Confidence Prediction from EEG Recordings in a Multisensory Environment

Jacobo Fernandez-Vargas
Brain Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom
j.fernandez-vargas@essex.ac.uk

Davide Valeriani
Department of Otolaryngology | Head and Neck Surgery, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA
davide_valeriani@meei.harvard.edu

Caterina Cinel
Brain Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom
ccinel@essex.ac.uk

Nitin Sadras, Parima Ahmadipour, Maryam M. Shanechi
Department of Electrical and Computer Engineering, University of Southern California, Los Angeles, CA, USA
nsadras@usc.edu, ahmadipo@usc.edu, shanechi@usc.edu

Luca Citi, Riccardo Poli
Brain Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, United Kingdom
lciti@essex.ac.uk, rpoli@essex.ac.uk

ABSTRACT

This paper investigates the possibility of decoding decision confidence from electroencephalographic (EEG) brain activity of human subjects during a multisensory decision-making task. In recent research we have shown that decision confidence correlates could be extracted from EEG recordings during visual or auditory tasks. Here we extend these initial findings by (a) predicting the confidence in the decision from EEG recordings alone, and (b) investigating the impact of multisensory cues on decision-making behavioral data. Our results obtained from 12 participants recorded at two different sites show that the decision confidence could be predicted from EEG recordings on a single-trial basis with a mean absolute error of 0.226. Moreover, the presence of a multisensory cue did not improve the performance of the participants, but rather distracted them from the main task. Overall, these results may inform the development of cognitive systems that could monitor and alert users when they are not confident about their decisions.

CCS Concepts

• **Applied computing→Bioinformatics**

Keywords

EEG; Confidence; Decision making

1. INTRODUCTION

In every-day life humans are continuously challenged with situations that require instantaneous decision making. In some occasions, the outcome of such decisions may have life-threatening consequences, for example for pilots, air traffic

controller or soldiers. It is therefore important to understand which mental states are associated to high performance in critical decision-making tasks.

Usually, mental states are estimated from behavioral measures. In decision-making, these measures include the decision confidence, which indicates how sure people are about their decisions [1], [2]. The better someone knows a task the more confident they will be, but also the more accurately they will estimate their confidence [3]. Still, the relationship between the confidence and the accuracy may vary depending on the subject, the task, or the conditions [4]. It has also been shown that the confidence correlates with reaction times (RT) [12]-[16] as well. A straightforward way to record decision confidence is asking directly to the participants after each decision. However, this approach is not viable in a real-world task as it would it would be time consuming. While it has already been shown that is possible to classify confidence in perceptual tasks [5]-[7], our proposed approach goes beyond binary confidence levels and focuses on continuous reporting featuring a more realistic task. The ability to predict the confidence in a decision using only the electroencephalographic (EEG), may be very useful in several applications. On one hand, predicting the confidence has a disadvantage compared to predicting the actual accuracy. While high levels of confidence in one's decisions are likely to indicate a high accuracy on that decision, very low levels of confidence indicate randomness in the decision (when talking about forced decision tasks, i.e. the answer cannot be "no answer") rather than being sure of being wrong. The predicted confidence is therefore more useful for high levels of confidence, while lower levels might not contain information about the task performance. On the other hand, being able to predict the confidence instead of the accuracy has the advantage that, in a forced decision with two-option tasks, there is a 50% probability of being correct. Thus, by predicting the confidence we are predicting the actual state of the participant. Another advantage of predicting the confidence, is that the confidence does not depend on the actual task, but only on the user's meta-cognition, while the accuracy is strongly tied to the task itself.

A second aim of this study was to investigate the effectiveness of cross-modal cues. It has been demonstrated that cross-modal cues

improve the accuracy [8]-[10], compared to unimodal cues, which, in turn, should alter other behavioral measurements. In this study, we analyze how different cross-modal cues altered accuracy, confidence and RTs. Moreover, we investigated whether the cross-modal cues affected the confidence prediction or if the prediction was independent of the cue.

2. MATERIALS AND METHODS

2.1 Participants

Twelve participants with normal or corrected-to-normal vision took part in the experiment (4 males, mean age 29.1 years, SD 11.4). Six participants were recruited at the University of Essex, UK, and the other six participants were recruited at the University of Southern California (USC), USA.

