Multimodal Concept for a New Generation of Screen Reader

Nadine Vigouroux, Bernard Oriola

IRIT-URA-CNRS 1399
Université Paul Sabatier
118, Route de Narbonne
F-31062 Toulouse Cedex
Phone: +33 61 55 63 14, Fax: +33 61 55 62 58
e-mail: vigourou@irit.fr

Abstract. One aim of this research is to explore how the selection of output modality and information presentation can be used to allow more appropriated access to electronic document by visually impaired persons. This paper shows how multi-modal interfaces can decrease certain difficulties linked with the visual disabilities. At IRIT our work consists in representing and developing multimodal access for the electronic document consultation. A representation space of output modalities is proposed in the goal of defining a multimodal user interface management system.

1. Introduction

Traditional interfaces are designed to allow direct manipulation of textual entities or graphic objects. They are typically command-oriented, completed with mouse to provide direct manipulation. An ordered set of manipulation is required to perform a given action. This requires familiarity with the system interface and the text processing functionalities. This fact is increasing instead of making them available for visually impaired persons (VIP).

The problem of having access to electronic documents (Eds) like newspapers, specialised magazines, books, ... [14, 23] by VIP, is completely renewed by the introduction of multimedia workstations [16] and multimodal interfaces [2, 18, 6, 1].

This paper presents the features of audio modality and the needs to define a modality model in order to specify a multimodal user interface management system (MUIMS). We will focus here particularly on the aural consultation of EDs by VIP. To perform this aural consultation, we need multimedia workstations provided with speech synthesis and/or speech recognition systems also as other communication devices such as adapted keyboards, mouse, adapted printers or displays, ... The users of these workstations can be disabled people like motor handicapped or visually impaired persons.

Years ago, Braille writing, allowing text access to the visually impaired, already realised an important progress. But this communication medium is still limited.

Indeed, only 10% of the visually impaired can read Braille. More over, they can only have access to a few transcribed and Braille printed texts, which are often old because the transcription takes a lot of time, while every day an important quantity of books, magazines and newspapers are published.

Recently, new progresses [25] have been possible thanks to the text-to-speech synthesis which allow reading of electronic information. The emergence of applications reading the contents of a screen offers the possibility to the blind to peruse and subsequently to converse with computers, to write programs or to read electronic texts [8]. We can list here the screen editors Edivox from Elan Informatique [7], Sonolect from Club Micro-Son, IBM Screen Reader [11], ...

But these aural devices are not yet sufficiently integrated in MUIMS, a fact which still restricts their scope. The paragraph 2 aims for a taxonomy of the modality concept. In paragraph 3, we focus on the aural modality features. The paragraph 4 describes our multimodal access system of EDs. Then, we lay the foundations of a referential space for the multimodal output presentation of EDs.

2. Taxonomy

Taxonomy of multimodal systems from an engineering point of view are emerging [12, 15]. First we should aim for a terminology that it is clear. Given that music, beeps, written and spoken natural language may be called ***modalities.*** These modalities are representational modalities. Each of this modality can be characterised by a set of features which serves to identify modalities from one another. These features are assumed for represented information: linguistic/non-linguistic, static/dynamic, and metaphoric/non-metaphoric.

For instance, the same linguistic information may be represented in either the graphical, sound or touch medium but the choice of the medium has influence on the perceptual qualities [10]. So perceptual modality qualities must be taken into consideration for the multimodal human-computer interfaces.

This paper reports first considerations from work undertaken on the modality concept for VIP. This concepts must be integrated in research on the information representation and exchange capabilities of multimodal interfaces. Since the beginning of the nineties, designers of interactive human-computer interfaces started to use increasing numbers of different, input/output modalities for the expression and the exchange of information between systems and their users.

Our intention is to propose a model of information-exchanging between user and system during **task performance in context,** defining the **input/output modalities** which constitute an optimal solution to the representation and exchange of the information.

Obtaining the modality model requires investigation in the following area:

- the study of aural modality,
- the definition of a conceptual and taxonomic space problem for analysing each type of mono- or multi-modal output representation (a),
- the integration of (a) in the human-computer interaction design.

3. Aural Modality

Speech recognition and speech synthesis are technologies of particular interest for their support of direct communication between user and computer.

3.1. Speech Input

The usability of Speech Recognition System is increasing. Speech will become an important component of the computer interface in dictation, report generation, automated telephone services, commands, ... For more information, see [21].

