ICANN/PASCAL2 Challenge: MEG Mind Reading — Overview and Results

Arto Klami[1], **Pavan Ramkumar**[2], **Seppo Virtanen**[1], **Lauri Parkkonen**[2], **Riitta Hari**[2], **Samuel Kaski**[1,3]

[1,2]Aalto University School of Science
[1]Department of Information and Computer Science
[1,3]Helsinki Institute for Information Technology HIIT
[2]Brain Research Unit, Low Temperature Laboratory
[3]University of Helsinki

Abstract

This report summarizes the modeling challenge held in conjunction with the International Conference on Artificial Neural Networks (ICANN) 2011 and sponsored by the PASCAL2 Challenge Programme. The challenge aimed at promoting awareness of the task "mind reading" or "brain decoding" based on magnetoencephalography (MEG) data. For neuroscientists, the task provides a practical tool for understanding brain process underlying perception, since any mechanism that can be used for inferring the stimulus on the basis of brain activity must be related to processing of the stimulus. For machine learners and other modelers, the challenge provides an interesting real-world application playground for solving active machine learning problems such as multi-view learning and covariate shift.

The task was to infer from one-second time windows the type of visual stimulus shown to the subject. The best brain decoders, out of the 10 submissions, reached almost 70% *accuracy in the task with mere* 23% *chance-level, proving that even a short MEG measurement can be sufficient for brain decoding tasks with a reasonable number of stimulus categories.*

1 Introduction

A grand challenge in neuroscience is to understand the neural basis of sensory and cognitive processing, even to the extent to predict brain correlates of novel stimuli. This challenge can be formulated as a decoding problem: given the brain signals, read out some information about the stimuli that generated (or modulated) them [1]. The information read out can be category specific, identity specific, or the entire stimulus itself—corresponding to the machine learning tasks of classification, identification, or regression/reconstruction. Such decoding tasks are often called brain/mind decoding, or multivariate/multivoxel pattern analysis (MVPA).

The majority of the reported brain decoding results derive from functional magnetic resonance imaging (fMRI), from attempts to decode relatively simple properties or to choose the correct alternative amongst a few choices. For example, Kamitani et al. [2] inferred the orientation of edges out of 8 possible alternatives and Formisano et al. [3] identified what (out of three vowels) and whom (out of three alternative speakers) the subject was listening to. Recent studies have shown significant progress in decoding more and more complex perceptual phenomena, resulting in successful identification of natural images [4] and the meaning of nouns [5] in setups where the set of possible alternatives is larger, in the order of tens. All of these works fall into the category of classification or identification. Miyawaki et al. [6] have studied the task of reconstruction of small binary images from local image patches decoded from brain signals, and Naselaris et al. [7] extended reconstruction tasks to natural images.

While fMRI has very high spatial resolution throughout the brain, it has poor temporal resolution and the blood oxygenation level dependent (BOLD) signal is an indirect measure of neuronal activity. Riger et al. [8] have shown that it is possible to apply decoding similarly to magnetoencephalography (MEG); they predicted on the basis of single-trial MEG signals whether the subject recognized and memorized a natural image. With MEG it will be possible to focus on shorter timescales. Of particular interest is the feasibility of brain decoding for continuous processes using e.g. speech or video stimuli. Besides attempting to decode external stimuli, MEG has also been used for decoding the direction of hand-movement [9] or reconstructing hand-movement trajectories [10]. Nevertheless, the task of brain decoding from MEG is still in its infancy.

From another point of view, the brain decoding task can be seen purely as a challenging machine learning problem. The recorded brain signals are very high-dimensional and noisy, and consequently advanced classification or regression methods are needed for solving the prediction task. This is also demonstrated in practical work, with focus on advanced Bayesian solutions [10, 11] and completely novel types of machine learning strategies, such as the zero-shot learning concept [12]. Furthermore, many of the current trends in machine learning are highly relevant for solving the brain decoding challenges: (1) the models need to handle covariate shifts (changes in the input distribution between training and test data) [13] with approaches like domain adaptation [14], (2) sparse solutions such as lasso regression [15] are likely to be effective for the high-dimensional data sources, (3) the prediction tasks should ideally combine information from multiple sources through multi-view learning, and (4) especially analysis of multiple subjects would benefit from multi-task learning methods [16].

