Modality Effects on Cognitive Load and Performance in High-Load Information Presentation

Yujia Cao
Human Media Interaction
University of Twente
P.O. Box 217, 7500AE,
Enschede, The Netherlands
y.cao@utwente.nl

Mariët Theune
Human Media Interaction
University of Twente
P.O. Box 217, 7500AE,
Enschede, The Netherlands
m.theune@utwente.nl

Anton Nijholt
Human Media Interaction
University of Twente
P.O. Box 217, 7500AE,
Enschede, The Netherlands
a.nijholt@utwente.nl

ABSTRACT

In this study, we argue that modality planning in multimodal presentation systems needs to consider the modality characteristics at not only the presentational level but also the cognitive level, especially in a situation where the information load is high and the user task is time-critical. As a first step towards automatic cognitive-aware modality planning, we integrated the effect of different modalities on cognitive load and performance, using a high-load information presentation scenario. Mainly based on modality-related psychology theories, we selected five modality conditions (text, image, text+image, text+speech, and text+sound) and made hypotheses about their effects on cognitive load. Modality effects were evaluated by two cognitive load measurements and two performance measurements. Results confirmed most of the predicted modality effects, and showed that these effects become significant when the information load and the task demand are high. The findings of this study suggest that it is highly necessary to encode modality-related principles of human cognition into the modality planning procedure for systems that support high-load human-computer interaction.

Author Keywords

modality effect, high-load information presentation, cognitive load, performance, heart rate variability

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces

INTRODUCTION

Intelligent human-computer interfaces are often multimodal, i.e. the human-computer communications are accomplished through multiple modalities, such as text, images, speech, sound etc. Following Bernsen's definitions, the term modality refers to "mode or way of exchanging information between humans and machines in some medium"([7] p. 95), and medium is the "physical realization of information at the interface between human and system"([7] p. 94). In order to generate multimodal presentations on the fly, intelligent user interfaces often require automatic modality planning, that is to automatically select and combine several available modalities for generating a multimodal message according to a given presentation goal. In existing multimodal presentation systems, modality planning is often guided by a set of predefined allocation rules embedded in structured presentation plans [2,13,22,23]. These rules make mappings from the domain of information to be presented to the domain of available modalities. For example, prefer graphics for concrete information (such as shape, color and texture) and prefer text for quantitative information (such as most, some, any, exactly, and so on) [1]; use graphics alone for location and physical attributes and text alone for abstract actions and relationships among actions [13]; prefer auditory modalities if the user is blind; prefer visual modalities if the environment is very noisy; prefer auditory modalities if the user's task requires continual body movement [9]. Essentially, the design of modality allocation rules is influenced by the following factors (based on [1]): 1) the presentation goal, 2) characteristics of the information to be conveyed, 3) characteristics of the modalities, 4) user profile and environmental profile, 5) the task to be performed by the user, and 6) resource limitation.

Generally, previous modality planning studies show a lack of attention on the cognitive characteristics of modalities, i.e. how information carried by different modalities is perceived and processed by the human perceptual-sensory system. As multimodal presentations are generated for human users to perceive, process and act upon, computer systems should understand not only how to convey information, but also how human minds are going to take in and analyze the information. Based on this knowledge, automatic modality planning can be conducted in a cognitive-aware manner, which means that the human cognitive resources are efficiently used while a user is perceiving and processing the multimodal messages. It has also been argued in [12] that human-computer interaction (interfacing) should be understood in the light of a general principle of human cognition.

Our interest in this research is high-load information presentation, a situation in which a large amount of diverse information needs to be presented in a limited amount of time.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
IUI 2009, February 8-11, 2009, Sanibel Island, Florida.
Copyright 2009 ACM 978-1-60558-331-0/09/02...$ 5.00.

The user task is also time-critical, which means users have a limited amount of time to perceive, analyze and react to the information that is presented to them. A typical example of this situation is real-time information presentation in a crisis response center. Facing a large display, crisis managers need to make fast decisions based on information that is continuously coming in (see [20] for an example). The full capacity of human cognitive resources is being challenged. In such applications, we believe that the cognitive characteristics of modalities become particularly important, because the amount of cognitive load that is imposed to the user might differ depending on which modalities are used, and the differences are enlarged when the information load is high. Therefore, modality-related principles of human cognition need to be encoded into the automatic modality planning procedure.

Cognitive psychologists have long been struggling to understand how the human mind perceives and processes information. Several modality-related theories and principles are empirically well-founded and have been successfully applied to explain educational phenomena and develop multimedia learning materials [10, 11, 17]. In this study, we intend to apply these theories and principles as a theoretical basis of automatic cognitive-aware modality planning. We designed an experiment to test the effect of different modalities on cognitive load and performance, using a high-load information presentation scenario. Mainly based on relevant cognitive theories, we selected five modality conditions and made hypotheses. If the predicted modality effects are confirmed, then we can conclude 1) modality planning in high-load information presentation systems needs to consider not only the representational properties, but also the cognitive properties and combinations of modalities, and 2) the applied theories and principles can be used to predict and manipulate a user's cognitive load in a cognitive-aware user interface.

