

Subjective Measurement of Workload Related to a Multimodal Interaction Task: NASA-TLX vs. Workload Profile

Dominique Fréard^{1,2}, Eric Jamet¹, Olivier Le Bohec¹, Gérard Poulain²,
and Valérie Botherel²

¹ Université Rennes 2, place Recteur Henri Le Moal
35000 Rennes, France

² France Telecom, 2 avenue Pierre Marzin
22307 Lannion cedex, France
{dominique.freard,eric.jamet,olivier.lebohec}@uhb.fr,
{dominique.freard,gerard.poulain,
valerie.botherel}@orange-ftgroup.com

Abstract. This paper addresses workload evaluation in the framework of a multimodal application. Two multidimensional subjective workload rating instruments are compared. The goal is to analyze the diagnostics obtained on four implementations of an applicative task. In addition, an Automatic Speech Recognition (ASR) error was introduced in one of the two trials. Eighty subjects participated in the experiment. Half of them rated their subjective workload with NASA-TLX and the other half rated it with Workload Profile (WP) enriched with two stress-related scales. Discriminant and variance analyses revealed a better sensitivity with WP. The results obtained with this instrument led to hypotheses on the cognitive activities of the subjects during interaction. Furthermore, WP permitted us to classify two strategies offered for error recovery. We conclude that WP is more informative for the task tested. WP seems to be a better diagnostic instrument in multimodal system conception.

Keywords: Human-Computer Dialogue, Workload Diagnostic.

1 Introduction

Multimodal interfaces offer the potential for creating rich services using several perceptive modalities and response modes. In the coming years, multimodal interfaces will be proposed for general public. From this perspective, it is necessary to address methodological aspects of new service developments and evaluations. This paper focuses on workload evaluation as an important parameter to consider in order refining the methodology.

In a multimodal dialoging system, various solutions can be encountered for implementation. All factors of complexity can be combined, such as verbal and non-verbal auditory feedback combined with a graphical view in a gestural and vocal command system. If not correctly designed, multimodal interfaces may easily

increase user complexity and may conduct to disorientations and overloads. Therefore, an adapted instrument is necessary for workload diagnostic. For this reason, we compare two multidimensional subjective workload rating instruments. A brief analysis of spoken dialogue conditions is presented and used to propose four configurations for information presentation to the subjects. The discrimination of subjects is intended depending on the configuration used.

1.1 Methodology for Human-Computer Dialogue Study

The methodological framework for the study of dialogue is found in Clark's socio-cognitive model of dialogue [2]. This model analyses the process of communication between two interlocutors as a coordinated activity. Recently, Pickering and Garrod [10] proposed a mechanistic theory of dialogue and showed that coordination, called alignment, is achieved by priming mechanisms at different levels (semantic, syntactic, lexical, etc.). This raises the importance of the action-level in the analysis of cognitive activities during the process of communication. Inspired by these models, the methodology used in human-computer dialogue addresses communication success, performance and collaboration. Thus, for the diagnostic, the main indicators concern verbal behaviour (e.g. words, elocution) and general performance (e.g. success, duration).

In this framework, workload is a secondary indicator. For example, Le Bigot, Jamet, Rouet and Amiel [7] conducted a study on the role of communication modes and the effect of expertise. They showed (1) behavioural regularities to adapt to - more particularly experts tended to use vocal systems as tools and produced less collaborative verbal behaviour - and (2) an increase in subjective workload in vocal mode compared to written mode. In the same way, the present study paid attention to all relevant measures. The present paper focuses on subjective workload ratings. Our goal is to analyze objective parameters of the interaction and to manipulate them in different implementations. Workload is used to achieve the diagnostic.

1.2 Workload in Human-Computer Dialogue

Mental workload can be described by the demand placed on user's working memory during a task. Following this view, objective analysis of the task gives an idea of its difficulty. This method is used in cognitive load theory, in the domain of learning [12]. Cognitive load is estimated by the number of statements and productions necessary to handle in memory during the task. This calculation gives a quantitative estimate of task difficulty. The workload is postulated to be a linear function of the objective difficulty of the material, which is questionable.