We organized the challenge for brain decoding based on MEG signals for four primary reasons. (1) To increase the awareness of the problem amongst both machine learning researchers and neuroscientists, (2) to study the feasibility of decoding continuous visual stimulus from short periods of MEG recordings, (3) to bring up some of the relevant methodological challenges for MEG brain decoding, and (4) to provide a simple benchmark data set. The challenge was organized in co-operation with the ICANN conference since it attracts machine learning researchers with interest in modeling neural processes. The motivations are largely shared by other recent attempts of promoting visibility of brain decoding in general, such as the 1st ICPR workshop on brain decoding, organized in conjunction with the 20th International Conference on Pattern Recognition.

2 Data

2.1 Stimuli

The brain decoding task in the challenge was to recognize the type of video stimulus shown to the subject. All videos were presented without audio, and five different types of stimuli were used:

1. Artificial: Screen savers showing animated shapes or text

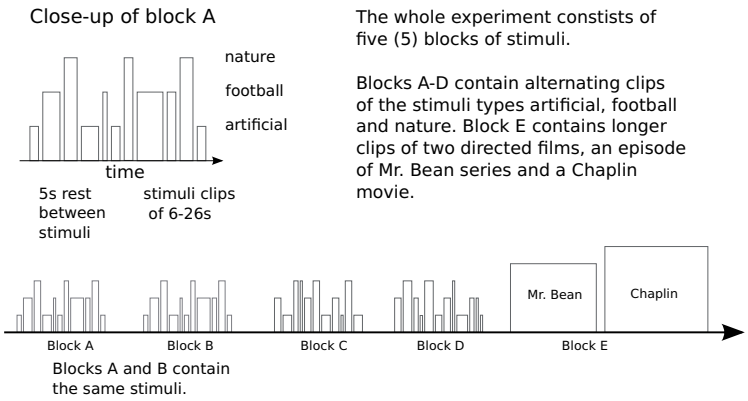

Figure 1. Illustration of the stimulus design. The subject viewed the same set of 5 blocks during two consecutive days. The first four blocks, labeled A-D, contained alternating short clips of artificial objects (animated shapes or text), football or nature documentaries, whereas the last block contained longer clips taken from a television series and a feature film. Within blocks A-D the different clips were separated by 5-s rest period showing a crosshair, and the clips lasted for 6-26 s. The two longer clips in block E, extracted from video content with a storyline, lasted for roughly 10 minutes.

2. Nature: Clips from nature documentaries, showing natural sceneries like mountains or oceans

3. Football: Clips taken from (European) football matches of Spanish La Liga

4. Mr. Bean: Clip from the episode “Mind the baby, Mr. Bean” of the Mr. Bean television series

5. Chaplin: Clip from the “Modern times” feature film

The stimuli were shown in five blocks (Figure 1). The first four blocks (A–D) contained alternating short clips of the first three stimulus types, so that each block contained a roughly equal number of clips for each stimulus type in random order. The clips lasted 6-26 s, and the different clips were separated by 5-s rest periods showing a crosshair in the center of the visual field. The first two blocks were identical, whereas blocks C and D contained different video clips.

After the four blocks described above, the subject viewed two continuous video clips containing a clear plot and storyline (clips from an episode of a television series and a feature film), each lasting roughly 10 min. These two clips were shown during the same experiment block.

2.2 Recording and preprocessing

We recorded MEG signals from one healthy 25-yrs old male who gave his written permission for releasing the data for the challenge.

