv* https://doi.org/10.1101/2020.11.05.370072 (Posted online 6 November 2020)
de Cheveigne A, Wong D D E, Di Liberto G M, Hjortkjaer J, Slaney M and Lalor E 2018 Decoding the auditory brain with canonical component analysis *NeuroImage* **172** 206–16
Di Liberto G M, Attaheri A, Cantisani G, Reilly R B, Ní Choisdealbha Aine, Rocha S, Brusini P and Goswami U 2023 Emergence of the cortical encoding of phonetic features in the first year of life *Nat. Commun.* **14** 7789
Di Liberto G M, Crosse M J and Lalor E C 2018a Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech *eNeuro* **5** ENEURO.0084–18.2018
Di Liberto G M and Lalor E C 2017 Indexing cortical entrainment to natural speech at the phonemic level: methodological considerations for applied research *Hear. Res.* **348** 70–77
Di Liberto G M, Nie J, Yeaton J, Khalighinejad B, Shamma S A and Mesgarani N 2021 Neural representation of linguistic feature hierarchy reflects second-language proficiency *NeuroImage* **227** 117586
Di Liberto G M, O'Sullivan J A and Lalor E C 2015 Low-frequency cortical entrainment to speech reflects phoneme-level processing *Curr. Biol.* **25** 2457–65
Di Liberto G M, Peter V, Kalashnikova M, Goswami U, Burnham D and Lalor E C 2018b Atypical cortical entrainment to speech in the right hemisphere underpins phonemic deficits in dyslexia *NeuroImage* **170** 70–79
Duchateau J, On Kong Y, Cleuren L, Latacz L, Roelens J, Samir A, Demuynck K, Ghesquière P, Verhelst W and Van hamme H 2009 Developing a reading tutor: design and evaluation of dedicated speech recognition and synthesis modules *Speech Commun.* **51** 985–94
Francart T, van Wieringen A and Wouters J 2008 APEX 3: a multi-purpose test platform for auditory psychophysical experiments *J. Neurosci. Methods* **172** 283–93
Hannun A *et al* 2014 Deep speech: scaling up end-to-end speech recognition (arXiv:1412.5567 [cs])
Holm S 1979 A simple sequentially rejective multiple test procedure *Scand. J. Stat.* **6** 65–70
Hrinchuk O, Popova M, and Ginsburg B 2019 Correction of automatic speech recognition with transformer sequence-to-sequence model (arXiv:1910.10697 [cs, eess])
Khalighinejad B 2017 Guilherme cruzatto da silva and nima mesgarani. Dynamic encoding of acoustic features in neural responses to continuous speech *J. Neurosci.* **37** 2176–85
Kovács A, Winkler I and Vicsi K 2017 EEG correlates of speech: examination of event related potentials elicited by phoneme classes *2017 8th IEEE Int. Conf. on Cognitive Infocommunications (CogInfoCom)* pp 000115–20
Lesenfants D, Vanthornhout J, Verschueren E, Decruy L and Francart T 2019a Predicting individual speech intelligibility from the cortical tracking of acoustic- and phonetic-level speech representations *Hear. Res.* **380** 1–9
Lesenfants D, Vanthornhout J, Verschueren E and Francart T 2019b Data-driven spatial filtering for improved measurement of cortical tracking of multiple representations of speech *J. Neural Eng.* **16** 066017
Luts H, Jansen S, Dreschler W and Wouters J 2014 Development and normative data for the Flemish/Dutch Matrix test (available at: https://lirias.kuleuven.be/retrieve/293640)
Mahapatra N C and Bhuyan P 2022 Multiclass classification of imagined speech vowels and words of electroencephalography signals using deep learning *Adv. Hum. Comput. Interact.* **2022** e1374880
Mesgarani N, Cheung C, Johnson K and Chang E F 2014 Phonetic feature encoding in human superior temporal gyrus *Science* **343** 1006–10
Monesi M J, Accou B, Francart T and Van hamme H 2021 Extracting different levels of speech information from EEG using an LSTM-based model *Proc. Interspeech 2021* pp 526–30
Monesi M J, Accou B, Montoya-Martinez J, Francart T and Van hamme H 2020 An LSTM based architecture to relate speech stimulus to EEG *ICASSP 2020—2020 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* pp 941–5
Oganian Y and Chang E F 2019 A speech envelope landmark for syllable encoding in human superior temporal gyrus *Sci. Adv.* **5** 11
Parhi M and Ahmed H T 2021 Classifying imaginary vowels from frontal lobe EEG via deep learning *2020 28th European Signal Processing Conf. (EUSIPCO)* pp 1195–9
Park D S, Chan W, Zhang Y, Chiu C-C, Zoph B, Cubuk E D and Quoc V L 2019 SpecAugment: a simple data augmentation method for automatic speech recognition *Interspeech 2019* pp 2613–7
Prinsloo K D and Lalor E C 2022 General auditory and speech-specific contributions to cortical envelope tracking revealed using auditory chimeras *J. Neurosci.* **42** 7782–98
Puffay C, Accou B, Bollens L, Jalilpour Monesi M, Vanthornhout J, Van hamme H and Francart T 2023 Relating EEG to continuous speech using deep neural networks: a review *J. Neural Eng.* **20** 041003
Puffay C, Van Canneyt J, Vanthornhout J, Van hamme H and Francart T 2022 Relating the fundamental frequency of speech with EEG using a dilated convolutional network *Proc. Interspeech 2022*
Sarmiento L C, Villamizar S, López O, Claros Collazos A, Sarmiento J and Bacca Rodríguez J 2021 Recognition of EEG signals from imagined vowels using deep learning methods *Sensors* **21** 6503
Somers B, Francart T and Bertrand A 2018 A generic EEG artifact removal algorithm based on the multi-channel Wiener filter *J. Neural Eng.* **15** 036007
van den Oord A, Yazhe Li and Vinyals O Representation Learning with Contrastive Predictive Coding 2019 (arXiv:1807.03748 [cs, stat])
Vanthornhout J, Decruy L, Wouters J, Simon J Z and Francart T 2018 Speech intelligibility predicted from neural entrainment of the speech envelope *J. Assoc. Res. Otolaryngol.* **19** 181–91
Verschueren E, Somers B and Francart T 2019 Neural envelope tracking as a measure of speech understanding in cochlear implant users *Hear. Res.* **373** 23–31
Wang R, Perreau-Guimaraes M, Carvalhaes C and Suppes P 2012 Using phase to recognize English phonemes and their distinctive features in the brain *Proc. Natl Acad. Sci. USA* **109** 20685–90