Some authors focus on the behaviour resulting in temporary overloads. In the domain of human-computer dialogue, Baber et al. [1] focus on the modifications of user's speech production. They show an impact of load increases on verbal disfluencies, articulation rate, pauses and discourse content quality. The goal, for the authors, is to adapt the system's output or intended input when necessary. Detection of overloads is first needed. In this way, a technique using Bayesian networking has been used to interpret symptoms of workload [5]. This technique is used to interpret the overall indicators in the same model. Our goal in this paper is not to enable this

kind of detection during a dialogue but to interpret workload resulting from different implementations of an application.

1.3 Workload Measurement

Workload measure can be reached with physiological clues, dual task protocol or subjective instruments. Dual task paradigms are excluded here because the domain of dialogue needs an ecological methodology, and disruption of the task is not desirable for the validity of studies. Physiological measures are powerful for their degree of precision, but it is difficult to select a representative measure. The ideal strategy would be to directly observe brain activity, which is not within the scope of this paper. In the domain of dialogue, subjective measures are more frequently used. For example, Baber et al. [1] and Le Bigot et al. [7] conduct the evaluation with NASA-TLX [3] since this questionnaire is considered as the standard tool for this use in Human Factors literature.

NASA-TLX. The NASA-TLX rating technique is a global and standardized workload rating "*that provides a sensitive summary of workload variations*" [3].

A model of the psychological structure of subjective workload was applied to build the questionnaire. This structure integrates objective physical, mental and temporal demands and their subject related factors into a composite experience of workload and ultimately an explicit workload rating. A set of 19 workload-related dimensions was extracted from this model and a consultation of users was conducted to select the most equivalent to workload factors. The set was reduced to 10 bipolar rating scales. Afterwards, these scales were used in 16 experiments with different kinds of tasks. Correlational and regression analyses were performed on the data obtained. The analyses identified a set of six most salient factors: (1) *mental demand*, (2) *physical demand*, (3) *temporal demand*, (4) *satisfaction* in performance, (5) *effort* and, (6) *frustration* level. These factors are relevant to the first model of the psychological structure of subjective workload.

The final procedure consists of two parts. First, after each task condition, the subject rates each of the six factors on a 20 point scale. Second, at the end, a pair-wise comparison technique is used to weigh the six scales. The overall calculation of task load index (TLX), for each task condition, is a weighted mean that uses the six rates for this condition and the six weights.

Workload Profile. Workload Profile (WP) [13] is based on the multiple resources model, proposed by Wickens [14]. In this model of attention, cognitive resources are organized in a cube divided into four dimensions: (1) *stage of processing* gives the direction: encoding as perception, central processing as thought and production of response. (2) *Modality* concerns encoding. (3) *Code* concerns encoding and central processing. (4) *Response mode* concerns outputs. With this model, a number of hypotheses are possible about intended performance. For example, if the information available for a task is presented with a certain code on a modality and needs to be translated in another code before giving the response, an increase of workload can be intended. The *time share hypothesis* is a second example. It supposes that it is difficult to share resources of an area in the cube between two tasks during the same time interval.

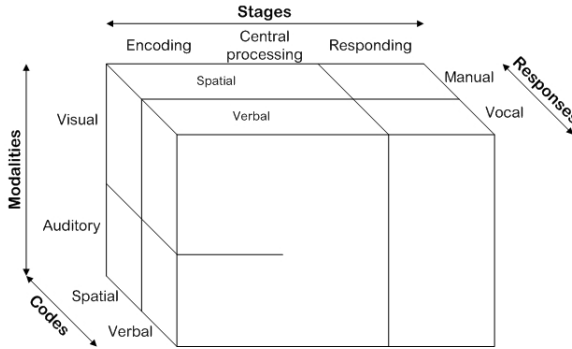


Fig. 1. Multiple resources model (Wickens, 1984)

The evaluation is based on the idea that subjects are able to directly rate (between 0 and 1) the amount of resources they spent in the different resource areas during the task. The original version, used by Tsang and Velasquez [13], is composed of eight scales corresponding to eight kinds of processing. Two are global: (1) *perceptive/central* and (2) *response processing*. Six concern directly a particular area: (3) *visual*, (4) *auditory*, (5) *spatial*, (6) *verbal*, (7) *manual response* and (8) *vocal response*.