MEG was acquired with a 306-channel Elekta Neuromag MEG system (Elekta Oy, Helsinki, Finland) with a basspand from DC to 330Hz and digitized at 1000 Hz. During the MEG recording, four small coils, whose locations had been digitized with respect to anatomical landmarks, were briefly energized to determine the subject's head position with respect to the MEG sensors. The continuous raw MEG data were further low-pass filtered at 50 Hz, and downsampled to 200 Hz. External interference was removed and head movements compensated for by using the signal-space-separation (SSS) method [17]. Finally, we applied piecewise mean and trend removal for each channel to compensate for very slowly varying signals that are likely to be artefacts.

Since identifying the videos would be relatively easy based on long sequences of MEG recordings, we chose to hand out only short 1-s signal epochs in random order. However, handing out only the raw measurement data would have resulted in a challenge that requires considerable expertise on MEG. In addition, it would have prevented reliable estimation of low-frequency waveforms because sharp filters could not be applied for signals as short as 1 s (200 samples). Consequently, we chose to precompute a number of features at different frequency bands. We applied a bank of 5 band-pass filters peaked at 2, 5, 10, 20, and 35Hz, and computed the envelopes of the signals at these frequencies by taking the absolute value of the Hilbert-transformed signal. The details of the filter bank are provided in Table 1.

For each sample (1-s epoch of the recording) the participants received six different data matrices, each containing 200 time points for 204 gradiometer channels of the MEG device. Those data matrices corresponded to the raw signals after the SSS preprocessing, and the envelopes at the five frequencies mentioned above.

3 Modeling problem

The modeling problem was to infer the stimulus from brain signals. Given the limited set of possible stimuli, this was a classification task: For each

Table 1. Details of the filter bank. The first column indicates the name of the filter, identified with a frequency within the band-pass area determined by the next two columns. The filters were Kaiser window FIR filters with stop bands increasing from 0.5Hz to 2Hz with increasing frequency. The order of the filters is shown in the last column.

Peak freq. (Hz)	Min freq. (Hz)	Max freq. (Hz)	Order
2	1	4	2009
5	4	7	2009
10	7	13	503
20	17	23	503
35	27	43	503

input signal the task was to infer the type of the stimulus. Consequently, the challenge was formulated as a classification problem. Given a set of labeled training examples, the task was to infer the labels for left-out test data.

For brain decoding, the generalization to new stimuli is critical. While the set of possible stimulus types needs to be limited to make inference possible, the actual stimulus content should be different for training and test samples. After all, the goal is not to recognize when the subject is watching a particular clip of a football match, but to identify the process of watching football in general. Besides generalizing to new stimuli, a brain decoding system will need to generalize over different recording sessions.

3.1 Data split

For studying the above properties, the data were split into training and test sets so that the following properties were satisfied:

- Some of the training and test instances were recorded using the same stimuli, whereas some test instances were taken from recordings of different stimuli of the same type. In total, 33% of the test samples consisted of recordings during stimuli not seen in the training phase.

- The training and test data were taken from different recording sessions. In particular, the training and test data were recorded during different days.

- A small portion of the test samples were labeled, to simulate brief train-

ing period during the test session and to enable studying possible differences between the data distributions.

- The samples were not continuous in time, to prevent attempts of ordering the samples given in random order.

The detailed split into training and test samples is described in Figure 2. In brief, both the training and test samples were of 1-s length and were separated from each other by 1 s. Out of the four blocks of short clips the blocks A, B and D recorded during day 1 were used for training and blocks A, B and C recorded during day 2 for testing. This resulted in 66% of the test samples having stimuli that exists also in the training data. The clips in block E were split to training and test data so that time (roughly) between 1:40 and 6:10 was used for training and time between 3:10 and 7:40 for testing, resulting in 68% of overlap between training and test data. Finally, a random class-balanced subset of 50 test samples were released with labels.

The training and test samples had, however, 1-s offset in timings. Hence, even the set of samples using the same stimuli are not exactly from the same time but instead are consecutive time points. If the time window between seconds 3 and 4 was used for training, then the window between seconds 4 and 5 was used for testing.

