Proceedings of the
8th Graz Brain-Computer Interface Conference 2019 DOI: 10.3217/978-3-85125-682-6-16

# PRESENTATION SPEEDS FOR A N400-BASED BCI

K.V. Dijkstra, J.D.R Farquhar, P.W.M Desain

Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands

E-mail: k.dijkstra@donders.ru.nl

ABSTRACT:
The N400 is an ERP sensitive to the semantic content of a stimulus, in relation to the active mental context of a subject. By repeatedly presenting stimuli (i.e., probe words), we can thus infer information about this context through decoding of this relatedness response. This can form the basis of a, so called, semantic Brain Computer Interface, allowing us to directly identify the concept on a users mind without spelling out the letters. The usability of such a BCI depends on how much information can be extracted from a single presentation, but also on how fast stimuli can be presented. Here we report results from a pilot study of an experimental design that aims to determine the effect of the time between such probes on the amplitude of this N400. The preliminary results show that an N400 can still be detected across subjects, even with stimuli presented at nearly twice the rate (1.7x) used in one of our previous studies.

## INTRODUCTION

The N400 is an Event Related Potential that is more negative for stimuli that are not related to the person's current mental context [1]. It is thought to reflect the (attempt to) incorporate new semantic information (e.g., from a stimulus) into this existing context. From a Brain Computer Interface (BCI) viewpoint we may be able to exploit this activity to infer information about this hidden mental context. Specifically, we can use the presentation of semantic stimuli (e.g., words) to learn what this active semantic context is, by decoding the N400 from each stimulus presentation, and accumulating information across presentations. In other words, the user thinks of a specific concept (we refer to this as the target word), and the BCI presents multiple consecutive words (referred to as probes) that may or may not be related to this target, and combines the information inferred from the brain responses with knowledge of the relationships between the probe words and any potential target word, to ultimately infer the target word.

This requires a ground truth database that tells us which concepts are or are not related. Such a database can be built by consulting people on whether concepts are related, or by asking people what words come to mind for a given concept (e.g., [2] dutch). However, it is not feasible to ask about all possible combinations of concepts in this way. An alternative approach is to use methods from computational linguistics such as Latent Semantic Analysis (LSA) or the more recently popular word2vec [3]. These methods learn word representations, in the form of n-dimensional vectors, from large text corpora, by looking at which words occur together or which words occur in similar contexts, depending on the method. From these vector representations, the relatedness of two words can then be calculated by computing the *cosine similarity*, i.e., the cosine of the angle between the two vectors. The advantage of these vector-based methods is that the relatedness between any two concepts can be measured, while the methods in which humans are queried result in sparse databases. Importantly, the relatedness scores from such vector-space models have been shown to correlate with the strength of the N400 response, at least as well as the human-elicited association databases [4].

The amount of information that can be extracted about the initial target word depends on both the accuracy with which the relatedness can be decoded, and the amount of stimuli that can be presented in a given time frame. Previous studies suggested the N400 response has high variability and can be decoded with only limited accuracy (50-75% on a binary problem, [5, 6]), making the speed at which stimuli can be presented particularly relevant. In the second of these studies, in which we tested the robustness of this N400 over multiple consecutive probes after a specific target word, we used a Stimulus Onset Asynchrony (SOA) of 1350 ms, i.e., presenting a stimulus every 1.35 seconds.

Here we present an experimental design aimed to determine how much information is lost per stimulus for decreasing SOAs. We test an SOA of 1250 (SLOW), an SOA of 750 ms (MEDIUM) and a 250 ms SOA (FAST). Elicitation of the N400 does not *require* active semantic analysis on the part of the participant, with N400s also having been evoked in passive tasks. However, the N400 is generally stronger in tasks that involve active analysis [7], while reducing the time until the next stimulus will make this harder or impossible for participants. On the other hand a N400 BCI paradigm that does not require an active task, would be preferable from a user point of view. The question therefore is how, for a given time-window, the ability to present more stimuli trades off with a potential reduction in response amplitude and decoding accuracy of a single stimulus.