A recent study from Rubio, Diaz, Martin and Puente [11] compared WP to NASA-TLX and SWAT. They used classical experimental tasks (Sternberg and tracking) and showed that WP was more sensitive to task difficulty. They also showed a better discrimination of the different task conditions with WP. We aim at replicating this result in an ecological paradigm.

2 Experiment

In Le Bigot et al's study [7] vocal mode corresponded to a telephonic conversation in which the user speaks (voice command) and the system responds with synthesised speech. On the opposite, the written mode corresponded to a chat conversation where the user types via keyboard (verbal commands only) and the system displays the verbal response on the screen. We aim at studying more detailed communication modes. The experiment focused on modal complementarity within output information: the user speaks in all configurations tested and the system responds in written, vocal or bimodal.

2.1 Analysis of Dialogic Interaction

Dialogue Turn: Types of Information. During the interaction, several kinds of information need to be communicated to the user. A categorization has been introduced by Nievergelt & Weydert [8] to differentiate *trails*, which refer to the past actions, *sites*, which correspond to the current action or information to give and *modes*, on the next possible actions. This distinction is also necessary when specifying a vocal system because, in this case, all information has to be given

explicitly to the user. For the same concepts, we use the words *feedback*, *response* and *opening*, respectively.

Dual Task Analysis. Several authors indicate that the user is doing more than one single task when communicating with an interactive system. For example, Oviatt et al. [9] consider multitasking when mixing interface literature and cognitive load problems (interruptions, fluctuating attention and difficulty). Attention is shared "*between the field task and secondary tasks involved in controlling an interface*". In cognitive load theory, Sweller [12] makes a similar distinction between cognitive processing capacity devoted to schema acquisition or to goal achievement. We refer to the first as the *target task* and to the second as the *interaction task*.

2.2 Procedure

Conforming to dual task analysis, we associate feedbacks with openings. They are supposed to belong to the interaction task. Responses correspond to the goal of the application, and they belong to the target task. Figure 2 represents the four configurations tested.

















Configuration	User command	Feedback	Response	Opening
AAA				
AVA				
VAV				
VVV				

Fig. 2. Four configurations tested

Subjects and Factors. Eighty college students from 17 to 26 years (M=19, 10 males and 70 females) participated in the experiment. They all had little experience with speech recognition systems.

Two factors were tested: (1) *configuration* and (2) automatic speech recognition (ASR) *error during the trial*. Configuration was administered in *between-subjects*. This choice was made to obtain a rating linked to the subject's experience with the implementation of the system rather than an opinion on the different configurations. ASR error trial was *within-subjects* (one with and one without) and counterbalanced across the experiment.

Protocol and System Design. The protocol was *Wizard of Oz*. The system is dedicated to managing medical appointments for a hospital doctor. The configurations differed only in information modality, as indicated earlier. No redundancy was used.

The wizard accepted any word of vocabulary relevant for the task. Broadly speaking, this behaviour consisted in copying an ideal speech recognition model. When no valid vocabulary was used ("Hello, my name's..."), the wizard of Oz sent the auditory message: "I didn't understand. Please reformulate".

The optimal dialogue consisted of three steps: *request*, *response* and *confirmation*. (1) The *request* consisted of communicating two research criteria to the system: the name of the doctor and the desired day for the appointment. (2) The *response* phase consisted of choosing among a list of five responses. In this phase, it was also possible to correct the request ("No. I said Doctor Dubois, on Tuesday morning.") or to cancel and restart ("cancel"...). (3) When a response was chosen, the last phase required a *confirmation*. A negation conducted to a new diffusion of the response list. An affirmation conducted to a message of thanks and dialogue ending.

Workload Ratings. Half of the subjects (40) rated their subjective workload with the original version of the NASA-TLX. The other half rated the eight WP dimensions and two added dimensions inspired from Lazarus and Folkman's model of stress [6]: *frustration* and *loss of control feeling*.