Overall, the setup resulted in 677 training samples with roughly class-balanced distribution (the number of samples for the five classes were: 140, 171, 96, 135, and 135), 50 labeled test samples, and 653 unlabeled test samples that the competitors needed to classify. The data are available at `http://www.cis.hut.fi/icann2011/mindreading.php`, and can be used for research purposes and scientific publications.

3.2 Machine learning concepts

Even though the main problem is that of regular classification, the particular setup of learning to decode MEG measurements leads to a number of more detailed machine learning challenges. Here we briefly overview the kind of aspects initially thought to be relevant for the task. The research on machine learning solutions for MEG mind decoding tasks would likely benefit from tackling these modeling issues, besides just working towards improved MEG signal analysis in general.

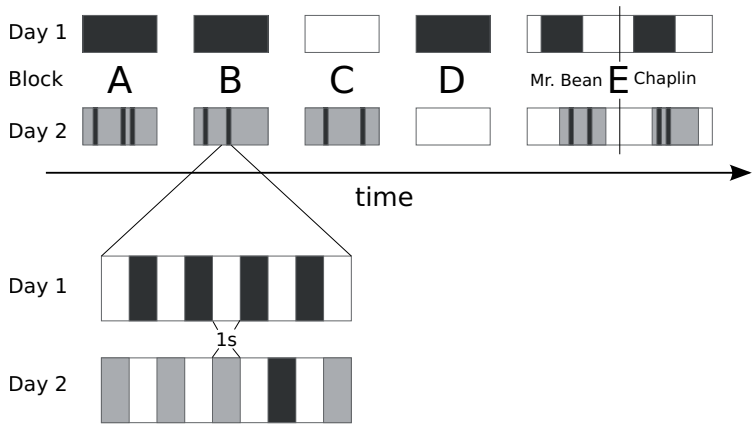

Figure 2. Illustration of the data split. The dark grey boxes correspond to the selection of data points for training data, the light gray boxes correspond to the choice of test samples, whereas the unshaded areas were not used in the challenge at all. The dark areas on the second day indicate the random choice of labeled test samples. Note that blocks A and B contained the same stimuli. The closeup shows how the 1-s samples were chosen with 1-s gaps between each other, and how the training and test samples taken from the same block were misaligned by 1 s.

Covariate shift/domain adaptation For real-use cases brain decoding systems need to work for new recording sessions, besides being able to predict merely new time points of existing recordings. Since (1) MEG instrumentation is subject to stochastic noise, and (2) since the state of the subject varies strongly from day to day, the data recorded during a different session generally do not follow the same distribution as the training data. Hence, computational models taking into account a change in the data distribution are needed. This problem is generally tackled under the term of domain adaptation [14], which is an active line of research in the machine learning community.

Multi-view learning MEG recording produces measurements for 204 gradiometer channels and 102 magnetometer channels, and for each signal we can extract multiple frequency bands or other types of features. Information encoded in different channels, frequency bands, and across different time scales is largely complementary. This suggests that multi-view learning methods could be useful for MEG decoding tasks. While it is possible to attempt decoding the stimuli from individual channels or based on simple predictors operating on all channels, there is reason to believe that clever integration of the different channels and frequency bands through

multi-view learning models could result in improved accuracy, as well as improved understanding of the underlying brain processes.

Multi-view learning methods have also been used for solving decoding tasks outside classification. For identification, multi-view learning methods based on canonical correlation analysis (CCA), such as the Bayesian CCA [18], can be used for extracting correlating projections of the brain activity and stimulus description, enabling direct comparison of brain measurements of test samples to the set of possible stimuli. Multi-view learning methods have also been used for extracting image bases for visual image reconstruction [19], as well as for inferring properties of natural music based on fMRI [20].