RELATED WORK

In order to address modality planning tasks in a unified and systematized manner, modality taxonomies were proposed to serve as theoretical foundations for understanding and describing modalities. They essentially support the automatic generation of multimodal presentations. Bernsen [6] proposed a modality taxonomy, based on the observation that different modalities have different representational power. This taxonomy identifies modalities by a set of properties, i.e. linguistic/non-linguistic, analogue/non-analogue, arbitrary / non-arbitrary, static/dynamic, and visual/auditory/tactile. It is claimed that this modality taxonomy is complete and unique [7]. Bachvarova [3] argued that a modality should be described by both the content it represents and its nature. Moreover, the nature of a modality should be looked at from several perspectives. Therefore, she reorganized the properties in Bernsen's modality theory and extended them into three levels: the information presentation level, the perception level and the structural level. The information presentation level describes the capability of a modality to represent certain types of information. The properties 'linguistic/non-linguistic' and 'analogue/non-analogue' belong to this level. The perception level determines how a modality is perceived and processed by the human perceptual-sensory system. This level distinguishes among being visual, auditory, haptic, olfactory, or gustatory. Static/dynamic is also a property at this level, because it determines how much time a modality allows to be perceived and processed. The structural level models the dependencies that can exist between composite modalities. The arbitrary/non-arbitrary property falls into this level.

At each level of this modality ontology, we can systematically identify modalities, describe their abilities, and (most importantly) derive rules for automatic modality planning. For instance, at the presentation level: linguistic modalities (e.g. text, discourse) surpass analogue modalities (e.g. images, graphics, diagrams) at explaining abstract concepts; while analogue modalities are better at expressing what things exactly look like [7]; at the perception level: use an additional auditory modality if using only a visual modality can cause overload in the visual perception channel [16]; and at the structural level: the combination of an icon and a map is able to describe both an object and its location [3].

Modality planning in the AI domain considers mostly the representational properties at the presentation level, while many modality-related psychology studies focus on the perceptional and cognitive aspects. In the multimedia learning domain, educational psychologists have spent much effort on investigating how to design multimedia learning materials that bring cognitive load benefits for the learner [10, 11, 17]. Their studies are mainly based on two theoretical foundations [8, 17]: dual-channel theory and dual-coding theory.

The dual-channel theory is derived from Baddeley's working memory model [4]. The model suggests that working memory has separated stores for visual information and auditory information (these can be understood as two separated perception channels), and each memory store has limited capacity. Therefore, in order to make better use of the working memory capacity, it is preferable to let the two perception channels share the perception task in a coordinated manner. Based on this theory, it has been shown that students learn better when their learning material combines animation and speech than when it combines animation and on-screen text [17]. This is because physically separated animation and on-screen text split a student's visual attention, and thus bring high load to the visual perception channel. When the perception of linguistic information is carried by the auditory channel, visual load is reduced and cognitive capacity is more efficiently used. This phenomenon is known as a split-attention effect.

The dual-coding theory from Paivio [18] states that mental processes and dynamic associative processes operate on a rich network of modality-specific verbal and nonverbal representations. This indicates that verbal and nonverbal materials are processed and mentally represented in separate but interconnected systems. Verbal materials contain visual, auditory, and other linguistic codes. Nonverbal materials include images, environmental sounds, actions, and other

non-linguistic objects and events.Therefore, the terminology 'verbal/nonverbal' is consistent with 'linguistic/analogue' in Bernsen's modality taxonomy. Multimedia learning studies have demonstrated that the associative processes between verbal and nonverbal systems play major roles in knowledge comprehension and memorization. Therefore, educational materials which contain associated verbal and nonverbal codes normally lead to better performance in comprehension, learning and memorization [11].

In this study, we took the dual-channel theory and the dual-coding theory as two theoretical assumptions, and applied them in our context and experimental setting. As our high-load information presentation scenario (discussed in the next section) does not require comprehension and long-term memorization, the same theoretical assumptions might yield different suggestions on the usage of modality (as will be explained below).

SCENARIO

We designed a crisis rescue scenario to simulate a high-load information presentation. After a massive earthquake, a rescue team arrive at an affected inhabited area. Rescue workers search for injured people and transfer them to safe places. Using mobile communication devices, they report the location of wounded victims ('patients') and dead bodies ('deaths') to the crisis response center. A user plays the role of a crisis manager located in the crisis response center, monitoring the rescue progress through a display. Patient and death reports are presented on the display with the background of a grid map (see figure 4). The user task is to guide a doctor to save all patients.

In this scenario, information load and workload can be regulated by three parameters: 1) The user task is time-critical. Each patient has a life value that decreases with time, thus patients die without timely treatment. When the life values are generally lowered, the task allows less reaction time for saving each patient, and in turn requires the user to be more mentally engaged into it. 2) The information load can be regulated by the presentation intensity: the number of patients/deaths appearing on the screen per minute. Higher intensity brings higher information load. 3) The difficulty of the task is also controlled by the patient/death ratio. When this ratio is low (few patients and many deaths), it is relatively difficult to identify the location of patients. Our parameter settings aim at inducing high workload but not making the task so difficult that users feel too frustrated to keep trying their best. Design choices have been made based on a pilot study. With the current parameter settings, at least 80% of patients can be saved if the user is fully engaged. Moreover, this crisis rescue scenario is also realistic in the sense that it matches two characteristics of real crisis management situations [21]: time urgency and high information load.

MODALITY CONDITION DESIGN

Different modalities are used in different trials to present patients and deaths. We consider the four most common and feasible modalities, i.e. text, image, speech and sound. The set of all combinations contains 15 (2^4-1) possible modality usages (table 1). Based on representational concerns and cognitive concerns, a subset of 5 have been selected to be experimental conditions, as follows.