We report preliminary results of seven participants of a pilot of this experimental design. While the aim for a full scale study would be to determine what, if any, decrease in information transfer rates occurs for faster SOAs, we restrict ourselves here to validating the experimental design. Specifically, we use Grand Average ERPs across subjects to ascertain that the baseline condition (SLOW), elicits a distinguishable N400 response for related and unrelated probes, as some experimental parameters have changed with regard to the previous study (in particular the participants' task and the ground truth model). Further, we aim to determine for the faster SOAs whether either of those elicit a detectable N400. Lastly, we evaluate the performance of participants on the behavioural task to determine whether it is suitable as a behavioural check of the participant's attention to the stimuli.

## MATERIALS AND METHODS

In the experiment, participants were presented with a target word to remember, and then shown multiple consecutive probe words that had a varying degree of relatedness to this target word. Words were drawn from the 5000 most frequent words in English (Corpus of Contemporary American English [8]). The speed at which these probe words were presented depended on the condition:

- SLOW: 250 ms stimulus, 1000 ms fixation cross
- MEDIUM: 250 ms stimulus, 500 ms fixation cross
- FAST: 250 ms stimulus, no fixation cross

A behavioural task was added to the experiment to ensure subjects kept the target word in mind during the presentation of the sequence. After a certain number of probe words, the subject was prompted to decide whether the probe word was related or unrelated to the target (see Fig. 1). A total of a 150 of these trials were presented across 6 blocks of ~10 minutes each (50 in each condition).

*Participants:* Seven participants completed the pilot study, ranging in age between 20 and 30. Despite the availability of a large population of native Dutch speaker at our lab, English was used for the stimuli as there are more good quality vector-space models available. To minimise problems with word comprehension in non-native speakers, participants were recruited who categorized themselves as *"speaking and understanding English" "Well"* or *'Very Well'*. In addition, we used *LexTale* (Lextale.com [9]), as an objective measure of English vocabulary knowledge. Two subjects scored low on this task (50-60%); the remaining five all scored above 80%.

*Stimuli and Task:* To determine relatedness we obtained pre-trained word2vec vectors, trained on a Google News corpus (GoogleNews-vectors-negative300.bin, from https://code.google.com/archive/p/word2vec/). Target words were selected at random from the 5000 frequent words. For each target word, between 1 and 30 probe words were selected of varying relatedness with the target word. We refer to the combination of a target word together with all of its probes as a trial.

Trials contained a varying number of probes (between 1 and 30), so that participants could not anticipate when a behavioural prompt would appear. Each condition (FAST, MEDIUM and SLOW) was created to have an approximately equal number of probes per trial, but consequently did not have an equal duration; a trial with 30 probes would last ~10s, ~25 or ~40s, for FAST, MEDIUM and SLOW, respectively. Every 5 trials, the condition changed, with the subject being notified of the next presentation speed. The order in which conditions appeared was randomized across subjects.

Participants were given feedback on their relatedness decision in the behavioural task. To account for disagreement between the word2vec model relatedness judgements and their own, they received a short explanation of word-embedding methods, and were tasked to predict the model's judgements, rather than reporting their own. To this end, the continuous cosine similarity scores from the word2vec model were discretized into three bins: -1 to 0.15, 0.15 to 0.3, 0.3 to 1, labeled 'unrelated', 'maybe related' and 'related' respectively. If the subjects choice matched the model's label, they received one point. To keep the task straightforward, participants did not have a 'maybe' option, instead we customized the points for this category: they received a point if they predicted a 'maybe' as related (this suggests they are able to predict the model well), and half a point if they predicted it as unrelated (to penalize them less for not recognizing a relationship). Participants were given a 1250 ms window to respond, and a point was deducted if no choice was made.

*Updated design:* After three subjects we inspected ERPs and noted that in the FAST condition, the expected location of the N400, in time, coincides with the stimulus response to the subsequent stimulus. To reduce this effect we added a jitter to the stimulus duration of between 1 and 100 ms. We jittered the stimulus duration, rather than the fixation cross duration, as, in the base case, the fastest condition did not have a fixation cross. As, on average, the SOA was now increased by 50ms we reduced the fixation cross duration by this amount to compensate (applicable only to the MEDIUM and FAST conditions).