Generalization and overfitting Another consequence of the high-dimensional nature of the MEG recording is that it is very easy to overfit to the available training data. Therefore a successful decoding solution will have to be very carefully regularized to control the degree of generalization to new data. Many of the decoding works apply Bayesian modeling techniques [10, 11], which provide a way of tackling the overlearning issue in a justified way, or apply sparse solutions such as lasso regression [15].

Multi-task learning The variability across subjects is large for all brain imaging techniques. Typical analysis methods will either assume that all subjects are identical, which is a simplifying but incorrect assumption, or will resort to subject-specific modeling resulting in no information being transferred from one subject to another. Multi-task learning [16] studies computational models that combine the strengths of both approaches, by learning separate predictive models for the subjects simultaneously, so that the similarities between the subjects are utilized for improved accuracy while still allowing subject-specific variation. In this challenge, we provided data only from a single subject, and hence such models could not be applied, but in general multi-task learning of decoding models is likely to be crucial. Recently, Alamgir et al. [21] demonstrated how multi-task learning improved accuracy for EEG-based brain computer interfaces.

Table 2. The list of participating teams in alphabetical order of the first author. The team Tu & Sun provided two different solutions.

Name	Authors / Institute
Van Gerven & Farquhar	M.A.J. Van Gerven, J. Farqugar Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen, the Netherlands
Grozea	C. Grozea Fraunhofer Institute FIRST, Germany
Huttunen et al.	H. Huttunen, J-P. Kauppi, J. Tohka Department of Signal Processing, Tampere University of Technology, Finland
Jylänki et al.	P. Jylänki, J. Riihimäki, A. Vehtari Dept. of Biomedical Engineering and Computational Science, Aalto University, Finland
Lievonen & Hyötyniemi	P. Lievonen, H. Hyötyniemi Helsinki Institute for Information Technology HIIT, Finland
Nicolaou	N. Nicolaou Dept. Of Electrical and Computer Engineering, University of Cyprus, Cyprus
Olivetti & Melchiori	E. Olivetti, F. Melchiori NeuroInformatics Laboratory (NILab), Bruno Kessler Foundation and University of Trento, Italy
Santana et al.	R. Santana, C. Bielza, P. Larrañaga Departamento de Inteligencia Artificial, Universidad Politécnica de Madrid, Spain
Tu & Sun	W. Tu, S. Sun Department of Computer Science and Technology, East China Normal University, China

4 Results

Overall, the challenge received 10 submissions from 9 different teams listed in Table 2. Multiple submissions per team were allowed if the solutions utilized significantly different modeling approaches.

4.1 Challenge results

The main criterion for evaluating the submissions was the classification accuracy on the test data. The baseline accuracy of predicting every sample to belong to the largest class in the training set would be 23%. The results of the participants are summarized in Table 3, showing that all

Table 3. The prediction accuracies (percent, bigger is better) of the competitors, sorted in the order of the overall accuracy that was the criterion for evaluating the submissions. The last three columns show the accuracy in separating the content with plot from the short clips (PvsC), the accuracy in predicting the short clip classes correctly (C), and the accuracy in identifying the longer clips with plot correctly (P). For all tasks the best accuracy has been boldfaced. A notable observation is that the best solution outperforms all others in making the correct predictions within both stimulus categories, but is only 7th best in making the split between the two categories. The last line shows the accuracy of majority voting based on the top nine submissions.

Team	Accuracy	PvsC	C	P
Huttunen et al.	**68.0**	89.7	**67.5**	**89.2**
Santana et al.	63.2	93.0	64.1	74.0
Jylänki et al.	62.8	93.0	56.8	85.8
Tu & Sun	62.2	**97.1**	50.1	87.0
Lievonen & Hyötyniemi	56.5	91.0	55.7	72.4
Tu & Sun (2)	54.2	96.6	44.3	75.8
Olivetti & Melchiori	53.9	94.6	41.4	85.4
Van Gerven & Farquhar	47.2	82.4	53.3	66.3
Grozea	44.3	88.5	39.1	67.7
Nicolaou	24.2	61.7	34.8	49.6
Pooled (top 9)	69.2	96.8	63.1	85.8

but one of the participants clearly surpass the baseline level, demonstrating successful brain decoding. The outlier submission falls at the chance level, suggesting either very heavy overlearning or mistakes in implementation. The range of accuracies, excluding the outlier, falls between 44% and 68%, demonstrating that there is a notable difference between the alternative decoding solutions. The solution of Huttunen et al. outperforms others by a margin of almost five percent, ending up as the clear winner, followed by three other solutions above 60% accuracy.