Index	Text	Image	Speech	Sound
1*	√			
2*		√		
3			√	
4				√
5*	√	√		
6*	√		√	
7*	√			√
8		√	√	
9		√		√
10			√	√
11	√	√	√	
12	√	√		√
13	√		√	√
14		√	√	√
15	√	√	√	√

Table 1. All possible modality usages based on the four uni-modalities. *: selected experimental conditions

First, we made a selection at the presentation level, based on the characteristics of the information and the representational properties of available modalities. A presentation unit (a report) contains an object type (patient or death) and an object location. All four modalities can easily present different types of objects. However, not all of them are suitable to convey the locations. When the area of interest is divided into many location units (e.g. a map contains many sub-zones), only using auditory modalities (speech or sound) without visual modalities is not effective to indicate a location. Speech can refer to a location by a row index and a column index, or a zone index. However, perceivers need to transfer this auditory information into a visual search task without explicit confirmation of the searching results. Sound can use its pitch or direction for location indication. However, it would be almost impossible to practise the pitch-location or direction-location mappings, especially when the possibilities are many (over 250 in this application). The conclusion is that at least one visual modality is necessary, and the choices 3,4, and 10 in table 1 are rejected. The situation is much easier when we have a visual modality and a map background. The text or image itself presents the object type and its location on the map indicates the location of the object. Between the two visual modalities, it is known that text, as a linguistic modality, is better at describing abstract concepts; while image, as an analogue modality, is better at describing concrete concepts [3]. As the object type in this scenario is a concrete concept, image is supposed to be more suitable than text. We selected both of them as experimental conditions, and expected to see the impact of representational characteristics on cognitive load and performance. Figure 1 illustrates the presentations used in the experiment. Text 'Patient' and 'Death' have the same font, size, and color. The two images also have the same size, color and similar shape. If the advantage of image over text can be confirmed when two similar images are used, it should be even

more notable when there is a large contrast in their colors, sizes, and shapes.

	Text	Image
Patient	Patient	
Death	Death	

Figure 1. Text and image presentations

Then, we made further selections at the perception level. Although one visual modality (text or image) is already sufficient to present a report, modality combinations might further bring cognitive benefits, i.e. imposing less cognitive load under the same task condition and allow better user performance. However, modality combinations can be beneficial only when they are used in a proper manner. First of all, if the combined modalities carry repeated information, the redundant information can unnecessarily cause extra cognitive load. This redundancy effect has been found in multimedia learning studies, showing that students understand a multimedia presentation better when words are presented as narration rather than as narration and on-screen text [17]. In our case, it is not wise to let each of the combined modalities carry exactly the same information (e.g. the combination of text 'patient' and speech 'a patient has been found'). Noticing that the main difficulty in the rescue task is to find out where the patients are, we provided extra information of patient locations with another modality. The screen is divided into two halves, the left half and the right half. Using a visual modality on the map as a basis, a second modality conveys which half of the screen contains a new-presented patient. These two modalities are temporally synchronized. In order to test the dual-channel theory, we need both visual-visual and visual-auditory combinations. In order to test the dual-coding theory, we need both verbal-verbal and verbal-nonverbal combinations. Between the two visual modalities (text and image), we chose to use text as the basis for combinations. As text was predicted to be less suitable than image, it would be better for showing the differences before and after being combined with another modality. Finally, three modality combinations were selected: text+image, text+speech, and text+sound. The text+image condition applies a left arrow or a right arrow with large size and striking color at the center bottom of the screen (the left arrow is illustrated in figure 2). The text+speech condition uses speech 'left' or 'right'. The text+sound condition plays an ambulance sound from the left or the right channel of a stereo speaker set.

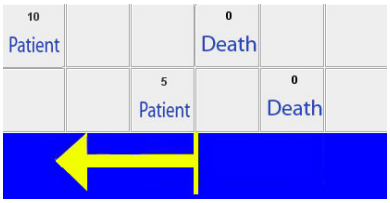

Figure 2. Example of the left arrow used in the text+image condition

According to the dual-channel assumption, a visual-visual combination in a high-load visual task might cause the split-attention effect and induce overload in the visual perception channel. Therefore, we predict that the visual-auditory combinations (text+speech and text+sound) impose less cognitive load than the visual-visual combination (text+image), thus also leading to better performance. Based on the dual-coding assumption, we predict that the verbal-verbal combination (text+speech) is superior to the verbal-nonverbal combinations (text+image and text+sound). This prediction seems to be against the suggestion from multimedia learning studies. This is due to the different characteristics between our user task and a learning task. A learning task requires information comprehension and long-term memorization, and the associative processes between verbal and nonverbal mental systems have been shown to be beneficial. However, our user task requires quick perception and reaction. A single word 'patient', 'death', 'left' or 'right' should be easy enough to be understood even without the associative efforts between the two mental systems. In this case, the associative processes might be unnecessary. Moreover, they use extra cognitive resources and this may slow down the user's reaction.

Moreover, we rejected all combinations with more than two modalities (choices 11 to 15 in table 1), because they are more likely to contain redundancy and induce cognitive overload, especially with a high presentation density. In summary, the five chosen modality usages were text, image, text-image, text-speech, and text-sound. Their impacts on cognitive load and performance were evaluated by one subjective measurement, two performance measurements and one physiological measurement (see the next section).

EXPERIMENTAL DESIGN

20 people participated in this experiment (15 men and 5 women). Their age ranges from 22 to 32. They are all university students (bachelor, master, or PhD) and daily computer users. With a within-subject design, each subject participated in all five modality conditions. The trial order is counter-balanced with Latin square size 5 [1].