Hypotheses. In contrast to Le Bigot et al [7], no keyboard was used and all user's commands were vocal. Hence, both mono-modal configurations (AAA and VVV) are intended to lead to equivalent ratings and bimodal configurations (AVA and VAV) are intended to decrease workload.

Given Rubio and al's [11] results, WP should provide a better ranking on the four configurations. WP may be explicative when NASA-TLX may only be descriptive. We argue that the overall measurement of workload with NASA-TLX leads to poor results. More precisely, the studies concluded that a task condition was more demanding than the other one [1, 7] and no more conclusions were reached. In particular, no questions emerged from the questionnaire itself giving reasons for workload increases, and no real diagnostic was made on this basis.

2.3 Results

For each questionnaire a first analysis was conducted with a canonical discriminant analysis procedure [for details, see 13] to examine the possibility to discriminate between conditions on the basis of all dependent variables taken together. Afterwards, a second analysis was conducted with ANOVA procedure.

Canonical Discriminant Analysis. NASA-TLX workload dimensions did not discriminate configurations since Lambda Wilks' was not significant ($\text{Lambda Wilk} = 0,533$; $F(18,88) = 1,21$; $p = .26$).

For WP dimensions a significant Lambda Wilks' was observed ($\text{Lambda Wilk} = 0,207$; $F(30,79) = 1,88$; $p < .02$). Root 1 was mainly composed of *auditory processing* (.18) opposed to *manual response* (-.48). Root 2 was composed of *frustration* (.17) and *perceptive/central processing* (-.46). Figure 3 illustrates these results. On root 1, the VVV configuration is opposed to the three others. On root 2, AAA configuration is the distinguishing feature. AVA and VAV configurations are more perceptive. The VVV configuration is more demanding manually, and the AAA configuration is more demanding centrally (perceptive/central).

ANOVAs. For the two dimension sets, the same ANOVA procedure was applied to the global index and to each isolated dimension. Global TLX index was calculated

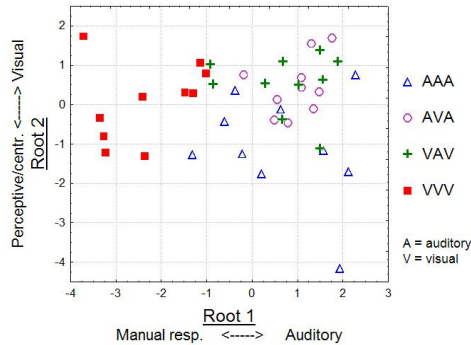


Fig. 3. Canonical discriminant analysis for WP

with the standard weighting mean [3]. For WP a simple mean was calculated including the two stress-related ratings. The plan tested configuration as the categorical factor and trial as a repeated measure. No interaction effect was observed between these factors in the comparisons. Thus, these results are not presented.

Effects of Configuration and Trial with TLX. The configuration produced no significant effect on TLX index ($F(3, 36) = 1,104$; $p = .36$; $\eta^2 = .084$) and no significant effect on any single dimension in this questionnaire. The trial gave neither significant effect on the global index ($F(1, 36) = 0,162$; $p = .68$; $\eta^2 = .004$) but among dimensions, some effects appeared: ASR error increased mental demand ($F(1, 36) = 11,13$; $p < .01$; $\eta^2 = .236$), temporal demand ($F(1, 36) = 4,707$; $p < .05$; $\eta^2 = .116$) and frustration ($F(1, 36) = 8,536$; $p < .01$; $\eta^2 = .192$); and decreased effort ($F(1, 36) = 4,839$; $p < .05$; $\eta^2 = .118$) and marginally satisfaction ($F(1, 36) = 3,295$; $p = .078$; $\eta^2 = .084$). Physical demand was not significantly modified ($F(1, 36) = 2,282$; $p = .14$; $\eta^2 = .060$). The opposed effect on effort and satisfaction in regard to other dimensions led global index to a weak representativity.

