# EEG analysis for implicit tagging of video data

Sander Koelstra
MultiMedia and Vision Group
Queen Mary, University of London, UK
Sander.Koelstra@elec.qmul.ac.uk

Christian Mühl
Human Media Interaction
University of Twente, NL
C.Muehl@utwente.nl

Ioannis Patras
MultiMedia and Vision Group
Queen Mary, University of London, UK
I.Patras@elec.qmul.ac.uk

## Abstract

*In this work, we aim to find neuro-physiological indicators to validate tags attached to video content. Subjects are shown a video and a tag and we aim to determine whether the shown tag was congruent with the presented video by detecting the occurrence of an N400 event-related potential. Tag validation could be used in conjunction with a vision-based recognition system as a feedback mechanism to improve the classification accuracy for multimedia indexing and retrieval. An advantage of using the EEG modality for tag validation is that it is a way of performing implicit tagging. This means it can be performed while the user is passively watching the video. Independent Component Analysis and repeated measures ANOVA are used for analysis. Our experimental results show a clear occurrence of the N400 and a significant difference in N400 activation between matching and non-matching tags.*

## 1. Introduction

Given the enormous amount of unannotated multimedia data available nowadays, the need for automatic categorisation and labelling of video material to enable efficient indexing and retrieval is evident. So far, the predominant method used for tagging video data is by manual annotation. This is a slow, labour intensive process that cannot keep up with the amount of newly generated multimedia data. Lately, research has focused on finding ways to automate the annotation of this data. The use of EEG in this process is interesting mainly because it offers the possibility of passive, implicit tagging. This means that tags can be generated by analysing the EEG data as subjects consume multimedia data, without active involvement or conscious effort on their part. While at the moment the recording of EEG measurements is still a quite cumbersome process, recent improvements in the development of dry electrodes may simplify the use of this modality and make it usable outside of the laboratory environment.

The use of EEG in annotating multimedia data is a very new research direction and so far only a few works have investigated this area. In [6], an oddball paradigm is used in which images of a forest environment were shown to subjects for 100 ms each. The goal was to detect a small subset of target images that contained pedestrians. The target images elicit a P300 event-related potential which was then classified using Fisher linear discriminant analysis. Another test was run without the EEG modality, where subjects pressed a button upon seeing the target images. The results showed no significant differences in target image detection accuracy between the use of the EEG modality and the use of buttons. In [8], categories of images are classified based on EEG measurements recorded as the images were presented. The used categories were faces, animals and inanimate objects. This was based on the notion that the human visual system responds very differently to these categories of images. The authors propose a vision-based algorithm that uses pyramid match kernels to initially classify the images. The EEG data is then combined with the vision-based features using a kernel-alignment method. The combination of the two modalities outperforms the individual methods. In [3] the RAPID system is proposed. The authors use ERP analysis in combination with eye tracking to assist intelligence analysts in rapidly reviewing and categorizing satellite imagery. The analyst is assigned a target category to look for in the images. When subjects see an image in the target category, an ERP occurs in the EEG data which is then classified. Eye tracking is used to determine points of interest within the images.

All of these works are based on image annotation where as we attempt validation of tags related to video data. Also, in contrast to these earlier works, we perform tag validation rather than trying to assign tags directly. We show that there are significant differences between the cases of matching and non-matching tag presentations. This approach can be used in combination with a vision-based indexing and retrieval system in order to validate and re-rank its output, or for validating tags added manually by users. Such a tag validation system could be especially helpful in cases were the content to be tagged and the label categories are too com-

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

plex (and only obvious from the incorporation of a wider context) to be classified by machine learning from the media directly. In that case the human (neural) responses can be used to indirectly classify the material. Many actions, such as for instance greeting a person, can vary greatly (e.g. waving, handshaking, hugging etc.) and be very difficult to detect via machine learning techniques. However, a human observer will have no difficulty in recognising these actions.

Another possible application is be the automatic recognition of social or affective content. In [9] an N400 response was observed for labels presented after musical excerpts. These words were very loosely attributed to the music in terms of associated objects (e.g. birds, needles), musical features, and moods. While these sub-categories were not analysed and reported separately, it is conceivable that the label information can entail categories of emotional content. As emotions are subjective in nature, the N400 approach to tag validation introduced here could in principle assess the subjective response to media content, thereby crossing a threshold insurmountable by a direct media analysis.

