Proceedings of the IASTED International Conference
HUMAN-COMPUTER INTERACTION
November 14-16, 2005, Phoenix, AZ, USA

USABILITY TESTING BASED ON THE COGNITIVE MISMATCH EEG SIGNALS

Koji Morikawa Shinobu Adachi
Advanced Technology Research Laboratories
Matsushita Electric Industrial Co., Ltd. (Panasonic)
3-4 Hikaridai, Seika-cho, Kyoto 619-0237, JAPAN
{morikawa.koji, adachi.shinobu}@jp.panasonic.com

ABSTRACT
This paper describes how electroencephalogram (EEG) signals can be utilized in usability testing, especially for evaluating the cognitive mismatch between a user's mental model and an appliance's working model. We aim to evaluate the usability of information appliances that interact with users. In addition to traditional usability testing, it is necessary to consider the user's mental state, that is, how users understand and predict an appliance's response. Two serial psychophysiological experiments were conducted to confirm whether the EEG signals obtained from a user reflect the cognitive mismatch that occurs when a user's expectations conflict with the actual response. We focused on the event-related brain potential (ERP) components measured by EEG, and we identified a positive ERP component with a peak latency of 550-600ms as an indicator of cognitive mismatch. These results suggest that the ERP component might reflect a user's mental status. By combining signals with behavioural observations, we can realize detailed evaluations that will improve product design.

KEY WORDS
Usability Testing, EEG, Cognitive Mismatch, Mental Model

1. Introduction

This paper describes a method for improving the usability testing through the use of electroencephalogram (EEG) signals, specifically the evaluation of cognitive mismatch between a user's mental model and an appliance's working model. Usability testing is applied to evaluate prototypes of a new product, to compare the ease of operations between appliances, and to quest for new improvements. In a sense, traditional usability testing methods [1, 2] are established from a methodological point of view, observation of user operations is a fundamental method, and the use of questionnaires and interviews are additional tools for understanding observational results. Nielsen and Molich [3] utilized heuristics for effective evaluations based on many experiences.

Wide varieties of products, from simple tools to computer GUIs (Graphical User Interfaces) are evaluated with usability testing. Since computers have great adaptability and program flexibility, product designers can create designs with almost any interface structure. This flexibility is the primary cause of the wide variety of mismatches between the product designer and the user. There is often a big difference between the working model of a product that reflects the intentions of the product designer and that of an actual user. The detecting such mismatches is particularly important when evaluating computer-based products, and there are few methods that evaluate users' actual mental models.

One possible solution is to measure brain activity using measuring instruments. EEG is one candidate that can explore the mental activities of a user almost directly. Some researchers in the field of Psychophysiology use EEG signals for human-computer interfaces, usability testing, etc. [4]. For example, Cremades et al. [5] used EEG signals for a human-computer interface and used frequency analysis to extract information about mental activity. Schalk et al. [6] used event-related potentials evoked just after a system error to improve the system's detection ability of a user's demands. This research was focused on EEG-based communication for those with severe motor disabilities. Conventional EEG signals are used with frequency analysis to investigate emotional states, sleeping states, and mental disorders, but frequency analysis is not suitable for detecting a short term response in HCI. There have been a few research studies on usability testing using EEG signals to detect cognitive mismatches during human-computer interactions.

The purpose of this study is to identify EEG signals that indicate a cognitive mismatch between a user's mental model and a system's working model under HCI conditions, and to show how these mismatch signals can be utilized for usability testing. We focus on event-related brain potentials (ERPs) since mismatch feelings occur just after response from a product. With this signals, we can evaluate whether or not a user is operating a product with a proper mental model. These investigations enable us to gain a deeper understanding of the behaviour of a user, and more detailed evaluations help us to properly modify

products. We designed two serial experiments to confirm whether a cognitive mismatch signal can be separated from the usual matching conditions. The results of the experiments showed that the signal differences are clear enough to detect a mismatch when the positive ERP component has a peak latency of 550-600ms.