All the participants signed a written, informed consent before taking part in the experiment. The research is part of a project funded by the UK's MoD through DSTL which received MoD and University of Essex ethical approval in 2017.

2.2 System

Participants sat comfortably during the whole experiment at about 80 cm from an LCD screen. The EEG system (BioSemi ActiveTwo) was the same in both locations (Essex and USC). However, the experiments done at the University of Essex were performed with an EEG cap of 64 electrodes, while the experiments performed at USC used a cap with 256 electrodes. It is important to note that the 64 electrodes are not a subset of the 256. Thus, to make it possible to compare both datasets, the USC dataset was interpolated into the same positions of the 64-electrode cap. First, we created a mesh using the theoretical position of the USC 256 electrodes (we will call them origin electrodes). Then, we located inside that mesh the theoretical position of each one of the Essex 64 electrodes (we will call them destination electrodes). For each one of the latter electrodes there were four possibilities:

1) The destination electrode coincided with one of the origin electrodes, in which case the signal of the destination electrode was the same as the origin electrode.

2) The destination electrode was exactly between two neighbor origin electrodes. In this case, the signal of the destination electrode was interpolated from the two origin electrodes using the bilinear method.

3) The destination electrode was between three origin electrodes. In this case, the signal of the destination electrode was interpolated from the three origin electrodes using the bilinear method.

4) The destination electrode was outside of the origin electrodes' mesh. In this case the electrode was removed from both datasets.

There were four electrodes that were located outside of the mesh: AF7, P9, AF8, and P10. This means that all the preprocessing and analysis of the data was done on 60 electrodes. All the channels were re-referenced to the mean signal recorded from two electrodes placed on the earlobes.

2.3 Task

Participants were asked to undertake a perceptual decision-making task consisting of 16 blocks of 48 trials, for a total of 768 trials. The experiment started showing the participants two images representing a soldier with a hat and a soldier with a baseball cap (Figure 1). Each trial (Figure 1) started with a fixation cross,

shown for 1000 ms, immediately followed by and image of an empty corridor with doors on each side for 500 ms. During this time, one out of four possible cues was presented (more details below). After the cue, the image of a soldier was presented in the corridor for 250 ms, followed by the message "Have you seen a HELMET (LEFT) or a CAP (RIGHT)?", where participants had to decide whether the soldier was wearing a helmet or a cap. They reported their decision by pressing the left or the right mouse buttons, controlled with their preferred hand. Finally, participants were asked to report the confidence in their decisions between 0% and 100% by scrolling the mouse wheel, which varied confidence in 10% steps. Participants were instructed to provide fast and accurate responses.
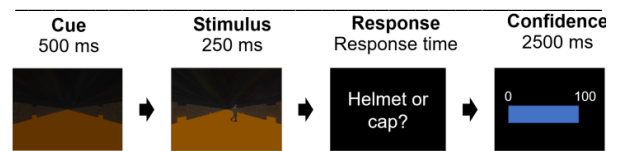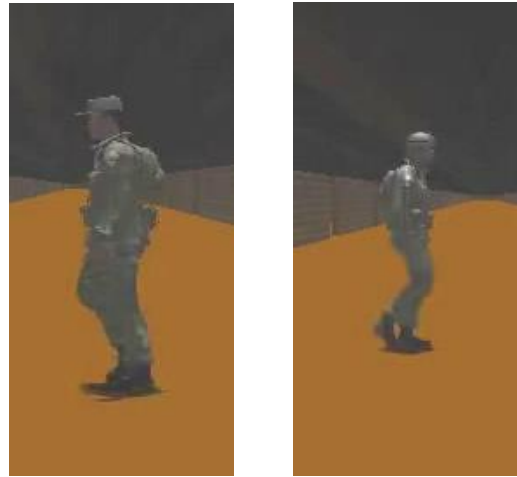
Figure 1 (Top) Examples of the two stimuli presented to the participants. The one in the right is the one considered target. (Bottom) Timeline of a single trial. Since the response times varies trial by trial, the duration of each trial was not fixed.