*Analysis:*

EEG was recorded with 32 sintered Ag/AgCl active electrodes (BioSemi ActiveTwo, Biosemi, Amsterdam, The Netherlands), at a sampling rate of 256 Hz. Two additional electrodes were placed on the mastoids, and four more electrodes were used to measure horizontal and vertical EOG.

For the analysis the recorded data was loaded and high-pass filtered at 0.1 Hz (4th order Butterworth filter). Data was then sliced into epochs with respect to each probe onset. These epochs were re-referenced to the two mastoid electrodes. To ensure any EEG signals were not contaminated with eye muscle activity, we regressed the signals from the EOG out of the EEG channels [10]. The data was subsequently low-pass filtered at 20Hz. To remove

DOI: 10.3217/978-3-85125-682-6-16

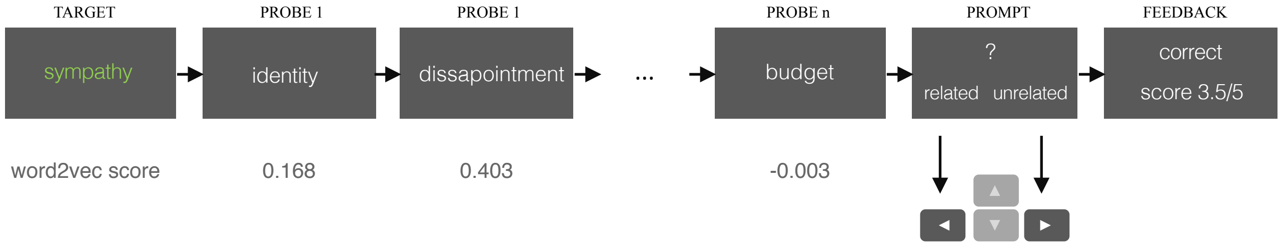

Figure 1: Experimental trials. A target word is presented, followed by several probe words. At the end of a trial, the participant is prompted to determine the relatedness of the most recent presented probe, in relation to the target word. Word2vec cosine similarity scores of the target and the respective probes are shown for this example trial.

any poorly-connected 'bad' channels, EEG channels that had a variance 3.5 standard deviations from the channel mean variance, were rejected, and replaced by an interpolated channel, using a spherical spline interpolation [11]. Epochs with abnormal activity were identified with the same 3.5 standard deviations of variance measure and excluded from further analysis.

## RESULTS

*Behavioural:* Participants were given a behavioural task to allow us to check that they were keeping the target word in mind, and attended the subsequent probe words. We report their scores as a fraction of the maximum obtainable score. Performance on this task can give us an estimate of their attention to the sequences, but is also dependent on their baseline ability to predict the word2vec model relatedness judgements. For this reason we included a 'wordpair' baseline task in which participants received a target with only a single subsequent probe. The behavioural scores for the three conditions and this baseline task can be found in Fig. 2.

The figure shows an increase in (median) score for faster presentation speeds, with participants performing similarly or higher on the fast sequences compared to the baseline task, but achieving lower scores on the MEDIUM and SLOW condition. Keep in mind that across the three conditions sequences were equally long with respect to the number of probes, and not the duration of the trial in seconds.

*EEG:*

To evaluate the suitability of the experimental design we look at grand average ERPs for each of the three conditions (FAST, MEDIUM and SLOW), contrasting responses to related and unrelated probes. For this purpose we select only epochs with probes that were highly unrelated or highly related according to the model: in the range $[-1, 0.1 >, < 0.3, 1]$ respectively. The ranges were chosen so that an approximately equal number of probes were assigned to the related as to the unrelated averages (200-250 probes per class). A large portion of stimuli fall outside this range, but have been included in the experiment to more closely resemble the probe distribution in a semantic BCI application, where probes could not be guaranteed to fall on only the 'extreme' sides of the relatedness spectrum, since the true target concept is unknown.