## 2. Methodology

We propose an approach to implicit tag validation through the use of EEG signals. In this approach, a subject is shown a video followed by a tag, and from the EEG signals recorded during tag display, we aim to discern whether the tag applies to the video content or not. Our hypothesis is that if the shown tag does not match the video content a ’mismatch negativity’ will occur in the form of an N400 event-related potential (ERP). It has been shown that in cases of two semantically mismatching categories an N400 event-related potential occurs at around 400 ms after the second stimulus is presented (or better: after the mismatch becomes obvious to the viewer). This N400 has been observed even when the stimuli originate from different modalities (e.g. audio and text or images and text) [14, 12, 9, 1]. We aim to show here that the mismatch negativity can also be observed when we combine the modalities of video and text by priming the subject by the display of video content, followed by the display of a semantically mismatched tag. To the best of our knowledge this is the first work combining the video and text (tag) modalities.

We collected a large dataset with 17 subjects, each recorded for 98 trials. We use independent component analysis to remove eye blinks and other artefacts in the data and then determine whether the signals for the two cases (matching and non-matching tags) are significantly different using a repeated measures ANOVA. We found that there are indeed significant differences in the signal between the two cases in certain areas of the brain. We will now describe each step of our analysis in detail.

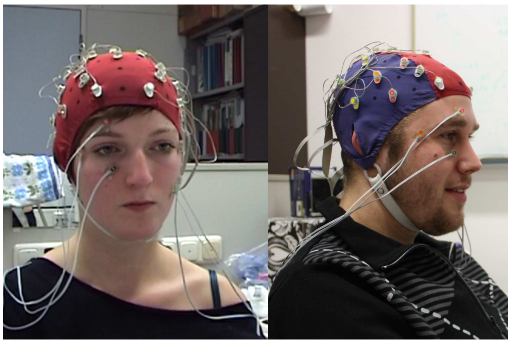


Figure 1. Subjects performing the experiment.

### 2.1. Experiment Setup

EEG was recorded using a Biosemi ActiveTwo system (www.biosemi.com) on a dedicated recording PC (P4, 3.2 GHz) using the BioSemi Actiview recording software. Stimuli were presented on a dedicated stimulus PC (P4, 3.2GHz) that sent synchronization markers directly to the recording PC. For presentation of the stimuli the Presentation software by Neurobehavioral systems (www.neurobs.com) was used. Subjects were seated in a comfortable chair, approximately 70 cm from the presentation monitor (a 20 inch Samsung Syncmaster 203B). In order to minimise eye movements, the video stimuli were all shown width a width of 640 pixels, filling approximately a quarter of the screen. Each subject signed an informed consent form and filled in a short questionnaire. They were then instructed to try to restrict any movement to the periods between trials to minimize movement artefacts in the EEG signal. Subjects were told they would be shown videos followed by tags, but were not given any further specific instructions as to the nature of the experiment. 32 active AgCl electrodes were used (placed according to the international 10-20 system) and the data was recorded at 512 Hz. Fig. 1 shows two subjects as they perform the experiment.

17 Subjects were each recorded for 98 consecutive trials. 12 subjects were male, 5 female. Ages ranged from 19 to 31, with a mean age of 25. All but two subjects were right-handed and all but three subjects viewed the tags in their native language. Each trial consisted of the following steps:

1. A fixation cross is displayed for 1000 ms (to minimise eye movements).
2. The video is displayed (ranging in duration from 6-10 seconds).
3. A fixation cross is displayed for 500 ms.
4. The tag is displayed for 1000 ms.
5. A fixation cross is displayed for 4000 ms before the start of the next trial.