In each trial, the participant received one of the following cues:

- No cue: There was no cue before the stimulus appeared.
- Audio: 500 ms before the stimulus appeared, a voice saying either "right" or "left", indicated at which side of the screen the stimulus would appear.
- Silent Head (SH): 500 ms before the stimulus appeared, the figure of a head would appear and remained there until the stimulus disappeared.
- Talking Head (TH): 500 ms before the stimulus appeared, the figure of a head would appear at the same time as a voice. The voice was the same as in the Audio condition. Differently from the SH condition, the head in this condition moved mimicking the motions of someone saying "right" or "left".

For a given block, all the trials had the same cue. For each participant, there were four blocks for every kind of cue. However, the order was randomized for each of them. The type of cue was presented to the participants before starting each block. Finally, after each block, the mean accuracy for that block was reported to the participants, so they would get feedback of their performance.

At the beginning of the experiment, participants practiced the task by undertaking one block of 10 trials for each type of cue.

2.4 Signal Processing

The EEG preprocessing used in experiment is the same that we used in past experiments [4]. The original data was sampled at 2048 Hz, then band-pass filtered between 0.15 and 40 Hz using a FIR filter resulted of convolving a low-pass filter with a high-pass filter. Then, the signal was downsampled by a factor of 16, resulting in a 128 Hz signal. In addition to this, a correction for eye-blink and other ocular movements was performed using a subtraction algorithm based on correlations to the average differences between FP1 and F1 and Fp2 and F2. [11]

After the preprocessing, the data were split into two types of epochs:

Stimulus Locked: Each epoch started at the onset of the cue and lasted 1.5 s.

Response Locked: Each epoch started 1.25 s before the response and lasted 1.5 s after. This representation is useful because, in a real-world application, the exact moment of the stimulus may not be available, while the time of response is usually known.

For each method, the data from each channel were baseline-corrected to the mean voltage of the same channel from 25 ms before to 25 ms after the zero.

2.5 Methods

The main goal of the experiment was to see if it was possible to predict the confidence using only EEG data. The reason for this is that, given that the task remains the same over time, changes in the confidence would indicate possible drops in attention or other performance-related processes. Being able to predict such changes would help to predict changes in the performance.

2.5.1 Confidence correlation

To see the correlation between the confidence and the accuracy, we grouped the responses into three confidence levels: low [0-0.3], medium [0.4-0.6], and high [0.7, 1]. Then, we calculated the mean accuracy and 95% confidence interval for each of the confidence levels, to see if they overlapped or not.

2.5.2 Grand average

For each participant, we calculated the mean EEG activity grouping the trials by the confidence level. Then, the mean activity across participants was calculated. The goal was to see if there was any significant EEG signal difference for the different confidence levels. A difference in EEG activity could mean that it is possible to extrapolate the participants confidence on their own decision from EEG alone without needing to ask them after each action. We performed this process using response locked epochs.

We performed a second analysis in which the data were grouped according to the pre-stimulus cue. A difference in the EEG signal would indicate different mental processing for the cues. This analysis was done using stimulus locked epochs because the goal was, precisely, to compare different stimulus.

2.5.3 Prediction

As mentioned before, the final goal was to predict the confidence using only the EEG. To be able to test our method, we compared it to a baseline method that would predict always the same

confidence. More specifically, the predicted confidence by the baseline method was the mean of the confidence during the training trials. On the other extreme, we compared our EEG-based method to another predictor that used as input the RT of each trial. As it has been mentioned before, the RT and the confidence are highly correlated. However, it is important to note that in a real-world application the RT of an action may or may not be available.

In the methods based only on EEG data, we extracted the features from response locked epochs. As features we extracted the Auto Regressive (AR) coefficients that model the epoch in an AR(1) process. A general AR(p) process is described as:

$$y(k) = w_0 + \sum_{n=1}^{p} w_n y(k - n)$$

where p is the order of the process, and w_n are the coefficients that fit the model.

The three methods used as predictor a linear regression. We performed the prediction and validation process using five different data sets. The first four contained only one of the conditions (No cue, Audio, TH, and SH), while the fifth contained all of them together. The goal of this approach was to see if any of the conditions presented and advantage/disadvantage to predict the confidence. Finally, to compare the accuracy of each method, we used the Mean Absolute Error (MAE) between the prediction and the prediction of each one of the predictors.

3. RESULTS

3.1 Behavioral

Table 1 shows the results for the RT, accuracy, and confidence for the four different experimental conditions.