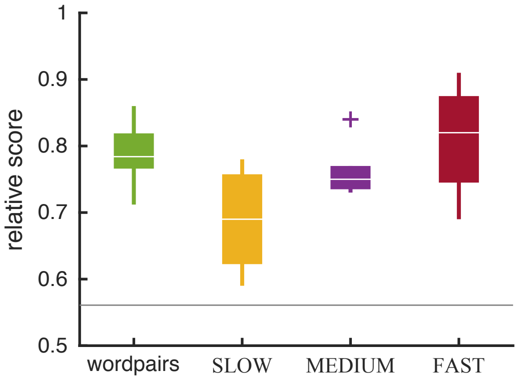

Figure 2: Participant accuracy on the behavioural prediction task. Score is relative to the maximum obtainable score. Scores are reported for each of the three speed conditions, plus a 'wordpair' condition that served as a baseline task. The grey line represents the mean accuracy achieved for pressing buttons randomly.

To test if there was a significant difference between related and unrelated probes for each condition, we used a non-parametric cluster-based permutation test (in Fieldtrip, [12]). The test was performed on all channels, in the 300-800 ms time range, using a bonferroni corrected one-sided alpha-level over conditions ($\alpha = 0.05/3 = 0.0167$). A significant effect was found for both the SLOW and MEDIUM condition ($p = 0.0010$ and $p = 0.0150$, respectively. The Grand Average ERPs across subjects can be found in Fig. 3. Fig. 3A and Fig. 3B show the average ERPs for the SLOW and MEDIUM condition. The grey areas mark the timepoints in the respective significant clusters identified by the permutation test. The average of unrelated probes is more negative around the 400ms range, as expected for the N400. The ERPs for these two conditions look largely similar in shape, though at the end of the 1 second window, for the MEDIUM condition, the (visually) evoked potential to the next stimulus appears to be visible (750ms SOA).

For the FAST condition (Fig. 3C, no significant cluster was identified ($\alpha = 0.0167$, $p = 0.0220$). In this condition four stimulus responses can be observed in the 1 second window, with the second one corresponding to the response to the stimulus triggering the ERP. The stimulus response to these epochs looks similar to those in the other conditions, but the stimulus response to the next stimulus co-occurs with the region in which the N400 was detected in the slower conditions. We introduced the jitter in stimulus duration, of 100ms, noted above, after S03 to reduce this overlap, but have averaged across all participants here (S01-S07).

The topographies reflecting the difference between related and unrelated ERPs, for the three conditions, are shown in Fig. 3D. The topographies represent the difference waves (unrelated − related), for (non-overlapping) periods of 200 ms, starting from 0 s. A negativity across centro-parietal sites, as expected for a N400, is visible in the 400-600ms slice, for all three conditions, though less pronounced for the FAST condition.

To visually investigate the degree of individual variability in ERP responses, per-subject and condition ERPs plots are shown in Fig. 4. Multiple channels in the centro-parietal region are shown. Note that standard deviations are large and not depicted here. Clearly the subject-to-subject variability in the responses is large – something masked by the grand-average responses. S01 shows a consistent negativity around 400ms for all three conditions. Other subjects also show a negativity in the expected timerange, though not necessarily across all conditions. S02, on the other end, shows very little of response that could be interpreted as an N400 at all. For the FAST condition, we would have expected to see a difference before and after introducing the stimulus duration jitter, but any effect is difficult to determine from this data. S06 and S07, for instance, have small stimulus responses even for the one stimulus the average was time-locked to.

## DISCUSSION

The goal of this preliminary study was to determine (1), whether the behavioural task was suitable for measuring the degree to which the subject was able to maintain attention during the repeated presentations of probes. (2), whether at the least the baseline condition (SLOW), elicited an N400, as expected. And (3), to give us a glimpse of the results we may expect in the full study with regard to the other two SOAs.