Measurements

We used one subjective measurement and one physiological measurement to assess cognitive load. The NASA Task Load Index (NASA-TLX [14]) was used to obtain the subjective cognitive load level (SCL). TLX contains six workload-related factors: mental demands, physical demands, temporal demands, own performance, effort and frustration. A 20-level rating can be performed on each of the six factors. However, in this experiment, we only applied the 'mental demands' dimension to access the subjective cognitive load. The users were asked to report how much mental effort they needed to devote to the task under each modality condition. They could rate this from 1 (very low) to 20 (very high).

As a physiological measurement, we used heart rate variability (HRV), which is one of the most commonly used indexes

[1] A Latin square is an n × n table filled with n different symbols in such a way that each symbol occurs exactly once in each row and exactly once in each column.

of cognitive workload. Previous studies suggested that HRV decreases with increased cognitive load demands [19,24,25]. HRV can be accessed by spectral analysis of the beat-to-beat interval sequence in the frequency domain. Three frequency bands have been associated with different biological control mechanisms: the very low band (VLF, 0.02Hz $\sim$ 0.06Hz) is associated with body temperature regulation; the low band (LF, 0.07 Hz $\sim$ 0.14 Hz) is related to short-term arterial pressure regulation; and the high band (Hf, 0.15Hz $\sim$ 0.50Hz) reflects respiration activities. Previous studies have shown that there is a systematic and reliable relationship between the power in the LF band and mental demands [15]. Higher levels of mental workload have been associated with decreased power in the LF band. We used the LF power as a physiological measurement of cognitive workload.

Two more measurements were used to evaluate the user performance: reaction time (RT) and number of dead patients (ND). RT is the time interval between the presentation and the treatment of a patient (in seconds). ND is the number of patients who died because they didn't receive treatment in time.

Apparatus and setup

Two PCs were used in the setup. PC-1 hosted the crisis rescue interface. User performance was logged as text files on this computer. Electrocardiograms (ECG) were collected by three flat-type active-electrodes placed on the torso. Via an A/D converter and a USB receiver, ECG signals were fed into PC-2 and recorded. There was also a parallel-port connection between PC-1 and the USB receiver. The rescue program on PC-1 sent event triggers at the beginning and the end of each trial, in order to synchronize the ECG recording. Figure 3 demonstrates the experiment setup.

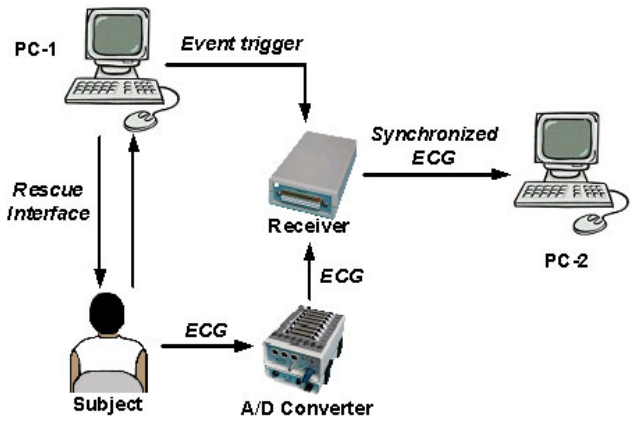

Figure 3. The experiment setup

User task

The map of the crisis scene was designed as a simple 20 $\times$ 13 array of cells in a grid (figure 4). A cell can contain at most one object at a time. The doctor is presented by an icon. The user task is to send the doctor to patients by mouse clicking on the presentations of patients (text 'Patient' or an image of a patient, see figure 1). After each mouse click, the doctor immediately moves ("jumps") to the patient's location and starts treating the patient (which takes a fixed interval of 1 second).

Figure 4. A screenshot of the rescue interface

Procedure

After entering the laboratory, a participant took a seat in front of PC-1. First, he/she listened to soothing music and looked at peaceful nature pictures while the electrodes were being placed. The music and pictures were supposed to make the participant relaxed. The resting period lasted for about 10 minutes. Then, the participant was asked to stay relaxed and try not to think about anything. A baseline ECG was recorded for a period of 5 minutes. Afterwards, an introduction plus training session was provided by a java program. Using combined narration and animation, the program introduced the rescue scenario, explained the user task, and presented the five modality conditions. In the training session, the participant practised the rescue task under all five conditions (1 minute per condition). Then, the participant went through 5 trials with different modality conditions, following a specified order. The length of each trial was 5 minutes. A 3 $\sim$ 5 minutes break was placed between each two trials. The participant was asked to fill in the cognitive workload questionnaire and have a rest during breaks. The whole experiment lasted for approximately 70 minutes.

Hypotheses

We constructed the following four hypotheses.

1. Based on the representational properties, the image condition is better than the text condition. By 'better', we mean that a modality imposes less cognitive load and allows better performance.

2. Based on the dual-channel assumption, the text+speech condition and the text+sound condition are both better than the text+image condition.

3. Based on the dual-coding assumption, the text+speech condition is better than the text+sound condition.

4. Combined modalities are better than single modality, because they convey extra useful information.

RESULTS

Due to the within-subject design, we applied the repeated-measure ANOVA on the experiment data, where the independent factor was modality condition with five levels, and dependent variables were SCL (subjective cognitive load), RT (reaction time), ND (number of dead patients), and HRV (heart rate variability). Trial order was treated as an extra between-subject factor. The ANOVA results showed that trial order did not have a significant effect on any of the four dependent variables ($p = 0.14, 0.64, 0.28, 0.68$ for SCL, RT, ND, and HRV, respectively). This indicated that there was no significant training effect, thus the trial order factor was ignored in further data analysis. Results from statistical analysis are shown below and discussions are given in the next section.