This paper consists of following chapters: chapter 2 describes the experimental design, results, and an interpretation of the results. We confirm the shape and characteristics of the ERP signals for the cognitive mismatch conditions. In the chapter 3, some hypothetical evaluations of the ERP signals are given. The combination of the ERPs and the behavioural results can reveal detailed evaluations of both specific functions and the whole appliance. In the chapter 4, we discuss the advantages and limitations of this method and mention other possible applications.

2. EEG experiments

Two experiments using EEG measurement are explained. The purpose of the experiments is to investigate the mismatch between a user's expectations and a system's response. We assume that a user operating a system must have some expectation of response feedback and that the user would experience a certain level of mental discomfort if the system responds differently to how the user expected.

In general, EEG signals can be analyzed by two main approaches: one is through frequency analysis and the other is using event-related potentials (ERPs). The former supposes that frequencies contain relevant information. The latter supposes that specific periods of time contain the relevant information. For example, brain waves evoked just after a specific stimulus (event) reflect the user's response to the stimulus, and the interpretations are assigned to each positive and negative peak with their latency as reflections of the cognitive processes [7].

Usually event-related brain potential (ERP) is measured under a passive stimulus. For example, in an auditory odd-ball task, participants have to listen without any action to a series of simple constant tones with occasional variant tones. For HCI research, user actions for retrieving a system response must be included. Nittono et al. [8, 9] proposes the mouse click paradigm; in it stimuli are elicited through the subject's action -- a mouse click. The results showed that the rare auditory stimuli elicited a positive peak around 300 ms, and comparing to passive stimuli, the amplitude of a positive peak was larger when the stimuli were triggered by the subject's mouse clicks.

In addition to the mouse click, we added an action selection step in our experiments. The action selection step is crucial for the human-computer interaction and predictions of feedback are essential for deciding which action may lead to expected feedback.

2.1 Methods

As explained above, both experiments included the action selection step. The difference between the two experiments is the feedback stimulus; experiment I uses a symbol as the answer to a selected action, while experiment II uses letters representing the selected action. These differences are prepared to confirm whether the measured ERPs are actually from the mental states properly or from the type of feedback response.

Experiment I

Participants were four normal right-handed volunteers (three men, one woman, 26-38 years old, M = 33.3 years). They provided written informed consent to participate. EEG was recorded from two midline scalp electrode sites (Cz and Pz according to the 10-20 system [10]) referenced to the nose tip. The bandpass filter was set at 0.03-20Hz and the sampling rate was 1000Hz. An LCD monitor was placed 1.0m in front of the participant. The epoch for averaging was defined from 100ms before to 900ms after mouse clicking. Baseline corrections were made by subtracting the mean amplitude of the 100ms pre-stimulus period from the amplitude at every time point along the averaged waveform. Trials in which the EEG or EOG exceeded 100μV were rejected from the ERP averaging.

Each trial consisted of three steps as shown in Figure 1: (1) Visual indication. An "L" or "R" alphabetic letter was randomly displayed (duration = 200ms); (2) Action selection. Participants selected the left or right mouse button according to the preceding visual indication and then clicked; (3) Visual feedback. Either an "O" symbolizing a correct action, or an "X" symbolizing an incorrect action was displayed (duration = 100ms). In 80% of the trials that a participant operated correctly, the "O" symbol (indicating a correct answer) was displayed, while, in the remaining 20%, the "X" symbol (indicating an incorrect answer) was displayed as unexpected visual feedback. ERPs were recorded for 130 trials, and a 20% mismatch condition was mixed during through the 31st to 130th trial for preparing the initial familiarization phase (the first 30 trials) and the test phase (the following 100 trials).