Table 1. Behavioral results

Condition	RT	Accuracy	Confidence
No Cue	.487	.889	.743
Audio	.522	.888	.741
SH	.546	.828	.647
TH	.522	.846	.704
Mean	**.519**	**.863**	**.709**

Table 2. Pairwise comparison p-values

Conditions	RT	Accuracy	Confidence
No Cue vs SH	**<.001**	**<.001**	**<.001**
No Cue vs Audio	**.025**	.959	.877
No Cue vs TH	.056	.065	**.011**
TH vs Audio	.91	**.015**	**.004**
TH vs SH	.46	.513	**.011**
Audio vs SH	.243	**.003**	**<.001**

The ANOVA analysis showed significant main effects for the RTs (F=7.58, p < 001), accuracy (F=11.66, <.001) and confidence (F=29.90, <.001). Holm-Bonferroni corrected p-values for the t-test comparing the conditions two-by-two are shown in
Table 2.
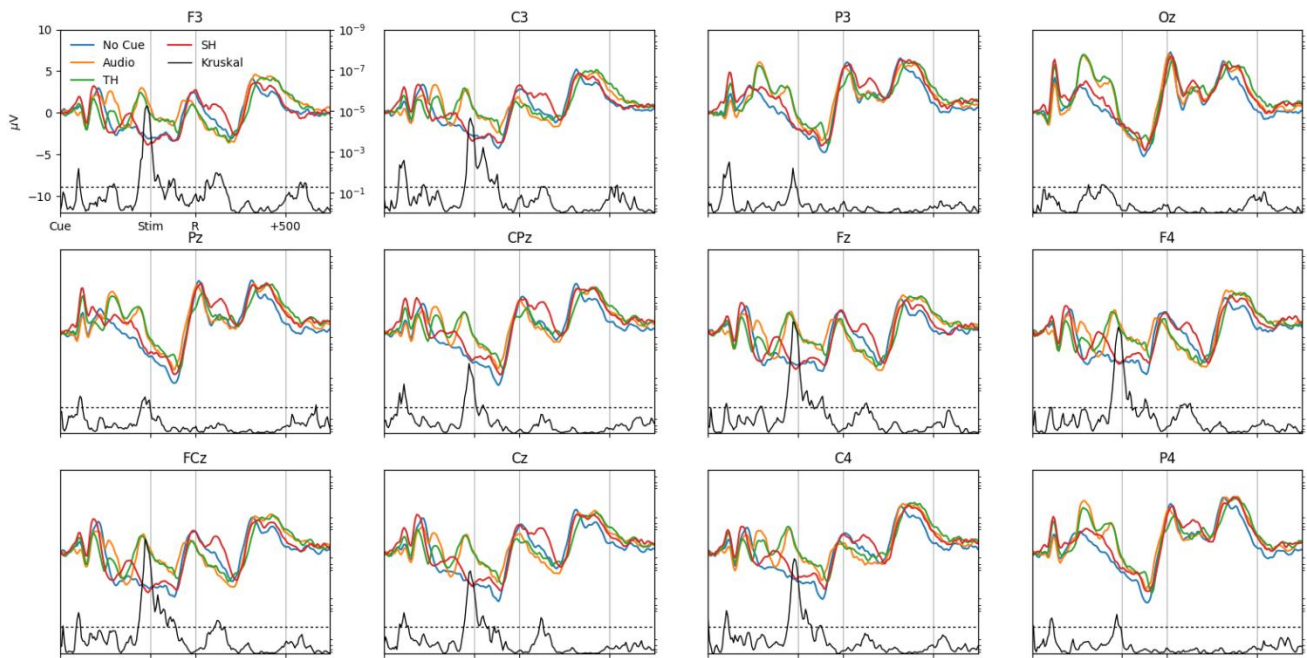
Figure 2. Stimulus locked grand averages comparing the four conditions for 12 different channels. Each colour line corresponds to one condition. The black line (that refers to the right y-axis), indicates the uncorrected p-value of the Kruskal-Wallis test comparing each individual point and channel. The horizontal dotted black line represents p=0.05 significant difference.