Behaviourally, we see that participants achieved scores well above chance level (though we did not test significance here), and that their performance decreases, rather than increases for slower presentation speeds. This is not what we expected, and suggests that the task in its current state does not benefit from the extra evaluation time of the probe prior to the behavioural prompt appearing. This could mean that conscious evaluation in that window is not important for the task, *or* that participants are not using this time for evaluating the relatedness to the original target (some comments made by participants suggest it is the latter). Additionally, all participants reported finding it harder to sustain attention during the SLOW condition, and some noted that it made them sleepy. This suggests that the decrease in performance is an effect of reduced

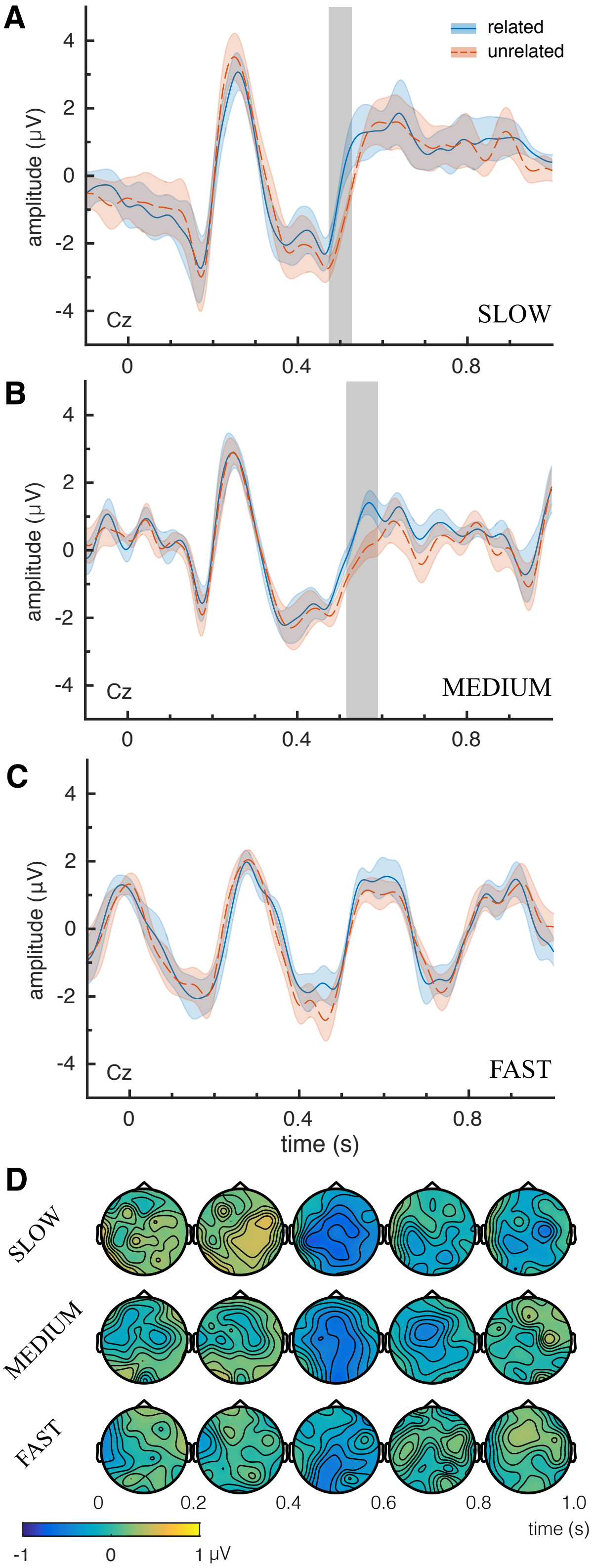

Figure 3: Grand Average ERPs across 7 subjects, for the central midline electrode (Cz), for **(A)**, SLOW, **(B)** MEDIUM and **(C)** FAST condition. Gray boxes denote significant clusters identified by a cluster permutation test. **(D)** topographies of the difference waves for each condition (unrelated − related), from 0-200,200-400,400-600 and 800-1000ms respectively.