Experiment II

Participants were eight normal right-handed student volunteers (eight women, 21-23 years old, M = 21.8 years). They gave written informed consent to participate. EEG was recorded from 2 midline scalp electrode sites (Cz and Pz according to the 10-20 system [10]) referenced to the nose tip. The bandpass filter was set at 0.15-20Hz and the sampling rate was 200Hz. Other measuring conditions were the same as experiment I.

Each trial consisted of three steps as shown in Figure 2: (1) Visual indication. An "L" or "R" alphabetic letter was randomly displayed (duration = 200ms); (2) Action selection. Participants selected the left or right mouse button according to the preceding indication and then

clicked; (3) Visual feedback. A “Left” or “Right” letter string was displayed (duration = 100ms) based on the clicked mouse button. In 80% of the trials that a participant acted correctly, the correct string was displayed (i.e. that corresponded with the button they clicked), while, in the remaining 20%, the other one was displayed as unexpected visual feedback. This enabled the user to react to the same amount of two stimuli; as the probabilities of “Left” or “Right” being displayed were equal. ERPs were recorded for 130 trials, and a 20% mismatch condition was mixed during from 31st to 130th trial, for preparing the initial familiarization phase (the first 30 trials) and the test phase (the following 100 trials).

In these experiments, participants expect the system’s response based on the mental model they constructed after receiving instructions from the experimenter. However, the system sometimes reacts differently to the given instructions. This corresponds to an artificial system-side error; the other possible mismatch is due to human-side error. We suppose that the cognitive mismatch signal is elicited in both conditions.

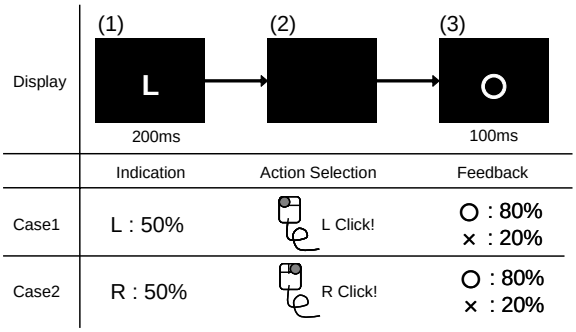

Fig. 1 Conditions in experiment I

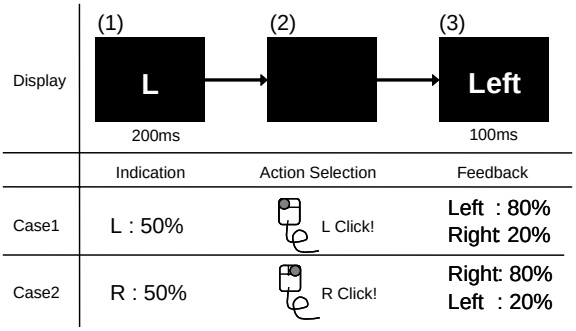

Fig. 2 Conditions in experiment II

2.2 Results

The results of experiment I are shown in Figure 3. The thin line designates the average waveform of correct feedback (i.e. no mismatch occurred) and the thick line designates the average waveform of incorrect feedback (i.e. a mismatch occurred). The horizontal axis shows the time of the feedback (ms), and the vertical axis shows the amplitude of the electrode attached on the position of Pz (μV). From figure 3, the cognitive mismatch correlated the positive peak with a latency of around 550ms. The results of experiment II are shown in Figure 4. Figure 4 also shows the positive peak with a latency of around 590ms in the case of a mismatch.

Figure 5 shows individual difference of the brain waves from four participants in experiment I. Although each plot shows different wave shapes, they have the same characteristics; the ERPs in the incorrect feedback condition have a positive peak of around 550-600ms when compared to the correct feedback condition.