Effects of Configuration and Trial with WP. The configuration was not globally significant ($F(3, 36) = 1,105$; $p = .36$; $\eta^2 = .084$) but planned comparisons showed that AVA and VAV configurations gave a weaker mean than VVV configuration ($F(1, 36) = 4,415$; $p < .05$; $\eta^2 = .122$). The AAA and VVV configurations were not significantly different ($F(1, 36) = 1,365$; $p = .25$; $\eta^2 = .037$). Among dimensions, perceptive/central processing reacted like global mean: no global effect appeared ($F(3, 36) = 2,205$; $p < .10$; $\eta^2 = .155$) but planned comparisons showed that AVA and VAV configurations received weaker ratings compared to VVV configuration ($F(1, 36) = 5,012$; $p < .03$; $\eta^2 = .139$); and AAA configuration was not significantly different to VVV configuration ($F(1, 36) = 0,332$; $p = .56$; $\eta^2 = .009$). Three other dimensions showed sensitivity: spatial processing ($F(3, 36) = 3,793$; $p < .02$; $\eta^2 = .240$), visual processing ($F(3, 36) = 2,868$; $p = .05$; $\eta^2 = .193$) and manual response ($F(3, 36) = 5,880$; $p < .01$; $\eta^2 = .329$). For these three ratings VVV configuration was subjectively more demanding compared to the three others.

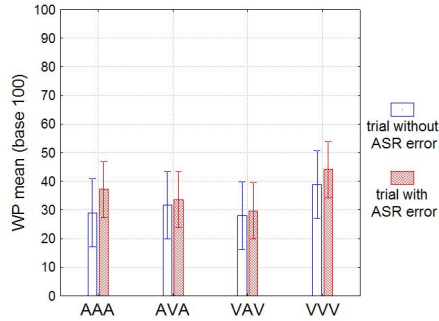


Fig. 4. Comparison of means for WP in function of trial and configuration

The trial with the ASR error showed a WP mean that was significantly higher compared to the trial without error ($F(1, 36) = 5,809$; $p < .05$; $\eta^2 = .139$). Among dimensions, the effect concerned dimensions related to stress: frustration ($F(1, 36) = 21,10$; $p < .001$; $\eta^2 = .370$) and loss of control ($F(1, 36) = 26,61$; $p < .001$; $\eta^2 = .451$). These effects were very significant.

Effect of Correction Mode. The correction is the action to perform when the error occurs. It was possible to say "cancel" (the system forgot information acquired and asked for a new request), and it was possible to directly correct the information needed ("Not Friday. Saturday"). Across the experiment: 34 subjects cancelled, 44 corrected and two did not correct.

A new analysis was conducted for this trial with correction mode as the categorical factor. No effect of this factor was observed with TLX ($F(1, 32) = 0,506$; $p = .48$; $\eta^2 = .015$). Within dimensions, only effort was sensitive ($F(1, 32) = 4,762$; $p < .05$; $\eta^2 = .148$). Subjects who cancelled rated a weaker effort compared to those who directly corrected.

WP revealed that cancellation is the most costly procedure. The global mean was sensitive to this factor ($F(1, 30) = 8,402$; $p < .01$; $\eta^2 = .280$). The ratings implied were visual processing ($F(1,30) = 13,743$; $p < .001$; $\eta^2 = .458$), auditory processing ($F(1,30) = 7,504$; $p < .02$; $\eta^2 = .250$), manual response ($F(1,30) = 4,249$; $p < .05$; $\eta^2 = .141$) and vocal response ($F(1,30) = 4,772$; $p < .05$; $\eta^2 = .159$).

3 Conclusion

NASA-TLX did not provide information on configuration, which was the main goal of the experiment. The differences observed with this questionnaire only concern the ASR error. Hypotheses have not been reached on user's activity or strategy during the task.