DOI: 10.3217/978-3-85125-682-6-16

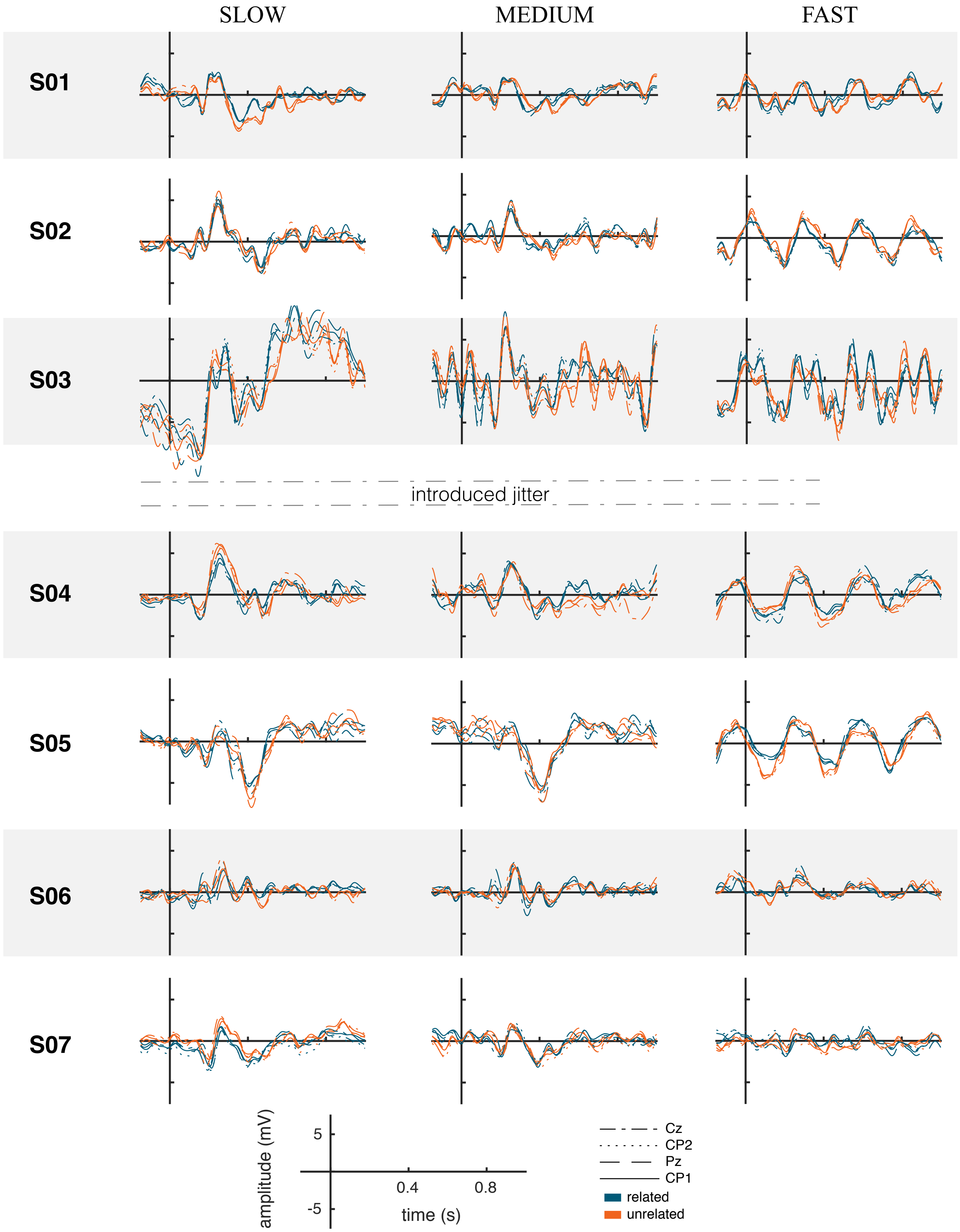

Figure 4: ERPs per individual subject (S01-S07), for the SLOW, MEDIUM and FAST condition, plotted for four centro-parietal electrodes (Cz, CP1, CP2, Pz). The introduction of stimulus duration jittering has been marked in the plot.

attention, however, this may not be intrinsically due to the SOA of this condition, but simply because the total accumulated time between the original target and the behavioural prompt was larger.