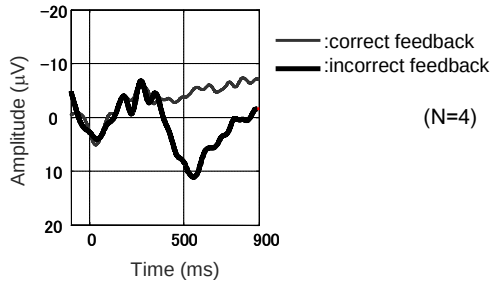

Fig. 3 Grand mean ERP waveforms in experiment I

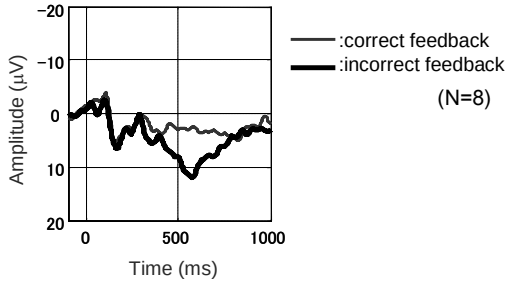

Fig. 4 Grand mean ERP waveforms in experiment II

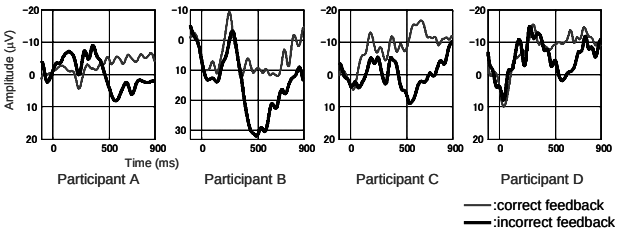

Fig. 5 Individual difference in experiment I

In both of the experiment I and II, similar signals were observed in the positive ERP components with a peak latency of 550-600ms in the mismatch condition. These results suggest that the signals reflect incorrect feedback and that they do not directly reflect the different types of stimuli, i.e. the correct or incorrect answer symbol, or the letter the participant selected. In experiment II, each visual feedback (“Left” or “Right”) was displayed with equal probability, suggesting that this ERP component may not reflect the difference in the probability of visual stimuli, but rather that it reflects the mental process, in contrast to the oddball task. Therefore, we assume that the incorrect feedback is closely related to the mismatch between expectation and actual response.

There are researches on the interpretation of brain waveforms in relation to late positive components [11], as most related components could be feedback error-related negativity (FB-ERN) [12]. It appears our results don't contain FB-ERN. We suppose the reason is the difference in the experimental conditions. Although the related research focuses on the ERPs elicited by the failure to predict a result, they use types of tasks that contain uncertainties. For example, they often use gambling or time-counting tasks, in which it is difficult to predict the exact system-side response. On the contrary, in our experiment and in HCI conditions, most of the system's responses are clear and sometimes a mismatch might occur that breaks the user's confident prediction. We think these differences could be the cause of the different brain waveforms.

3. Application for usability testing

In this chapter, we describe the way in which the cognitive mismatch signals measured in chapter 2 can be utilized for usability testing. Cognitive mismatch signals can be used as additional information in conventional usability tests. We will show some examples of this integration using observational results, which are fundamental information of the usability test.

Table 1 shows the interpretation of an integrated evaluation with an observational test. The table consists of two components. One shows the detection of a cognitive mismatch signal and the other represents the presence of a correct operation. In the table 1, there are four combinations of integrated evaluations.

Cell (A) displays a case of incorrect expectation and correct operation. The user may not have had prior knowledge of the operation but could perform the operation successfully. In this case, the user uses guess work to operate. This is appropriate for product evaluation, because users often don't have much prior knowledge of a new product's with new function. A product designer can increase the chance of a user operating a device correctly through a good design.

Cell (B) shows a case of correct expectation and correct operation. The user knows how the function works properly.

Cell (C) is a case of incorrect expectation and incorrect operation. The user in this cell had an incorrect expectation with confidence and failed to correctly operate the product, therefore, a mismatch occurred between the user's mental model and the product's working model. This case is the most noteworthy, because the product might have given the wrong impression about how it operated.