Those p-values considered significant (p-values < 0.05) are shown in boldface. The 'No Cue' condition showed the fastest RT, with the higher accuracy and confidence. This difference was statistically significant only compared to the 'SH' condition (the slowest one, with the lowest accuracy and confidence). The 'Audio' condition had a significant higher accuracy than the 'TH' and the 'SH' conditions, however, this difference was not significant in the RT. The variable that showed the greatest number of statistical differences was confidence, where all the comparison were statistically different except for the No Cue vs Audio.

3.2 Grand Averages

In Figure 2, the grand averages for stimulus locked analysis are presented for 12 channels, which cover from the frontal to the occipital area in the central line, as well as both laterals from the frontal to the parietal areas. In addition, the p-values of the Kruskal-Wallis analyses comparing the four conditions are shown. The response locked results were not presented in this paper, as there were no consistent significant differences for any of the electrodes or times.

Figure 3 (top) shows the differences in the EEG activity before the decision for the three confidence levels. The low p-values (in some cases <.0001) of the Kruskal-Wallis test, and both the temporal and spatial consistency of these values (the differences are not localized in a very specific point of time or space), confirm that the differences in the figures were not due to chance.

3.3 Confidence Prediction

The results presented in Figure 4 (Top) show a strong relationship between confidence and accuracy. This shows the ability of the participants to correctly self-evaluate their decisions. It is interesting to note how the interpretation and use of the level '0' for the confidence changes across participants (no indication was given to the participants about the meaning of "0" confidence).

For 7 out 12 of them, a confidence of '0' had a mean accuracy of around 50%, suggesting that for them the lowest confidence meant random choice, or that they did not know the answer (however, they were forced to provide an answer). On the other hand, for the remaining five participants, a confidence of '0' had a mean accuracy of 0%, suggesting that these five participants used the lowest level to report that they knew that they had made a mistake. This difference makes that the 95% confidence interval for the confidence level '0' so big compared to the other confidence levels.

MAEs of confidence prediction using the three different methods and the five sets of data can be seen in Figure 4 (Bottom). First, we performed a Kruskal-Wallis analysis comparing the four conditions (No Cue, Audio, SH, and TH) to see if there was a difference predicting the confidence for each one of the conditions. The result of this analysis resulted in a p-value of >.5. Next, we performed the Kruskal-Wallis analysis comparing the three methods using the complete set of data. This resulted in a p-value < 0.001. The MAEs of this data set were .229, .207, and .245 for AR, RT, and Baseline methods respectively. The paired Wilcoxon test analysis with Bonferroni correction showed that all the methods were significantly different between them (p-value <0.001 in every case).

4. DISCUSSION
4.1 Behavioral

In terms of accuracy, the behavioral results showed that the No Cue condition was the best one closely followed by the Audio condition, while the TH was worst than both conditions (even if only significantly different than the No cue condition), but still better than the SH condition. This suggest that the presence of the head is a distractor, while the audio information provides some useful information that can balance the negative effect of the head to some extent. Regarding the confidence, we can see that the results followed a similar pattern, with the No cue being the

condition with the highest confidence closely followed by the Audio condition, the SH condition being the worst condition with only .647 confidence, and the TH somewhere in between. Differently from the accuracy, the p-value of the paired t-tests showed statistical difference in all the comparison except for the No cue vs Audio test. Finally, in terms of RTs, the data showed something slightly different. The No Cue condition was still better than any other condition (with no significant difference with the TH), however instead of the Audio's RT being similar to the No Cue condition, it was similar to the TH condition, again the SH condition is the worse of them with the slowest responses. This would suggest, that the time required to process the audio signal is interfering with the processing of the stimulus. Then, comparing the grand averages in Figure 2, we can see that in those parts that the Kruskal-Wallis test indicates bigger differences, the No Cue and the SH behave in a similar way and different to the TH and the Audio conditions. This behavior is probably showing the semantic processing of the audio cue, which requires some time.