Our grand average ERP results show that a difference between related and unrelated probes can be identified from the SLOW condition, in the time range associated with the N400. This confirms that experimental design is able to elicit N400 responses, despite the use of non-native speakers as participants, and the use of the word2vec similarity scores as a relatedness measure. However, the amplitude difference looks small, and the individual ERPs also show only small amplitude differences in the N400 range, if at all. Anecdotally, this ERP difference looks smaller than in our previous study [6](preprint), but this may also be due to the fact that in the other study the task explicitly encouraged subjects to evaluate each probe before a behavioural prompt appeared. Here, we removed this aspect, as we assumed that the shorter SOAs would not give a participant an opportunity to do this, but this potential decrease in N400 amplitude may thus reflect the difference between an active and passive task (as established in other research [7]).

With regard to the other two conditions, we found a significant difference for the MEDIUM condition. This is an encouraging result, though due to the small *N* we do not yet have the statistical power to compare the size of this response to that of the SLOW condition. We find no related/unrelated difference for the FAST condition, in this pilot data, but statistical power is also a limiting factor here.

## CONCLUSION

Overall we can conclude that the experimental design is suitable to answer our question. Furthermore, the ability to present stimuli at the speed of the MEDIUM condition, rather than at the SLOW speed, would increase the rate of stimulation of the potential BCI by 1.7x. The fact that an N400 could still be detected reliably (across subjects), at this speed, is thus an encouraging preliminary result.

A full scale study, together with more sophisticated analysis of single subject data, e.g, regression or classification analysis and determining the information transfer over time, will be required to determine, in detail, the tradeoff between accuracy and speed of probe presentation (in terms of SOA). This in turn will determine the speed with which a BCI paradigm that exploits the N400 response can infer the concept on a user's mind, and hence the suitability of this approach for potential applications.

## REFERENCES

[1] 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, no. 1, pp. 621–647, 2011.

[2] S. D. Deyne and G. Storms, "Word associations: Norms for 1,424 Dutch words in a continuous task", en, *Behavior Research Methods*, vol. 40, no. 1, pp. 198–205, Feb. 2008. (visited on 10/09/2017).

[3] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality", in *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2013, pp. 3111–3119.

[4] C. Van Petten, "Examining the N400 semantic context effect item-by-item: Relationship to corpus-based measures of word co-occurrence", *International Journal of Psychophysiology*, vol. 94, no. 3, pp. 407–419, Dec. 2014.

[5] J. Geuze, M. A. J. van Gerven, J. Farquhar, and P. Desain, "Detecting Semantic Priming at the Single-Trial Level", *PLoS ONE*, vol. 8, no. 4, e60377, Apr. 2013. (visited on 10/06/2014).

[6] K. Dijkstra, J. Farquhar, and P. Desain, "Semantic Probing: Feasibility of using sequential probes to decode what is on a user's mind", [arXiv preprint doi:10.1101/496844, December 2018], Dec. 2018.

[7] D. Cruse, S. Beukema, S. Chennu, J. G. Malins, A. M. Owen, and K. McRae, "The reliability of the N400 in single subjects: Implications for patients with disorders of consciousness", *NeuroImage: Clinical*, vol. 4, pp. 788–799, Jan. 2014.

[8] M. Davies, "The Corpus of Contemporary American English (COCA): 560 million words", 1990-present, 2008. [Online]. Available: `Availableonlineathttps://corpus.byu.edu/coca/..`

[9] K. Lemhöfer and M. Broersma, "Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English", en, *Behavior Research Methods*, vol. 44, no. 2, pp. 325–343, Jun. 2012.

[10] G. Gratton, "Dealing with artifacts: The EOG contamination of the event-related brain potential", en, *Behavior Research Methods, Instruments, & Computers*, vol. 30, no. 1, pp. 44–53, Mar. 1998.

[11] F. Perrin, J. Pernier, O. Bertrand, and J. F. Echallier, "Spherical splines for scalp potential and current density mapping", *Electroencephalography and Clinical Neurophysiology*, vol. 72, no. 2, pp. 184–187, Feb. 1989.

[12] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data", *Journal of Neuroscience Methods*, vol. 164, no. 1, pp. 177–190, Aug. 2007.