SCL

In the five trials, the mean subjective cognitive load levels mostly fall in the higher half ($10 \sim 20$) of the 20-level rating scale, indicating that the rescue task generally has a high mental workload demand. Figure 5 shows the mean cognitive workload levels of the five modality conditions. Text appears to be the most difficult condition and text+speech is the easiest. The figure also suggests that the five conditions can be grouped into two clusters. Text and text+image form a relatively high cognitive load cluster, and the other three conditions form a relatively low cognitive load cluster.

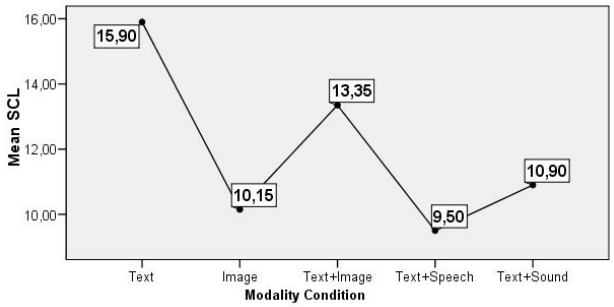

Figure 5. Average subjective cognitive load in five modality conditions

ANOVA results indicated that the subjective cognitive load level was significantly affected by the modality condition, $F(4, 76) = 16.91$, $p < 0.001$. We further applied a post hoc procedure (Bonferroni test) to make pairwise comparisons. The following six significant effects were found: 1) the text condition imposed higher cognitive load than the image condition; 2) the text condition imposed higher cognitive load than the text+speech condition; 3) the text condition imposed higher cognitive load than the text+sound condition; 4) the text+image condition imposed higher cognitive load than the image condition; 5) the text+image condition imposed higher cognitive load than the text+speech condition; and 6) the text+image condition imposed higher cognitive load than the text+sound condition.

RT

The average reaction time of all trials is shown in figure 6. On average, it took subjects between 1.9 seconds and 3.1 seconds to react to a patient. Text was the slowest condition and text+speech was the fastest one.

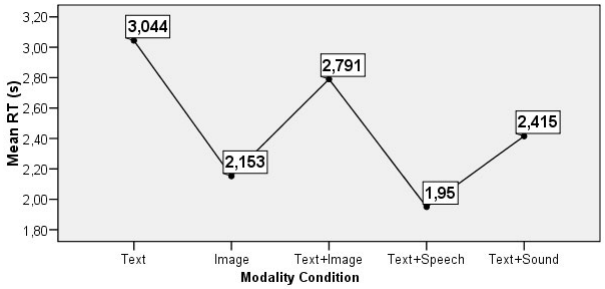

Figure 6. Average reaction time in five modality conditions

ANOVA results revealed a significant modality effect on reaction time, $F(2.87, 54.51) = 12.76$, $p < 0.001$. A post hoc test showed five significant effects: 1) subjects were slower in the text condition than in the image condition; 2) subjects were slower in the text condition than in the text+speech condition; 3) subjects were slower in the text+image condition than in the image condition; 4) subjects were slower in the text+image condition than in the text+speech condition; and 5) subjects were slower in the text+sound condition than in the text+speech condition.

ND

On average, the number of dead patients in each condition was between 2 and 12 (see figure 7). As 100 patients were presented in each trial, the percentage of saved patients was between 88% and 98%. Most patients were saved in the text+speech condition, and least were saved in the text condition.

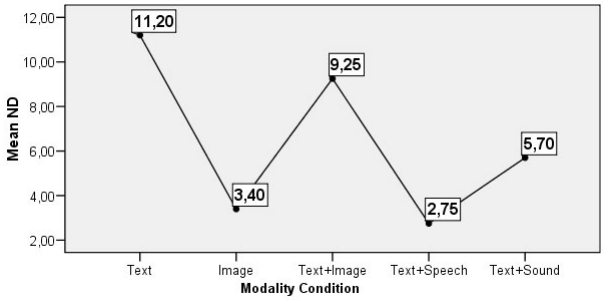

Figure 7. Average number of death patients in five modality conditions

ANOVA results indicated that there was an significant modality effect on the number of dead patients, $F(2.36, 44.84) = 16.81$, $p < 0.001$. Pairwise comparisons from a post hoc test showed five significant effects: 1) fewer patients were saved in the text condition than in the image condition; 2) fewer patients were saved in the text condition than in the text+speech condition; 3) fewer patients were saved in the text condition than in the text+sound condition; 4) fewer patients were saved in the text+image condition than in the image condition; and 5) fewer patients were saved in the text+image condition than in the text+speech condition.

The similar patterns in figure 5, 6, and 7 indicate that these three measurements might be positively correlated. Correlation tests confirmed this observation. Positive correlations exist at the 0.01 confidence level for ND-RT and ND-SCL, and at the 0.05 confidence level for SCL-RT. These results indicate that a higher subjective cognitive load is associated with lower reaction times and more dead patients. This consistency further confirms the modality effects on mental workload and performance.

HRV

No significant effects were found by the ANOVA of HRV data, $F(2.12, 40.29) = 1.17$, $p > 0.05$. This result can be interpreted as either that the modality condition factor does not have a significant impact on mental workload, or that the power in the LF band is not a sensitive mental workload indicator. As the modality effects on cognitive load have been confirmed by the other three measurements, we would rather trust the latter interpretation. We further applied a t-test on the experimental conditions (all five trials) and the baseline conditions. The result showed that the LF power was significantly higher in the baseline condition (Mean = 645.10 ms^2) than in the experimental conditions (Mean = 211.78 ms^2), $p < 0.001$. This in turn means that mental workload during the resting period was indeed significantly lower than the rescuing period. Therefore, we may conclude that the power in LF band, as an mental workload indicator, is only sensitive to major differences in the task demand. However, it failed to measure relatively minor differences between the five conditions in this experiment. The same conclusion has been found in [15] that the relationship between LF power and task demands "is generally found for relatively large differences in task difficulty"(p. 311).