This would explain why the RTs of these two conditions were similar, even if the accuracy and the confidence are significantly different. This led us to consider that a longer interval between cue and stimulus presented should therefore increase both the accuracy and the RT of the 'Audio' and 'TH' conditions.
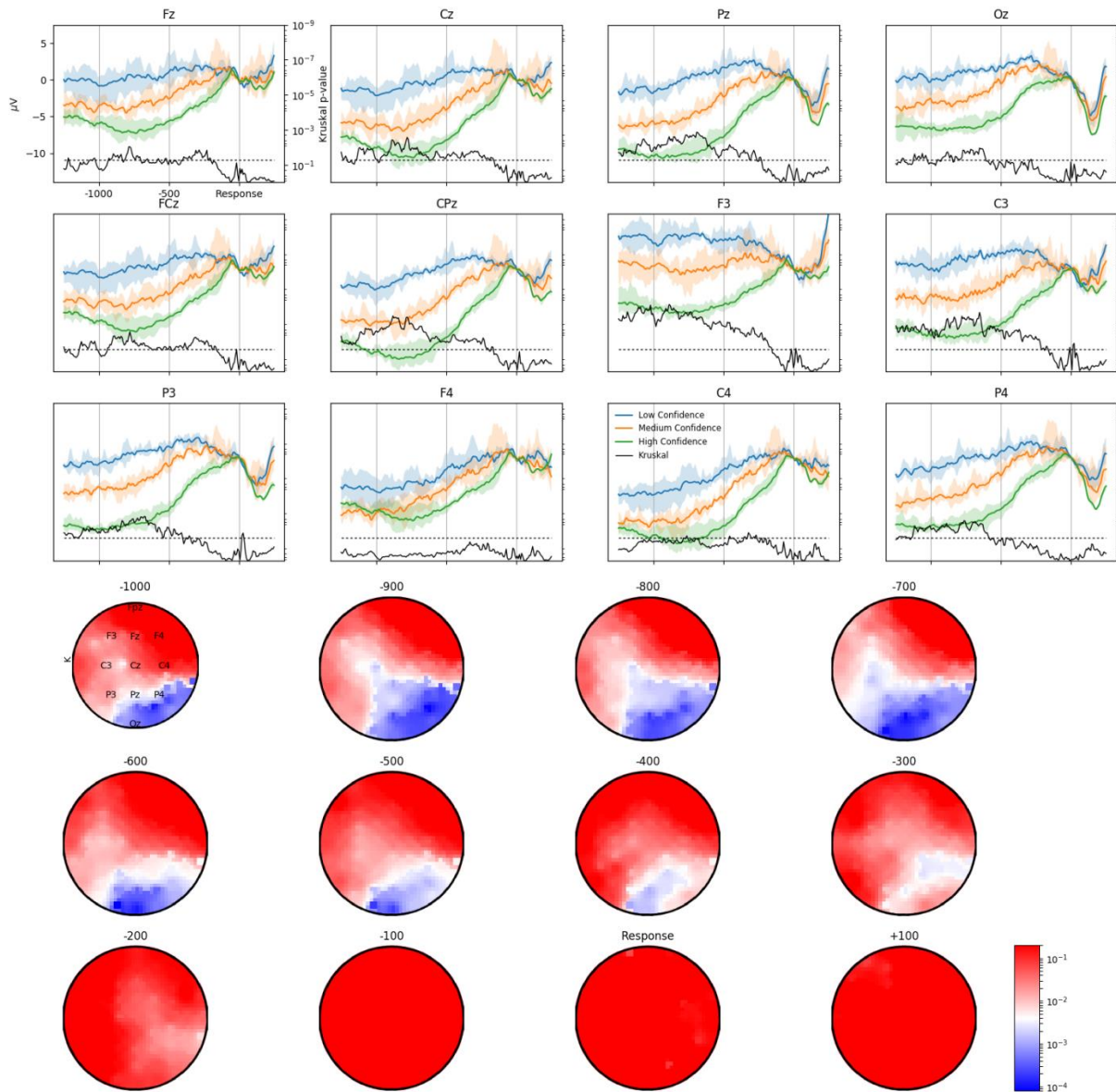
Figure 3. (Top) response-locked grand averages of EEG for Low, Medium and High levels of confidence. The colored lines correspond to the mean activity for all participants and epochs for the three levels of confidence defined. The shaded areas correspond to the range of the mean when the EEG activity is grouped by confidence levels and condition, i.e. the mean of each one of the four conditions for a given level would be inside the shaded area. The black line refers to the right y-axis, and indicates the p-values resulting from the Kruskal-Wallis test for each individual point and channel. The dotted black line represents the significant-difference level at p-value 0.05. Each panel represents one channel. (Bottom) p-value scalp maps for response-locked epochs. Each scalp map corresponds to one-time sample. Blue colors correspond to lower p-values. Note that the scale for the p-values is logarithmic.
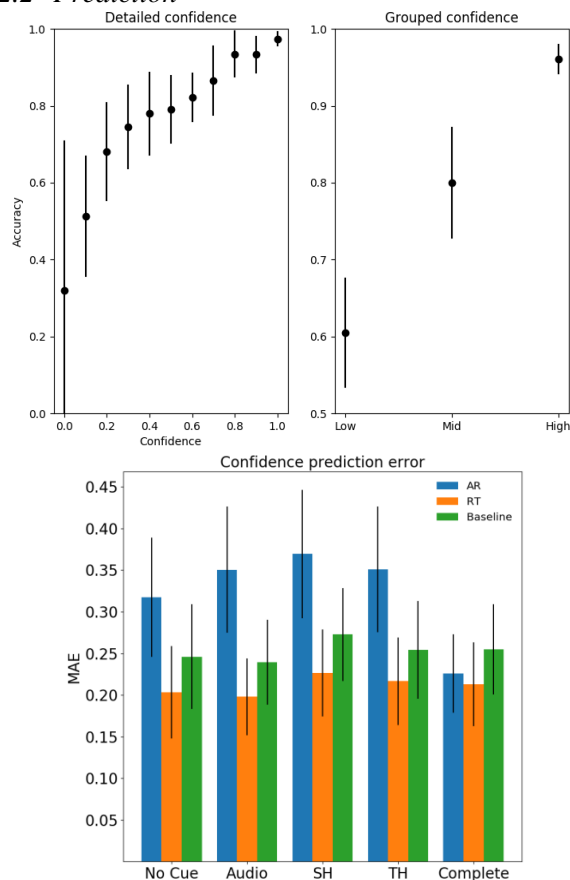
4.2 Confidence