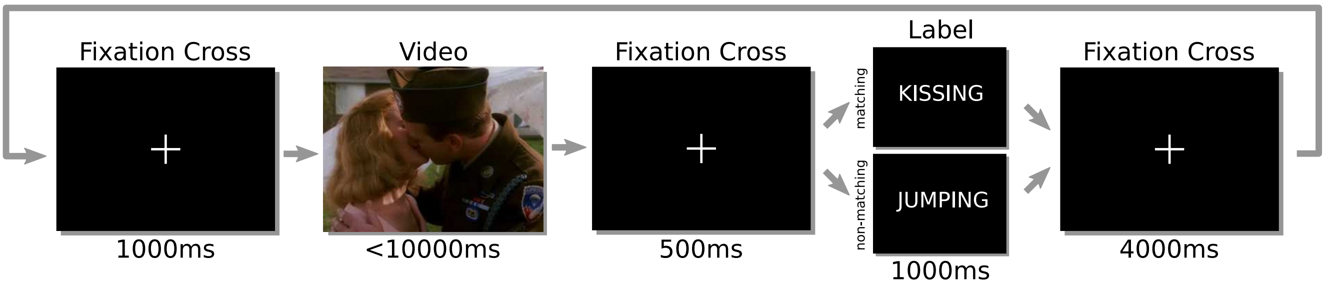


Figure 2. Order and timing of the experiment.

The stimuli were presented in 3 blocks of 32-33 trials. In between the blocks, subjects were given breaks and could move freely, reseat themselves or have a drink of water in order to avoid any muscle straining or fatigue. Fig. 2 illustrates the order and timing of the experiment.

49 Videos from seven different categories were used as stimuli, with 7 videos in each of the 7 categories. Each video has a duration of ten seconds or less and was shown twice, once followed by a matching tag and once followed by an incorrect tag. Table 1 gives an overview of the different video categories and their sources. The categories were chosen according to two criteria. Firstly, the categories should encompass events which do not vary too much in appearance within one category (to facilitate an eventual vision-based analysis). Secondly, we selected categories with human faces, animals and inanimate objects, following [8], who indicate that these categories can be separated reasonably well by analysing the EEG signals from subjects watching the videos.

### 2.2. Analysis

As a preprocessing step, the data was referenced the common average (CAR). Also, the data was bandpass-filtered between 0.5 and 40Hz to remove DC drifts and suppress the 60Hz power line interference. We extracted epochs for further analysis ranging from 500 ms before tag display to 1000 ms after. To remove interference caused by eye blinking and other artefacts, we perform spatial filtering using Independent Component Analysis (ICA). ICA has been used before in EEG data analysis with good results (e.g. [7]). Components containing only noise were manually selected and removed from the data. Fig. 3(a) is an example of a component that is strongly correlated with eye blinks. This is evident because the activation occurs in isolated periods (blinks) that are not correlated across trials. Also, the component is mostly active in the frontal electrodes. Such components are removed. Fig. 3(b) shows an example component correlated with the N100 and P200 ERP. The activation is concentrated in the occipital lobe (which is concerned with vision tasks), the component shows a resemblance to a typical ERP curve and there is a strong correlation between trials.

After removing the components that are due to blinks and other artefacts, we perform a repeated measures ANOVA to determine whether significant differences occur in the recorded EEG signal between the cases of matching and non-matching tags. For this purpose, we only consider the period of 300-500 ms after tag display, during which the strongest N400 response can be expected.