Cell (D) shows a case of no expectation and incorrect operation. In this cell, a product designer must think about giving more information to the user regarding what functions the product can achieve and must assist in the user's construction of a correct expectation or a mental model.

In conventional tests, the experimenter can only know whether or not the user operates the function correctly. If the experimenter combines this knowledge with EEG signals as we explained, more detailed information can be extracted, and this information can be exploited in order to improve products.

Table 1 Integrated evaluation with an observational test

		Cognitive Mismatch? (ERP measurement)	
		Yes	No
Correct Operation? (Observational test)	Yes	(A) Operated without prior knowledge	(B) Operated properly
	No	(C) Operated with incorrect expectation	(D) Operated randomly

Table 2 shows virtual examples of the evaluations of functions within one appliance. Information appliances such as TVs and DVD players have many functions and an evaluation has to be done for every function. In the example of table 2(a), there are four functions, and the number in each cell shows the percentage calculated from the usability test for some user groups. Table 2(b) also shows an evaluation table without cognitive mismatch signals, namely, the conventional evaluation. The number in each cell is the percentage, corresponding to the table 2(a), which is calculated if the same function is evaluated without EEG signals. For example, in function 1 of the conventional evaluation, 80% of the users operated correctly and 20% of the user operated incorrectly. In the integrated evaluation, 80% of users could be separated into two groups – the 70% of users who produced no mismatch signals and the 10% of users who produced mismatch signals.

Table 2 Example of the integrated evaluation

(a) Integrated evaluation

	Correct Operation (%)		Incorrect Operation (%)		Evaluation
	Match	Mismatch	Match	Mismatch	
Function 1	70	10	10	10	Good
Function 2	10	70	10	10	Good Affordance
Function 3	10	10	70	10	Incomprehension (mo mental model)
Function 4	10	10	10	70	Wrong mental model

(b) Conventional evaluation

	Correct Operation (%)	Incorrect Operation (%)	Evaluation
Function 1	80	20	Good
Function 2	80	20	Good
Function 3	20	80	Not Good
Function 4	20	80	Not Good

In the conventional evaluation (Table 2(b)), the results of the evaluation of function 1 and function 2 show the same percentage of correct operations, that is, both are evaluated as good. On the contrary, in the integrated evaluation, the evaluations are not the same, because

function 1 is operated correctly with correct expectation in most of cases. Function 2 is used correctly with incorrect expectation. Therefore, function 1 should be evaluated as "good", and the function 2 should be evaluated as "good due to appearance design", which means user could operate it correctly without having a correct expectation and that the user chose the correct operation thanks to the appearance or the affordance of the appliances.

Similarly, function 3 and function 4 have the same evaluation in the conventional evaluation (table 2(b)), but have different evaluations by integrated evaluation. Function 3 should be evaluated as the user failing to have any expectations, which means it is necessary to notify the functional ability of function 3 and to assist to construct the correct mental model. Function 4 should be evaluated as the user having incorrect expectations -- function 4 attracted misleading operations. Consequently, both function 3 and function 4 have the same evaluation "modification needed", but the different aspects must be modified.

In this way, the characteristics of this integrated evaluation using cognitive mismatch signals offer detailed evaluations of the user's mental model. We can't know the user's exact mental model with this evaluation method, however, we can perceive the difference in specific functions that exist between a user's mental model and a system's working model. Equally, the grand average of the waveform of users and functions can be utilized in the design of appliances. We can judge which appliance is more misleading, which one has easier an operational structure without prior knowledge, and which appliance needs more detailed explanations about its functions. Furthermore, the average waveform of each user can be used in evaluating users. We can judge who has an incorrect mental model, who can operate with a perfect mental model, and who has no mental model of an appliance.

Users often encounter mental model mismatches when using a system or an appliance. A user may operate a Mac computer with the mental model of a Windows PC, or operate a DVD recorder with the mental model of a VHS VCR, etc. We guess users often produce cognitive mismatch EEG signals, which can be characterized as having a positive peak at around 500-600ms.