For many classification tasks combining several classifiers results in improved performance. While various advanced solutions, such as boosting, can be used for obtaining maximal benefit from multiple classifiers, already a simple majority voting of the results provides often a reasonably good model. Here, the combination of all 10 solutions results in accuracy of 68.9% and the combination of the 9 solutions exceeding the chance level gives 69.8%. Both figures are better than the best solution, but the margin is smaller than the difference between the individual solutions.

As the stimuli to be decoded consisted of two distinct categories, directed

Huttunen et al.

	1	2	3	4	5
1	94	29	16	10	1
2	22	100	10	18	1
3	25	16	51	10	0
4	3	4	12	85	21
5	2	2	4	3	114

Santana et al.

	1	2	3	4	5
1	67	54	14	15	0
2	25	110	5	11	0
3	19	14	57	12	0
4	1	1	5	59	59
5	1	0	0	4	120

Jylänki et al.

	1	2	3	4	5
1	67	32	43	8	0
2	36	89	18	8	0
3	30	6	61	4	1
4	6	6	11	78	24
5	1	0	1	8	115

Tu & Sun

	1	2	3	4	5
1	56	55	36	3	0
2	30	96	21	4	0
3	33	22	46	1	0
4	4	3	3	95	20
5	1	0	0	11	113

Figure 3. Confusion matrices of the top four submissions. The rows correspond to the true classes, whereas the columns are the predicted classes. The labels are 1:artificial, 2:football, 3:nature documentary, 4:Mr.Bean, 5:Chaplin.

films with clear storyline and short video clips, we can also look at the success rate in separating these two categories as well as the accuracy in classifying the samples within either category (Table 3). The accuracy in separating the two categories is computed as the binary classification accuracy, whereas the accuracy within each category is measured with the ratio of correct assignment amongst all samples for which both the true and predicted class are within that category. Interestingly, the best submission is not amongst the top ones in the easier task of separating the clips with plot from the rest, but has the best accuracy within both stimulus categories. One possible reason is that the other solutions have overfitted to solving the easier task of binary separation between the two categories. This is illustrated by the confusion matrices of the best four solutions in Figure 3.

The best solutions are described in more detail in the separate articles following this overview. Overall, the solutions focused quite strongly in feature selection, either by careful validation of possible alternative features or by building classifiers with automatic feature selection, such as L1-regularized lasso models. One team, Santana et al., tried an ensemble of more than one classifier. Three of the competitors, Olivetti & Melchiori

and both submissions by Tu & Sun, focused on solving the domain adaptation problem with advanced machine learning techniques, each having reasonable performance but not reaching the top positions, while many of the other teams addressed the shift in input distributions by placing more weight on the labeled test examples when validating the learned classifier.

4.2 Alternative prediction tasks

Even though the challenge was defined as decoding the stimulus based on 1-s MEG epochs, we can estimate how well the solutions would have fared with longer observations by pooling the predictions given for consecutive samples. For this purpose, we looked at the predictions obtained by majority vote for each short clip (classes 1-3), averaging as 8 observation per clip, and for each collection of 8 consecutive samples for the longer clips (classes 4-5). The best submission then gives 80% accuracy in predicting the class correctly for each clip or 8s period (Table 4), supporting the intuitive belief that solving the decoding task is easier based on longer observations. These accuracies provide a lower bound for the accuracy the competitors could have obtained if they had access to such 8s observations and had explicitly developed predictors for solving this alternative task.