4.2.1 Grand averages

The grand averages of the three confident levels in Figure 3, show that there is a clear and statistically significant difference in the EEG activity before making the decision. This supports the idea that the confidence is integrated in an online way [17]. These differences are consistent across time and space. The biggest differences are between 800 and 600 ms before the decision, with a stronger effect on the left and parietal areas of the brain. The right hemisphere also shows some of these differences, especially in the parietal/occipital area and, to some extent, in the central area. The frontal area does not show any significant difference. It is interesting to note as well, that the EEG activity shows a coherent shift on the grand average activity correlated with the confidence level. We can see that the low confidence has the highest EEG activity followed by the medium confidence and the high confidence level show the lowest activity the three levels.

In addition to this, the fact that there is a difference between the means when grouped only by confidence levels but not when grouped by condition, suggest that the effect of the confidence level is bigger than the differences due to the four different cues, making it possible to analyze the confidence as a whole without taking into consideration the four different cues.

4.2.2 Prediction

Figure 4 (Top) Mean accuracy and 95% confidence interval for each of the confidence level as for the three defined confidence ranges. (Bottom) MAE and 95% confidence interval for the three methods proposed to calculate the confidence, for each one of the five data-sets used for the training and validation.

The upper panel in Figure 4 shows that there are big differences ($p\text{-}value < 0.0001$) in the accuracy not only between the different confidence ranges, but also in the 11 confidence levels. This results expand previous studies where it was shown the difference in terms of accuracy of binary confidence levels [5]–[7], by showing that this difference happens also when using graded confidence levels.

The lower panel in Figure 4 (in addition to the statistical analysis performed), shows that the error of predicting the confidence when dividing the data by the cue condition is independent from the condition. Thus, making it possible to use all the data samples together to predict the confidence. Interestingly, merging the four conditions results in a significant decrease of the error when using the AR model (this doesn't occur with the other approaches). The most likely reason lies within the number of parameters that need to be fitted by the model. Each one of the conditions has only a fourth of the data samples to train the linear regression. In the case of the RT, only two parameters need fitting (the RT and the bias). In the case of the AR model, the number of parameters is substantially bigger as it is two times the number of electrodes (120 in total) plus the bias. The lower accuracy for separated conditions, suggests that there were not enough samples to fit appropriately the model to each of the conditions.