4. Discussion

In this chapter, we discuss the advances, limitations, and the necessary investigations of our usability testing methods.

The merit of this usability testing method is that it can add extra information that cannot be detected through the conventional method. The information matches the user's model and the system's model. Usually, information about a user's mental state is extracted from questionnaires during or after the usability test. These questionnaires are a useful and convenient method, but they require time and labor on the part of the user. The most crucial defect in the questionnaire method is that it interrupts the user's natural continuous cognitive processes if it is done during a test, and it relies on the user's unreliable memory if performed after a test. EEG measurements don't interrupt the user's cognitive processes during the usability testing.

On the other hand there are some difficulties to be solved, and we now discuss two of them. The first one is weak brain wave intensity. Since the amplitude of ERP signal is so small, it is necessary to average many waveforms under the same conditions, usually between 10-20 times. If there are difficulties in performing repeat evaluations of a specific function by a single user, there must be several participants.

The averaging also prevents the immediate detection of mismatch signals just after the feedback. If the signal could be detected without the addition of the waveforms, more applications could be realized. For example, immediate evaluation can affect the next evaluation function, such as adaptive evaluation. This adaptive evaluation is more effective in computer-aided education. In particular, a correct answer without correct expectation could be modified at the time it occurs. Some researchers are aiming to detect ERP signals from single trials using independent component analysis [13], Hidden Markov Models (HMMs) [14], wavelet denoising [15], etc.

The second difficulty is the attaching the electrodes on the head. It requires time and a certain level of skill. For example, we used seven electrodes in our experiments, and this required more than 20 minutes preparation time. However, this is an easier and simpler preparation compared to other methods for measuring brain activities such as fMRI (functional Magnetic Resonance Imaging) and MEG (Magnetoencephalography).

There are other issues to be investigated. One is whether these ERP signals can be observed in complex stimuli. In our experiments, we used simple symbols or letters such as "X" or "Right". However, the responses of the products we would like to evaluate are not only simple ones but also more complex ones. For example, in the case of evaluating a TV, the response to the user may be motion pictures on the screen, or unfamiliar technical words, even in a motion picture. We have to investigate the characteristics of the ERP signals in these kinds of complex stimuli.

The clarification of the relationship between ERP signals and a user's mental model is also on the list for investigation. In cognitive science, one of the primary research fields includes user modeling during interacting with computers. Norman [16] defined the user-centered system design, which handles simple model for user's

operating appliances. His model claims to fill the gap between the mental world of the user and the physical world of the system. The ERP signals in this paper might reflect the mismatch at such a gap.

Card, Moran and Newell proposed the GOMS model [17], which consists of "Goals", "Operators", "Methods" and "Selection rules". In their model, cognitive mismatch signals may correspond to the incorrect selection of "Selection rules", or the incorrect choice of "Methods". Currently, even HCI professionals make own user models based on these approaches [18], therefore, it can be said that there is no definitive user model. Still, we believe that the cognitive mismatch signal represents a kind of mismatch between a user's mental model and a system's model, and even at this level of interpretation, these signals can be used to improve usability testing by evaluating new aspects that conventional method cannot understand.

5. Conclusion

This paper describes how EEG signals can be utilized in usability testing, especially in the detection of a cognitive mismatch between a user's mental model and an appliance's working model. The results of the two experiments showed the possibility of measuring the cognitive mismatch in a human-computer interactive condition. The results show that the observational data of user operations can be divided into two parts, showing which operations contain cognitive mismatch signals, and which do not. If the cognitive mismatch signal is present, we can find evidence of a problem even if a user can operate an appliance successfully.