As described in Section 3.1, the data set was split so that some of the test samples were picked from the same clips as the training samples (though with 1-s offset) while some were not. Even though the competitors were not aware which samples matched the training samples, we can inspect whether the accuracy of decoding differs from the two sets. Table 4 shows how almost all participants were more accurate in predicting the samples taken from the same clips that were available in training, providing a quantification of the increase in difficulty in brain decoding due to completely new stimulus content. On average, the accuracy was 6.3 percentage points higher for the samples included in the training data.

5 Discussion

The primary task in the challenge was to decode the type of the video stimulus from MEG data. Nine out of ten submissions succeeded in this task significantly above the chance level, showing that it is possible to decode

Table 4. Results of alternative decoding tasks (not part of the competition), sorted in order of the performance in the challenge results. For each task the best accuracy is boldfaced. The first column shows the accuracy for predicting correctly the whole clips by majority voting based on the samples within each clip (on average 8 samples per clip). For all but one participant the accuracy is better than when decoding the label for 1s samples, as expected. The second column gives the accuracy in the challenge decoding task for test samples taken from the clips used also in the training set, whereas the last column gives the accuracy for the test samples from clips not seen in the training set. For all contestants except one, the accuracy is better for the first group, showing clearly how generalizing to the new stimulus content makes the decoding task harder. Still, the accuracies for the new content are well above chance level.

Team	Full clips	Within train	Not in train
Huttunen et al.	**79.7**	**69.9**	**64.2**
Santana et al.	68.5	65.1	59.6
Jylänki et al.	76.0	66.2	56.0
Tu & Sun	70.7	64.4	57.8
Lievonen & Hyötyniemi	62.2	59.8	50.0
Tu & Sun (2)	57.0	59.5	43.6
Olivetti & Melchiori	61.8	55.6	50.5
Van Gerven & Farquhar	55.9	49.0	43.6
Grozea	41.3	44.4	44.0
Nicolaou	25.0	23.7	25.2

the various kinds of stimuli already from short 1-s windows of MEG data. The difference in accuracies between the approaches was considerable, with the best solution reaching near 70% accuracy while the majority of the solutions had around 50% success, showing that carefully developed machine learning solutions will achieve improved accuracy in brain decoding. Still a clear gap exists between the best solution and perfect accuracy, demonstrating that the task is far from trivial and especially that perfect decoding results are unlikely to be obtained with such brief signals, probably because of the low signal-to-noise ratio of the single-trial MEG. By pooling the competitors results for longer (8 s) periods of observations, the accuracy of the best solutions increases close to 80%. In future research it could be advisable to directly study the accuracy on multiple timescales, to better estimate the amount of data needed for inferring different types of stimuli.

The majority of the competitors focused on good feature selection and cross-validation of the learned models, demonstrating once again the importance of carefully controlling overlearning. In this challenge this aspect was particularly important due to the relatively big change in input data distribution between training and test data. For example, the top team explicitly mentioned in their submission that some of the more advanced features were neglected for that reason. Many of the teams also addressed the domain adaptation problem seriously. Some of the competitors handled the adaptation by giving more weight for the labeled test samples in cross-validation, whereas some teams applied more advanced techniques for correcting for the shift in the distribution, using methods of EasyAdapt, transfer-priority cross validation and transferable discriminant analysis.

In future, it would be interesting to see challenges with brain signals from more subjects. This would enable studying more advanced modeling concepts such as multi-task learning, while also providing information on to which extent the perceptual processes that are best for decoding the stimuli are shared by individuals. However, prior to releasing such data sets it could be beneficial to create a more finely processed feature set, since otherwise the amount of data becomes infeasible. Now the data of just one subject took a total of roughly 6 gigabytes in compressed format, and started to become a technical difficulty for some competitors.