## 3. Results

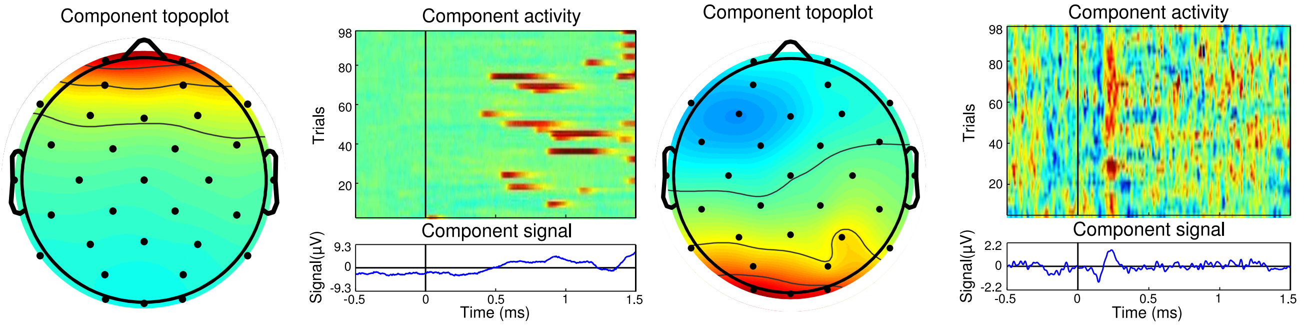
Table 2 shows the results of performing the repeated measures ANOVA. Results that have a $p$-value lower than $0.01$ are deemed significant. Electrodes that show significant differences ($p \leq 0.01$) between the cases of matching tags and non-matching tags are highlighted. The fourth column shows the mean signal difference between the cases in $\mu V$ (the mean signal in the case of matching tags minus the mean signal for the case of non-matching tags). Electrodes showing a significantly higher/lower negativity for non-matching tags are shaded light red and darker blue respectively.

Fig. 4 shows the location of observed differences in signal values. We can see that the differences are spatially mainly localised in two regions. The main region is located around the occipital and parietal lobe (covering electrodes CP1, Pz, PO3, CP2, C4 and Cz), where a more negative voltage deflection occurs when displaying non-matching tags than when displaying matching tags. The occipital lobe is concerned primarily with vision tasks and the parietal lobe is, among other things, concerned with the location of visual attention [10, 4]. The other region showing a significant difference in signal values is located in the left temporal lobe around electrodes AF3, FC5, T7 and F7. One of the functions of the left temporal lobe is the recognition of words, possibly explaining the activation there. In this case, the observed voltage is less negative for the case of non-matching tags than for the case of matching tags.

Fig. 5 depicts the grand average waveforms for the 9 electrodes exhibiting the most significant differences be-

| Category/Label | Source |
|---|---|
| Airplane take off | Plane spotter homevideos (http://www.flightlevel350.com/) |
| People kissing | Hollywood movies dataset [11] |
| People getting out of cars | Hollywood movies dataset [11] |
| Mice drinking water | Mouse behaviour dataset [5] |
| Cats opening doors | Pet homevideos (http://www.youtube.com) |
| Jawdrop (posed facial expression) | MMI facial expression database [13] |
| Laughing people (spontaneous facial expression) | AMI meeting corpus [2] |

Table 1. The different video event categories used in the experiment and their sources.



(a) An independent component that is strongly correlated with blinks. The component activity is concentrated in the frontal area and there is no correlation between trials.

(b) An independent component that is correlated with ERPs in the occipital cortex related with early visual processes. We can primarily see the activation here of the N100 and P200 ERP.

Figure 3. Visualisation of two independent components. In each of the subfigures: On the left is a topoplot of the component activation. In the top right the component activation is shown for 98 trials of one subject. In the lower right the average component signal is displayed.

tween the two cases. The first four plotted electrodes show less negativity for non-matching tags than for matching tags. The remaining electrodes show the opposite behaviour and display a higher negativity for the case of non-matching tags than for matching tags. Clear examples of the N400 ERP can be observed. The differences are most clear in the 300-500 ms period after tag display.

From these results it is clear that the N400 occurs when subjects are shown a combination of stimuli from the modalities of video and text (in the form of a tag). Furthermore, significant differences are present in a considerable number of electrodes between the cases of non-matching and matching tags. However, the effect size ($\leq 1\mu V$) is smaller than that found in other studies (e.g. [12, 1]). This can be due to the semantic categories, the stimulus material, or other parameters of the experiment used here.

## 4. Conclusions

In this work, we have collected and analysed a dataset to investigate the use of EEG for passive, implicit tag validation. Data was collected for 17 subjects and each subject was shown 98 videos, 49 followed by with matching tags and 49 followed by non-matching tags). Independent Component Analysis was used to remove noise (including eye blink artefacts) from the data. A repeated measures ANOVA showed significant differences in the EEG signal between the two cases of congruent and incongruent tags. This implies that the two cases can be successfully distinguished by analysis of the EEG signal. The next step in our research is to determine for single data trials whether the tag matches the video content. Successful single trial analysis would mean we can use this technique as a feedback mechanism in video analysis for indexing and retrieval. Other uses could include validating unreliable user-generated tags and possibly determining user reactions to the content (such as liking or disliking the content or other affective reactions).