WP provided the intended information about configurations. Perceptive/central processing was higher in mono-modal configurations (AAA and VVV). Subjects had more difficulties in sharing their attention between the interaction task and the target task in mono-modal presentation. Besides, VVV configuration overloaded the three

visuo-spatial processors. Two causes can be proposed. First, the lack of perception-action consistency in the VVV configuration may explain this difference. In this configuration, subjects had to read system information visually and to command vocally. Second, the experimental material included a sheet of paper, giving schedule constraints. Subjects had also to take this into account when choosing an appointment. This material generated a split-attention effect and thus led to the increase of load. This led us to reinterpret the experimental situation as a triple task protocol. In the VVV configuration, target, interaction and schedule information were visual, which created the overload. This did not occur in the AVA configuration, where only target information and schedule information were visual. Thus, overloaded dimensions in WP led to useful hypotheses on subjects' cognitive activity during interaction and to a fine diagnostic on the implementations compared.

Regarding workload results, the bimodal configurations look better than monomodal configurations. But performance and behaviour must be considered. In fact, VAV configuration increased verbosity and disfluencies and led to a weaker recall of the date and time of the appointments taken during the experiment. The best implementation was AVA configuration, which favoured performance and learning, and shortened dialogue duration.

Concerning the ASR error, no effect was produced on resource ratings in WP, but stress ratings responded. This result shows that our version of WP is useful to distinguish between stress and attention demands.

For user modeling in spoken dialogue applications, the model of attention structure, underlying WP, seems more informative than the model of psychological structure of workload, underlying TLX. Attention structure enables predictions about performance. Therefore, it should be used to define cognitive constraints in a *multimodal strategy management component* [4].

References

- [1] Baber, C., Mellor, B., Graham, R., Noyes, J.M., Tunley, C.: Workload and the use of automatic speech recognition: The effects of time and resource demands. *Speech Communication* 20, 37–53 (1996)
- [2] Clark, H.H.: *Using Language*. Cambridge University Press, Cambridge (1996)
- [3] Hart, S.G., Staveland, L.E.: Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In: Hancock, P. A., Meshkati, N. (eds.) *Human mental workload*, North-Holland, Amsterdam, pp. 139–183 (1988)
- [4] Horchani, M., Nigay, L., Panaget, F.: A Platform for Output Dialogic Strategies in Natural Multimodal Dialogue Systems. In: *Proc. of the IUI, Honolulu, Hawaii*, pp. 206–215 (2007)
- [5] Jameson, A., Kiefer, J., Müller, C., Großmann-Hutter, B., Wittig, F., Rummer, R.: Assessment of a user's time pressure and cognitive load on the basis of features of speech, *Journal of Computer Science and Technology* (In press)
- [6] Lazarus, R.S., Folkman, S.: *Stress, appraisal, and coping*. Springer, New York (1984)
- [7] Le Bigot, L., Jamet, E., Rouet, J.-F., Amiel, V.: Mode and modal transfer effects on performance and discourse organization with an information retrieval dialogue system in natural language. *Computers in Human Behavior* 22(3), 467–500 (2006)

- [8] Nievergelt, J., Weydert, J.: Sites, Modes, and Trails: Telling the User of an interactive System Where he is, What he can do, and How to get places. In: Guedj, R. A., Ten Hagen, P., Hopgood, F. R., Tucker, H. , Duce, P. A. (eds.) *Methodology of Interaction*, North Holland, Amsterdam, pp. 327–338 (1980)
- [9] Oviatt, S., Coulston, R., Lunsford, R.: When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In: *ICMI'04*, State College, Pennsylvania, USA, pp. 129–136 (2004)
- [10] Pickering, M.J., Garrod, S.: Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27 (2004)
- [11] Rubio, S., Diaz, E., Martin, J., Puente, J.M.: Evaluation of Subjective Mental Workload: A comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied Psychology* 53(1), 61–86 (2004)
- [12] Sweller, J.: Cognitive load during problem solving: Effects on learning. *Cognitive Science* 12(2), 257–285 (1988)
- [13] Tsang, P.S., Velasquez, V.L.: Diagnosticity and multidimensional subjective workload ratings. *Ergonomics* 39(3), 358–381 (1996)
- [14] Wickens, C.D.: Processing resources in attention. In: Parasuraman, R., Davies, D.R. (eds.) *Varieties of attention*, pp. 63–102. Academic Press, New-York (1984)