DISCUSSION

All the significant modality effects found on the four measurements are summarized in table 2. We discuss these results and evaluate the hypotheses, as follows.

Measure	Effect	Sig.
SCL	text > image	p<0.001
	text > text+speech	p<0.001
	text > text+sound	p<0.001
	text+image > image	p<0.05
	text+image > text+speech	p<0.001
	text+image > text+sound	p<0.05
RT	text > image	p<0.001
	text > text+speech	p<0.001
	text+image > image	p<0.01
	text+image > text+speech	p<0.01
	text+sound > text+speech	p<0.05
ND	text > image	p<0.001
	text > text+speech	p<0.001
	text > text+sound	p<0.01
	text+image > image	p<0.01
	text+image > text+speech	p<0.01

Table 2. Significant modality effects on all four dependent measurements

Text vs. Image

Results of SCL, RT, and ND all suggest that image is more suitable than text in this scenario, thus the first hypothesis has been clearly confirmed. Image, as an analogue modality, is better for presenting concrete concepts, such as a wounded victim or a dead body. This representational characteristics indeed brought cognitive benefit (lower cognitive load) and performance benefit (faster reaction and fewer dead patients), even though the contrast between the two images was minor. We believe that the advantage of image over text will become even more notable when the two images contain larger contrasts in color, shape, and size.

Comparing the three modality combinations

First, we compare the visual-visual combination (text+image) with the visual-auditory combinations (text+image and text+sound). The results clearly show that text+speech is a more proper combination than text+image. A significant advantage of text+sound over text+image is only found in SCL, although the average RT (figure 6) and the average ND (figure 7) both show preference to the text+sound condition. The user task in this experiment imposes a high load on the visual channel. Based on the dual-channel theory, an extra arrow image, instead of functioning as a performance aid, further splits up the visual attention and causes distraction. Indeed, during informal interviews after experiments, many subjects mentioned that they had to ignore the arrow in order to concentrate on the rescue task. Only few subjects found the image aid somehow helpful, because they were able to perceive the arrow with the side view. However, when the performance aid was given in the auditory channel, it could be of real help without imposing any extra load on the visual channel. Therefore, the text+speech and the text+sound conditions imposed lower cognitive load and subjects performed better. Overall, the second hypothesis is confirmed.

Second, we compare the verbal-verbal combination text+speech with the verbal-nonverbal combinations text+sound. Only the RT results show a significant effect between these two conditions, indicating that subjects reacted faster in the text+speech condition than in the text+sound condition. Although conclusions can not be made from the SCL results and the ND results, the comparisons of average values still suggest that the subjective cognitive load was lower and fewer patients died in the text+speech condition than in the text+sound condition. When asked to compare these two conditions, most subjects mentioned that they preferred the speech slightly more, because it helped them to maintain a short queue of newly-coming patients in memory while searching for a current one. Only few subjects clearly preferred sound, because they were not quick enough to associate the words 'left' and 'right' with the directions; then a sound coming from one side was a more explicit indication for them.

As predicted, the principle of using a verbal-nonverbal combination to design better learning materials doesn't hold in our high-load information presentation scenario. Text+speech turns out to benefit memorization. Sometimes new patients were presented while a subject was still busy with searching for a previous one. In the speech condition, most subjects

could memorize the queue of 'left's and 'right's, and went for these patients after they found the one they were currently searching for. But they found it much harder to do the same in the text+sound condition. We try to understand this interesting result based on the working memory theory and the dual-coding theory. In addition to the dual-channel assumption, another very important statement of Baddeley's working memory theory is that "one characteristic frequently assigned to short-term memory is its reliance on speech coding - most models of short-term memory involve some process of rehearsal, usually via sub-vocal speech, to maintain the memory trace."([5] p. 49). Therefore, the speech 'left' and 'right' can be easily maintained in the short-term memory via sub-vocal rehearsal. However, the direction of a sound, as a nonverbal code, needs to be translated into a verbal code in order to be maintained. Based on the dual-coding theory, this translation requires associative processes between the verbal and nonverbal mental systems, thus requires extra cognitive resources. As a consequence, subjects found it hard to maintain a queue of un-treated patients in the text+sound condition. Moreover, there is actually no contradiction between the findings of this study and the multimedia learning studies, because the memorization in the learning context refers to the long-term memory, not the short-term memory (working memory). Therefore, a verbal-nonverbal combination helps long-term memorization in the learning context; while in our context, it assists short-term memorization only if the verbal modality is auditory.

Overall, the third hypothesis is also confirmed. However, we have discovered that only using the dual-coding theory is not sufficient to explain why the text+speech condition is better than the text+sound condition. Furthermore, a few subjects disliked the ambulance sound. It makes people vigilant, but can be irritating in the long run. It might be possible to improve the performance of the sound by using another more user friendly sound, e.g. a beep.