In order to achieve a working tag validation system several parameters will have to be studied and optimized. Questions that need to be answered include: how long after a stimulus does a non-matching tags still elicit the ERP? What types of categories elicit the most robust mismatches? Does a subliminal presentation, not consciously perceived by the viewer, also elicit N400 responses? Can we also use a frequency analysis to judge how subjects implicitly judge the semantic meaning of the video?

In similar P300 experiments usually the EEG signal of several trials is averaged to increase the signal-to-noise ratio and increase the accuracy of ERP detection. This strategy could in principle also be used for the evaluation of label validity. However, it has to be ensured that multiple presented tags really are associated with the media content and

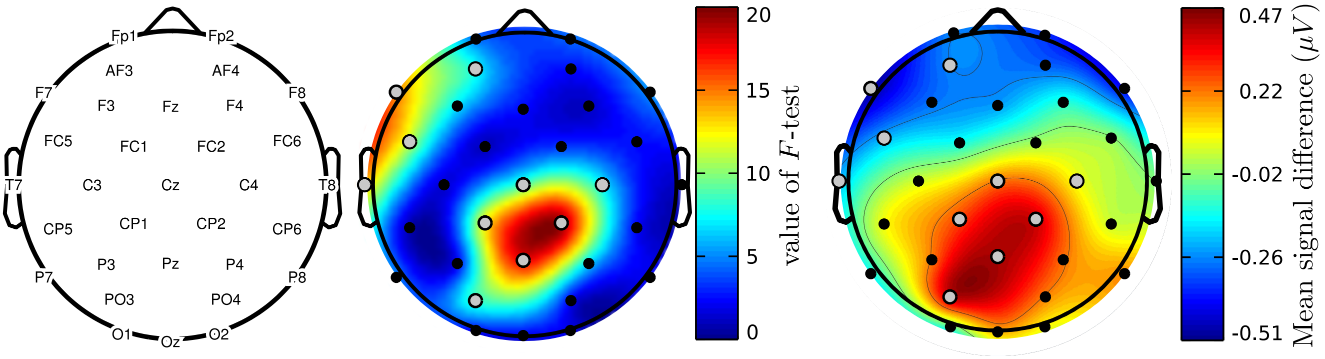


Figure 4. Left: Topoplot of Electrode locations, Middle: Topoplot of Significance of difference ($F$-test value), Right: Topoplot of the Grand-average differences between 300 and 500 ms for all 17 subjects. Electrodes with significant differences are highlighted in grey.

| Electrode | $F(1,16)$ | $p$-value | MSD ($\mu V$) |
|---|---|---|---|
| CP2 | 19.98 | 0.000 | 0.776 |
| Pz | 17.59 | 0.000 | 0.819 |
| CP1 | 11.74 | 0.001 | 0.535 |
| Cz | 08.15 | 0.004 | 0.480 |
| PO3 | 07.32 | 0.007 | 0.616 |
| C4 | 06.86 | 0.009 | 0.364 |
| F7 | 15.25 | 0.000 | -0.948 |
| T7 | 14.15 | 0.000 | -0.758 |
| FC5 | 11.76 | 0.001 | -0.640 |
| AF3 | 07.84 | 0.005 | -0.557 |
| F3 | 06.50 | 0.011 | -0.482 |
| P4 | 06.30 | 0.012 | 0.478 |
| P7 | 04.53 | 0.034 | -0.443 |
| F8 | 04.42 | 0.036 | -0.455 |
| Fp2 | 03.68 | 0.055 | -0.417 |
| Fp1 | 03.65 | 0.056 | -0.429 |
| FC2 | 02.29 | 0.130 | 0.260 |
| Fz | 01.69 | 0.194 | -0.261 |
| AF4 | 01.61 | 0.204 | -0.250 |
| FC6 | 01.57 | 0.210 | 0.203 |
| CP6 | 01.33 | 0.249 | 0.236 |
| P3 | 01.18 | 0.278 | 0.196 |
| P8 | 01.09 | 0.297 | 0.209 |
| O2 | 00.71 | 0.399 | 0.182 |
| PO4 | 00.63 | 0.427 | 0.180 |
| O1 | 00.63 | 0.428 | -0.173 |
| C3 | 00.43 | 0.514 | 0.094 |
| F4 | 00.09 | 0.761 | 0.055 |
| T8 | 00.08 | 0.783 | 0.053 |
| Oz | 00.06 | 0.808 | 0.056 |
| CP5 | 00.03 | 0.870 | -0.027 |
| FC1 | 00.00 | 0.995 | 0.001 |