Some limitations were also discussed. The EEG signals are slight, and must be averaged out in order to judge whether or not the supposed signal exists. If we had a method that did not requiring averaging, we could use this testing method under real time conditions, which would mean a system could provide more timely information to assist the construction and modification of a user's mental model. Another restriction is that users must have EEG electrodes attached to their heads. Clearly, this is not a usual condition for users when watching a TV or DVD at home, however, it is allowable in the evaluation phase at the usability test lab. Nevertheless findings from these tests can be used in improving product design from the aspects of user's mental models.

We are currently conducting the usability testing with consumer products, not merely with the simple mouse and the simple display. In future, we will improve this method so that it can be applied to a more realistic environment.

Acknowledgements

The authors wish to acknowledge H. Nittono for helpful comments about the interpretation of the ERP signals and S. Araki for discussions about experimental design and for supporting this research.

References

[1] J. Nielsen, Usability Engineering, Morgan Kaufmann, 1994
[2] C. D. Wickens, J. G. Hollands, Engineering psychology and human performance (3rd ed.). Upper Saddle River, NJ, Prentice-Hall, 2000
[3] J. Nielsen, R. Molich, Heuristic evaluation of user interfaces, Proceeding of CHI'90, 1990, 249-256
[4] E. Donchin, A. F. Kramer, C. D. Wickens, Applications of brain event-related potentials to problems in engineering psychology. In M. G. H. Coles, E. Donchin & S. W. Porges (Eds.) Psychophysiology: Systems, processes, and applications. New York, Guilford Press. 1986, 702-718
[5] J. G. Cremades, A. Barreto, D. Sanchez, M. Adjouadi, Human–computer interfaces with regional lower and upper alpha frequencies as on-line indexes of mental activity, Computers in Human Behavior, 20, 2004, 569-579
[6] G. Schalk, J. R. Wolpaw, D. J. McFarland, G. Pfurtscheller, EEG-based communication: presence of an error potential, Clinical Neurophysiology, 111, 2000, 2138-2144
[7] T. C. Handy, Event-Related Potentials: A Methods Handbook, The MIT Press, 2004
[8] H. Nittono, P. Ullsperger, Event-related potentials in a self-paced novelty oddball task, Neuroreport, 11(9), 2000, 1861-1864
[9] H. Nittono, A. Hamada, T. Hori, Brain potentials after clicking a mouse: a new psychophysiological approach to human-computer interaction, Human Factors, 45(4), 2003, 591-599
[10] H.H. Jasper, The ten twenty electrode system of the international federation, Electroencephalography and Clinical Neurophysiology, 10, 1958, 371-375
[11] J. Dien, K.M. Spencer, and E. Donchin, Parsing the late positive complex: mental chronometry and the ERP components that inhabit the neighborhood of the P300," Psychophysiology, vol. 41, no. 5, pp. 665-78, September 2004.
[12] M. Rushsow, J. Grothe, M. Spitzer, M. Kiefer, Human anterior cingulate cortex is activated by negative feedback: evidence from event-related potentials in a guessing task, Neuroscience Letters, 325, 2002, 203-206
[13] A. Delorme, S. Makeig, EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis, Journal of Neuroscience Methods, 134, 2004, 9-21
[14] B. Obermaier, C. Guger, C. Neuper, G. Pfurtscheller, Hidden Markov models for online classification of single trial EEG data, Pattern Recognition Letters, 22, 2001, 1299-1309
[15] R. Q. Quiroga, H. Garcia, Single-trial event-related potentials with wavelet denoising, Clinical Neurophysiology, 114, 2003, 376-390
[16] D. A. Norman, Cognitive engineering. In D. A. Norman, & S. W. Draper (Eds.), User Centered System Dsign, Hillsdale, NJ: Lawrence Erlbaum, 1986, 31-61
[17] S. K. Card, T. Moran, A. Newell, The psychology of human-computer interaction, Hillsdale, NJ: Lawrence Erlbaum Associates, 1983
[18] T. Clemmensen, Four approaches to user modeling - a qualitative research interview study of HCI professionals' practice. Interacting with Computers, 16, 2004, 799-829