Single modality vs. Combined modalities

The results tell us that a modality combination is not necessarily better than a single modality, even though the combination actually carries more useful information. For example, results of SCL, RT, and ND all show that image is a more suitable modality than text+image. This indicates that a suitable modality which carries less information (image) can be still more beneficial than an improper modality combination which carries more information (text+image). On the other hand, text+speech is a better modality usage than text only. When we compare the mean values of all trials (table 5, 6, and 7), we see that text+speech is the best modality condition among the five. This indicates that a less suitable modality (text) can be significantly improved by being properly combined with another modality (speech). The fourth hypothesis has thus been partially confirmed.

In summary, the results from this experiment have shown us the power of a proper modality usage. In order to define a proper modality usage, one should follow principles at both the presentational level and the perception level.

Low load vs. High load

We further investigated whether the modality effects mentioned above would also occur without the high-load condition. At the beginning of each trial, no objects were on the grid map, thus the user task was relatively easy. As more and more objects were presented, it got more and more difficult to identify a patient in the crowded surroundings. After approximately one minute, the task difficulty reached the maximum (40% of the grid cells contain objects) and remained the same for the rest of the trial. The performance log shows that the first dead patients occurred after 60 seconds in all trials of all subjects. Therefore, we took the first 60 seconds as a relatively low-load period and recalculated the average reaction time during this period. Comparing figure 8 and figure 6, we see a similar main trend in the two plot lines, which indicates that the relative difficulty between conditions remains unchanged. However, the difference between the fastest condition (text+speech) and the slowest condition (text) is about 0.15s in figure 8, which is only around 14% of the difference calculated from the whole trial (1.09s, figure 6).

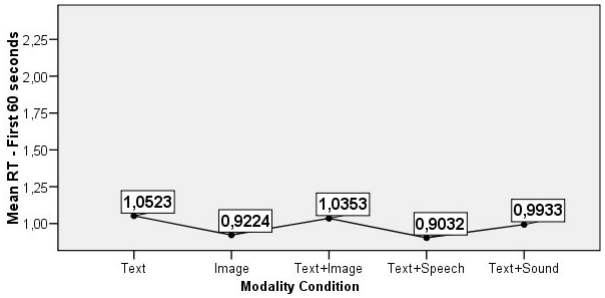

Figure 8. Average reaction time from the first 60 seconds in five modality conditions

ANOVA analysis showed that the modality condition factor does not have a significant effect on the reaction time during the first 60 seconds of a trial, $F(4,76) = 1.61$, $p > 0.05$. The modality effects on cognitive load and performance seem to hold also for a low-load conditions, but they become significant under a high-load condition. This result suggests that it is highly necessary to encode modality-related principles at both the presentation level and the perception level into the modality planning procedure in high-load information presentation systems.

CONCLUSION

In this study, we investigated the modality effects on cognitive workload and performance, using a high-load information presentation scenario. We put modality principles from both the presentational and the cognitive perspective into test.

At the presentational level, it is confirmed that image is more suitable than text for presenting concrete concepts, such as whether an earthquake victim is wounded or dead. Therefore, subjects experienced less mental workload and performed better in the image condition than in the text condition.

At the cognitive level, first, when the user task demands large capacity in the visual channel, visual-auditory combinations (text+speech, text+sound) impose less cognitive load and enable better performance than a visual-visual combination (text+image). This result is in line with the dual-channel theory which states that working memory has separated stores for visual information and auditory information, and each store has limited capacity. Second, between the two visual-auditory combinations, the text+speech condition has been shown to be better than the text+sound condition. It imposed less cognitive load, allowed faster reaction, and assisted the memorization of un-treated patients. This result can be explained by the dual-coding theory and the working memory theory. The dual-coding theory states that verbal and nonverbal presentations are presented in separate but interconnected mental systems. The working memory theory indicates that the maintenance of memory traces in the short-term memory is based on the sub-vocal speech. Speech can be directly rehearsed and maintained in the short-term memory; while the direction of sound needs to be translated from a nonverbal code into a verbal code. The associative processes between the two mental systems occupy extra cognitive resources and slow down the reaction.

We used two cognitive workload measurements and two performance measurements. Strong positive correlation has been found between subjective cognitive workload level (SCL), reaction time (RT) and the number of dead patients (ND). However, the heart rate variability feature extracted from the power in the low frequency band failed to reflect the differences in mental workload level under different conditions. This measurement is only sensitive to major mental workload changes, e.g the difference between baseline and experimental conditions.

The findings of this research bring several implications to the usage of modality in multimodal user interfaces. First, the cognitive properties of modalities influence the cognitive efficiency of the presentation. Cognitive properties explored in this study are the perception channel (visual / auditory) used by a modality and the mental system (verbal / nonverbal) used to process the information carried by a modality. Different modality usages can impose difference levels of cognitive load to the user even if they are used to convey the same information. In this study, a proper modality combination - text+speech - brings significant cognitive benefit and allows performance enhancement; while an improper combination - text+image, instead of providing extra performance aid, induces overload and causes distraction. Second, the cognitive effects of modalities might not be crucial to the quality of interaction when the cognitive load demand is low. However, they are enlarged and become significant under a high-load interaction situation where the information load is high, the user task is time critical, and the full capacity of human cognition is being challenged. Third, several modality-related cognitive principles, such as the dual-channel and dual-coding theory, can be used to predict and manipulate users' cognitive load level. In cognitive-aware user interfaces, these theories can be encoded as modality allocation rules and embedded in presentation planning structures.