5. CONCLUSIONS

In this paper, we showed that there was a strong relationship between accuracy, RTs, and confidence in a multisensory decision-making task. Moreover, providing unisensory or multisensory cues does not increase the accuracy nor decrease the confidence of the participants, but it may increase the RTs. This suggests that multisensory cues increase the cognitive load and, in our experiment, they did not provide any help to the participant to increase the task performance. Future studies should investigate whether longer cues or cue-to-stimulus intervals could change these results.

We also validated our hypothesis that EEG signals carry enough information for machine learning algorithms to estimate the decision confidence of the participant. However, RTs provide better estimates of the decision confidence than the EEG alone. Future studies should investigate whether combining these two measures could further increase the accuracy of the confidence estimates.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] P. Grimaldi, H. Lau, and M. A. Basso, "There are things that we know that we know, and there are things that we do not know we do not know: Confidence in decision-making," *Neuroscience and Biobehavioral Reviews*, vol. 55. Elsevier Ltd, pp. 88–97, 01-Aug-2015.

[2] L. Aitchison, D. Bang, B. Bahrami, and P. E. Latham, "Doubly Bayesian Analysis of Confidence in Perceptual Decision-Making," *PLoS Comput. Biol.*, vol. 11, no. 10, 2015.

[3] N. Yeung and C. Summerfield, "Metacognition in human decision-making: confidence and error monitoring," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 367, no. 1594, pp. 1310–1321, May 2012.

[4] D. Valeriani, C. Cinel, and R. Poli, "Group Augmentation in Realistic Visual-Search Decisions via a Hybrid Brain-Computer Interface," *Sci. Rep.*, vol. 7, no. 1, p. 7772, Dec. 2017.

[5] S. Gherman and M. G. Philiastides, "Neural representations of confidence emerge from the process of decision formation during perceptual choices," *Neuroimage*, vol. 106, pp. 134–143, Feb. 2015.

[6] M. Graziano, L. C. Parra, and M. Sigman, "Neural correlates of perceived confidence in a partial report paradigm," *J. Cogn. Neurosci.*, vol. 27, no. 6, pp. 1090–1103, Jun. 2015.

[7] M. A. Pisauro, E. Fouragnan, C. Retzler, and M. G. Philiastides, "Neural correlates of evidence accumulation during value-based decisions revealed via simultaneous EEG-fMRI," *Nat. Commun.*, vol. 8, p. 15808, Jun. 2017.

[8] D. R. Begault, "Head-up auditory displays for traffic collision avoidance system advisories: A preliminary investigation," *Hum. Factors*, vol. 35, no. 4, pp. 707–717, 1993.

[9] J. J. McDonald, W. A. Teder-Sälejärvi, and S. A. Hillyard, "Involuntary orienting to sound improves visual perception.," *Nature*, vol. 407, no. 6806, pp. 906–8, Oct. 2000.

[10] C. Spence and J. Driver, *Crossmodal Space and Crossmodal Attention*. Oxford University Press, 2012.

[11] B. W. Jervis, E. C. Ifeachor, and E. M. Allen, "The removal of ocular artefacts from the electroencephalogram: a review," *Med. Biol. Eng. Comput.*, vol. 26, no. 1, pp. 2–12, Jan. 1988.

[12] R. D. (Robert D. Luce, *Response times : their role in inferring elementary mental organization*. Oxford University Press, 1986.

[13] R. Ratcliff and J. J. Starns, "Modeling confidence judgments, response times, and multiple choices in decision making: Recognition memory and motion discrimination.," *Psychol. Rev.*, vol. 120, no. 3, pp. 697–719, 2013.

[14] J. V. Baranski and W. M. Petrusic, "The calibration and resolution of confidence in perceptual judgments," *Percept. Psychophys.*, vol. 55, no. 4, pp. 412–428, Jul. 1994.

[15] R. Kiani, L. Corthell, and M. N. Shadlen, "Choice Certainty Is Informed by Both Evidence and Decision Time," *Neuron*, vol. 84, no. 6, pp. 1329–1342, Dec. 2014.

[16] R. Kiani and M. N. Shadlen, "Representation of confidence associated with a decision by neurons in the parietal cortex.," *Science*, vol. 324, no. 5928, pp. 759–64, May 2009.

[17] D. Dotan, F. Meyniel, and S. Dehaene, "On-line confidence monitoring during decision making," *Cognition*, vol. 171, pp. 112–121, Feb. 2018.