For this challenge we used decoding accuracy as the primary criterion and evaluated the submissions additionally based on methodological nov-

elty of the approach. The challenge was also primarily advertised for modelers. Consequently, the submissions focused on these aspects and no neuroscientific interpretations were made. In future challenges it could be a good idea to value also neuroscientific findings when determining the winners, to encourage tighter interaction between modelers and neuroscientists as well as to provide insights into the perceptual processes revealed by succesful decoding.

Acknowledgments

We gratefully acknowledge the PASCAL2 European Network of Excellence, in particular their Challenge Programme, for sponsoring the challenge. We also thank the ICANN conference for hosting the challenge workshop.

The organizers have additionally received support from Academy of Finland (project numbers #133818 and #218072, and National Centers of Excellence Program 2006-2011) and from the aivoAALTO research project of the Aalto University.

Finally, we thank Kranti Kumar Nallamothu for preparing the stimuli for the experiment.

Bibliography

[1] J-D. Haynes and G. Rees. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, 7:523–534, 2006.

[2] Y. Kamitani and F. Tong. Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685, 2005.

[3] E. Formisano, F. De Martino, M. Bonte, and R. Goebel. "Who" is saying "What"? Brain-based decoding of human voice and speech. *Science*, 322(5903):970–973, 2008.

[4] K.N. Kay, T. Naselaris, R.J. Prenger, and J.L. Gallant. Identifying natural images from human brain activity. *Nature letters*, 452(20):352–356, 2008.

[5] T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nounds. *Science*, 320(1191), 2008.

[6] Y. Miyawaki, H. Uchida, O. Yamashita, M. Sato, Y. Morito, H.C. Tanabe, N. Sadato, and Y. Kamitani. Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. *Neuron*, 60:915–929, 2008.

[7] T.Naselaris, R.J. Prenger, K.N. Kay, M. Oliver, and J.L. Gallant. Bayesian reconstruction of natural images from human brain activity. *Neuron*, 63:902–915, 2009.

[8] J.W. Rieger, C. Reichert, K.R. Gegenfurtner, T. Noesselt, C. Braun, H-J. Heinze, R. Kruse, and H. Hinrichs. Predicting the recognition of natural scenes from single trial MEG recordings of brain activity. *Neuroimage*, 42:1056–1068, 2008.

[9] S. Waldert, H. Preissl, E. Demandt, C. Braun, N. Birbaumer, A. Aertsen, and C. Mehring. Hand movement direction decoded from MEG and EEG. *Journal of Neuroscience*, 28(4):1000–1008, 2008.

[10] A. Toda, H. Imamizu, M. Kawato, and M.A. Sato. Reconstruction of two-dimensional movement trajectories from selected magnetoencephalography cortical currents by combined sparse Bayesian methods. *NeuroImage*, 54(2):892–905, 2011.

[11] K. Friston, C. Chu, J. Mourao-Miranda, O. Hulme, G. Rees, W. Penny, and John Ashburner. Bayesian decoding of brain images. *NeuroImage*, 39(1):181–205, 2008.

[12] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1410–1418, 2009.

[13] H. Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90(2):227–244, 2000.

[14] H. Daume III and D. Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126, 2006.

[15] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society B*, 58(1):267–288, 1996.

[16] R. Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, 1997.

[17] S. Taulu and J. Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in Medicine and Biology*, 51(7):1759–1768, 2005.

[18] A. Klami and S. Kaski. Local dependent components. In *Proceedings of ICML 2007, the 24th International Conference on Machine Learning*, pages 425–432, 2007.

[19] Y. Fujiwara, Y. Miyawaki, and Y. Kamitani. Estimating image bases for visual image reconstruction from human brain activity. In *Advances in Neural Information Processing Systems 22*, pages 576–584, 2009.

[20] S. Virtanen, A. Klami, and S. Kaski. Bayesian CCA via group sparsity. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

[21] M. Alamgir, M. Grosse-Wentrup, and Y. Altun. Multitask learning for brain-computer interfaces. In *Proceedings of 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 17–24, 2010.