LIMITATIONS AND FUTURE WORK

Future work is considered based on the limitations of the current study. First, the information to be conveyed was relatively simple, which especially limited the usage of text. In this specific case, using single words was already sufficient to convey an object type, thus there was no need to construct sentences with redundant information. However, the representative power of text modality is far beyond the single word level. For example, in situations where abstract information (such as casual relations between events or quantitative information) needs to be conveyed, text is very likely to be a more suitable modality than image. This limitation also goes for the other three modalities. They are able to represent much more than just an object type or location. Therefore in future experiments, we intend to use more complex information to further investigate the cognitive effects of modalities.

Second, the user task requires only the perception of the target objects (patients), which is a relatively low level cognitive task. We intend to further investigate the cognitive effects of modalities with more challenging user tasks, such as comprehension, reasoning and decision making. Moreover, when higher level cognitive tasks are involved, there might be factors other than the usage of modality (e.g. presentation order and structure) that influence the user performance and the cognitive load demand of the interaction. It would be also useful to investigate the interaction between the usage of modality and other factors.

ACKNOWLEDGMENTS

This research is part of the Interactive Collaborative Information System (ICIS) project. ICIS is sponsored by the Dutch Ministry of Economic Affairs, grant nr: BSIK03024. We thank C. Mühl for his help with the physiological recording. We thank E. L. Abrahamse and B. van Dijk for their advice on setting up the experiment. We also thank the 20 participants for their effort and time.

REFERENCES

1. Andre, E. The Generation of Multimedia Presentations. *Handbook of Natural Language Processing*, 305–327, 2000.

2. Arens, Y., Hovy, E. and van Mulken, S. Structure and Rules in Automated Multimedia Presentation Planning. In *Proc. IJCAI, '93*, 1253–1259, 1993.

3. Bachvarova, Y., van Dijk, B. and Nijholt, A. Towards a Unified Knowledge-Based Approach to Modality Choice. In *Proc. Workshop on Multimodal Output Generation (MOG) '07*, 5–15, 2007.

4. Baddeley, A.D. and Hitch, G.J. Working Memory. *The Psychology of Learning and Motivation: Advances in Research and Theory*, 8, 47–89, 1974.

5. Baddeley, A.D. Essentials of Human Memory. Psychology Press, 1999.

6. Bernsen, N.O. Foundations of Multimodal Representations: a Taxonomy of Representational Modalities. *Interacting with Computers*, 6(4), 347–371, 1994.

7. Bernsen, N.O. Multimodality in Language and Speech Systems - From Theory to Design Support Tool. *Multimodality in Language and Speech Systems*, Kluwer Academic Publishers (2002), 93-148, 2002.

8. Brunken, R., Plass, J.L. and Leutner, D. Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1), 53-61, 2003.

9. Buxton, W. Speech, Language and Audition. *Readings in Human Computer Interaction: Toward the Year 2000*, Morgan Kaufmann Publishers (1995), 525–570, 1995.

10. Chandler, P. and Sweller, J. Cognitive Load Theory and the Format of Instruction. *Cognition and Instruction*, 8(4), 293–332, 1991.

11. Clark, J.M. and Paivio, A. Dual Coding Theory and Education. *Educational Psychology Review*, 3(3), 149–210, 1991.

12. Elouazizi, N. and Bachvarova, Y. On Cognitive Relevance in Automatic Multimodal Systems. In *Proc. IEEE International Symposium on Multimedia Software Engineering '04*, 418–426, 2004.

13. Feiner, S.K. and McKeown, K.R. Automating the Generation of Coordinated Multimedia Explanations. *IEEE Computer*, 24, 33–41, 1991.

14. Hart, S.G. and Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Human Mental Workload*, 1, 139-183, 1988.

15. Kramer, A.F. Physiological Metrics of Mental Workload: a Review of Recent Progress. In Damos, D.L. ed. *Multiple-task performance*, 279–328, CRC Press, 1991.

16. Mayer, R.E. and Moreno, R. A Split-Attention Effect in Multimedia Learning: Evidence for Dual Processing Systems in Working Memory. *Journal of Educational Psychology*, 90(2), 312–320, 1998.

17. Mayer, R.E. and Moreno, R. Nine Ways to Reduce Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1), 45–52, 2003.

18. Paivio, A. Mental Representations: A Dual Coding Approach. Oxford University Press, 1986.

19. Scerbo, M.W., Freeman, F.G., Mikulka, P.J., Parasuraman, R. and Di Nocero, F. The Efficacy of Psychophysiological Measures for Implementing Adaptive Technology. *NASA Langley Research Center TP-2001-211018 NASA*, Hampton, 2001.

20. Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., MacEachren, A.M. and Sengupta, K. Speech-gesture Driven Multimodal Interfaces for Crisis Management. In *Proc. of the IEEE*, 91(9), 1327–1354, 2003.

21. Turoff, M., Chumer, M., Van de Walle, B. and Yao, X. The Design of a Dynamic Emergency Response Management Information System (DERMIS). *Journal of Information Technology Theory and Application*, 5(4), 1–35, 2004.

22. Wahlster, W. SmartKom: Symmetric Multimodality in an Adaptive and Reusable Dialogue Shell. In *Proc. Human Computer Interaction Status Conference*, 3, 47-62, 2003.

23. Wahlster, W., Andre, E., Bandyopadhyay, S., Graf, W. and Rist, T. WIP: The Coordinated Generation of Multimodal Presentations from a Common Representation. *Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues*, 121-144, 1992.

24. Wickens, C.D. Engineering Psychology and Human Performance (3rd edition). Prentice Hall, 1999.

25. Wilson, G.F. and Eggemeier, F.T. Psychophysiological Assessment of Workload in Multi-task Environments. In Damos, D.L. ed. *Multiple-task performance*, 329–360, CRC Press, 1991.