Table 2. ANOVA Results per electrode. MSD stands for Mean Signal Difference. Significant differences ($p \leq 0.01$) are highlighted.

not with previously presented labels.

Using a single trial analysis, we hope to build a tag validation system that will achieve an efficiency close to that of manual tagging without active user involvement. However, given the low bitrate usually achieved by BCI systems, this task seems rather daunting. Also, mere tag validation does not compare to a complete manual tagging. Nevertheless, we envision a system that will be a useful addition to current tagging methods, especially given the absence of the requirement for active user involvement.

## Acknowledgement

The research leading to these results has received funding from the Seventh Framework Programme under grant agreement no. FP7-216444 (PetaMedia).

## References

[1] V. Bostanov and B. Kotchoubey. The t-CWT: A new ERP detection and quantification method based on the continuous wavelet transform and Students t-statistics. *Clinical Neurophysiology*, 117(12):2627–2644, 2006.

[2] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007.

[3] A. Cowell, K. Hale, C. Berka, S. Fuchs, A. Baskin, D. Jones, G. Davis, R. Johnson, R. Patch, and E. Marshall. Brainwave-Based Imagery Analysis. *Digital Human Modeling: Trends in Human Algorithms*, pages 17–27, 2008.

[4] A. Cummings, R. Čeponienė, A. Koyama, A. Saygin, J. Townsend, and F. Dick. Auditory semantic networks for words and natural sounds. *Brain research*, 1115(1):92–107, 2006.

[5] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *IEEE Int'l Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 65–72, 2005.

[6] A. Gerson, L. Parra, and P. Sajda. Cortically coupled computer vision for rapid image search. *IEEE Trans. Neural Systems and Rehabilitation Engineering*, 14(2):174–179, 2006.

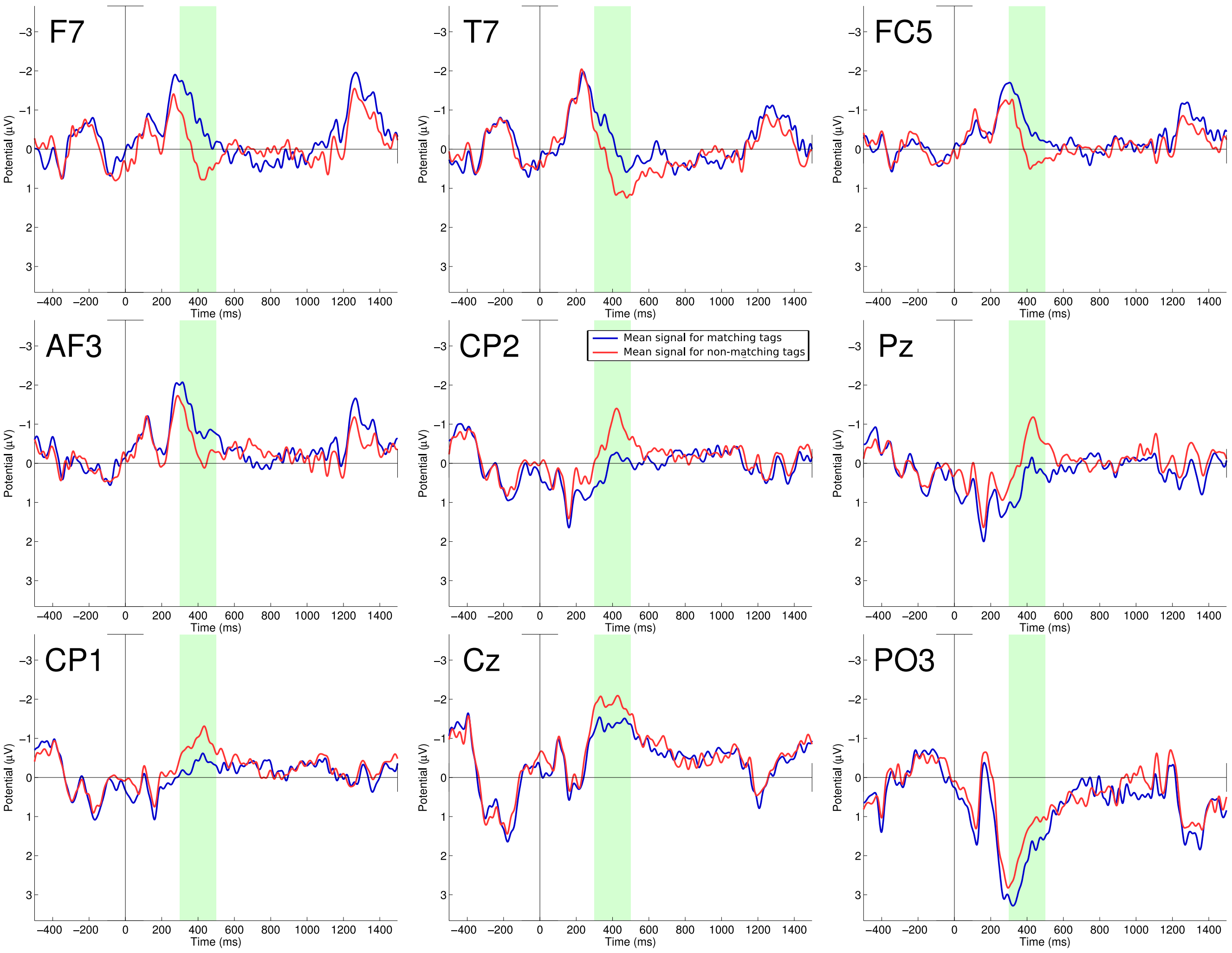


Figure 5. Grand average waveforms for the period 500 ms before to 1500 ms after tag presentation for the 9 electrodes with the most significant differences. The signal is averaged over all trials and subjects. The red line shows the average signal during presentation of matching tags and the blue line shows the average signal for non-matching tags. Differences in signal values between the two categories can be observed in each plot around the 400 ms mark. The light-green shaded area is the window used for ANOVA analysis. A 30 Hz low-pass filter was used in these plots, for display purposes only. Note that for the y-axis, negative is up.

[7] B. Kamousi, Z. Liu, and B. He. Classification of motor imagery tasks for brain-computer interface applications by means of two equivalent dipoles analysis. *IEEE Trans. on Neural Systems and Rehabilitation Engineering*, 13(2):166–171, 2005.

[8] A. Kapoor, P. Shenoy, and D. Tan. Combining Brain Computer Interfaces with Vision for Object Categorization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[9] S. Koelsch, E. Kasper, D. Sammler, K. Schulze, T. Gunter, and A. Friederici. Music, language and meaning: brain signatures of semantic processing. *Nature Neuroscience*, 7:302–307, 2004.

[10] M. Kutas and S. Hillyard. Event-related brain potentials to grammatical errors and semantic anomalies. *Memory & Cognition*, 11(5):539–550, 1983.

[11] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[12] G. Orgs, K. Lange, J. Dombrowski, and M. Heil. Is conceptual priming for environmental sounds obligatory? *International Journal of Psychophysiology*, 65(2):162–166, 2007.

[13] M. Pantic, M. Valstar, R. Rademaker, and L. Maat. Web-based database for facial expression analysis. In *Proc. IEEE Int. Conference on Multimedia and Expo*, pages 317–321, July 2005.

[14] C. van Petten and H. Rheinfelder. Conceptual relationships between spoken words and environmental sounds: Event-related brain potential measures. *Neuropsychologia*, 33(4):485–508, 1995.