With regards to the technology thus involved, both technological advances, proper, and user behaviour toward this technology, make it reasonable to contemplate it as the possible interactive mode in various areas of voice applications [9].

There are some opportunities for using:

- shortcut rather than accessing a file by crossing many levels of hierarchical menus, a user can say "OPEN FILE" for example,
- information retrieval systems graphical user interfaces are awkward for specifying constraint based retrieval,
- preferable to keyboard in difficult situations because of free hand task.

CNET [25] experiments on using speech input as a means of running input/output voice servers, so to show that users now tend to prefer voice to keyboard interaction. More and more, speech input is favoured over DTMF telephone keyboards.

3.2. Sound Output

Different outputs based on sound modality may be used in the auralization of the human computer interaction such as speech synthesis, sound and/or music cues.

Speech synthesis offers an output channel in cases where visual displays are either not possible, insufficient or awkward. The constraints, set by voice-response systems, depend upon the latter types. There are two sorts of voice synthesis systems: the encoded speech systems and the text-to-speech systems.

The first ones yield a very good restitution of voice, with natural tone and prosody since they correspond to a digitised voice recording but takes a lot of place on the hard disk. To give an idea, when digitised through this method, a daily newspaper will require some 64 MB (with a 24 KB adaptive coding) [26]. For this reason, and because this kind of technique requires a numeration phase (long time consumable), the text-to-speech systems are preferable in the case of short-living documents of variable length.

The currently available text-to-speech systems include a grapheme-to-phoneme transcription module [24] that actualises technological and linguistic compromises, yielding an acceptable pronunciation for some 90% of the words making up a simple text. But this average drops sharply when tackling either technical texts, or small ads, telex, that involve acronyms, abbreviations, foreign words and/or linguistic exceptions. A good intelligibility of this type of texts demands considerable improvement of their pronunciation. In general, supplying lexicons of exceptions and/or abbreviations may afford a solution within the context of specific applications.

In order to overcome such limitations, an environment for the pre-processing of linguistic texts has been developed: TEXOR [4], a system operating on the orthographic string to be synthesised. This system uses both lexical and phonological knowledge, a set of morpho-orthographic rules and a bi-class grammar.

Not only a high quality voice speech synthesis is important for text comprehension but also all the spatio-temporal features and typographical characteristics contributed to intelligibility.

It is why sound modality can be released under different forms: sound, music or speech. In this context, Karsmer & al. [13] have demonstrated the advantage of using systems of music notes and chords, in order to make available menus of interactive systems for visually impaired persons.

3.3. The problems posed by an exploration in sound

We now give some characteristics that will be taken into account in a multidimensional sound space. Using a sound component as a means of feed back information requires that the following characteristics are to be respected:

- *access:*
  the aural information units are received sequentially, whereas visual entities are grasped (semi-)globally; the eye being able to scrutinise at once several elements of a structure ;
- *perception:*
  sound **restitution is but partial** as contemporary text-to-speech synthesis systems do not take into account all the semantic information that is associated to a linguistic statement (for more information, see [4]) - namely, the morpho-dispositional and typographical properties. Numerous studies have demonstrated the importance of morpho-distributional representation with respect to memorisation and understanding [27] ;
  the sound modality is **perceptible**, even if the operator is not in front of the source or awaiting information. Introducing the sound modality into processing control systems such as remote surveillance goes to show the intrinsic interest of this type of modality.
  Both of these latter points raise fundamental questions about communication theory and Searle-type language acts ([22, 19]).
- *nature of the modality:*
  sound is volatile, as opposed to text data or to static image, both of which are enduring. So, memorisation increases largely the cognitive load of the user. That's why we need to provide the user with redundant information by different output channels: aural, tactile and kinaesthetic.

4. Multimodal interface for electronic documents

We have chosen to put on our interface for VIP the greater number of the technical devices available at the present time:

- As input devices, the user has the possibility to use, in addition to the standard keyboard, a Braille keyboard and a speech recognition system ;
- As output devices, he has, in addition to the screen and the computer buzzer, a text-to-speech synthesis and a Braille display.

Our remarks will be more particularly turned towards the multimodal output devices. For more details concerning the different functionalities of our IODE system see [17, 18].

Reading a document has two phases; the first one concerns the choice of the text to be read and the second one the reading itself. These two stages determine two well distinguished behaviours:

- The navigation through the different menus is more pleasant on a Braille display than with the text-to-speech synthesis. As a matter of fact, it is possible to peruse a menu entry reading only a few letters while the synthesis will read it on its integrity which soon becomes boring.
- During the text consulting itself, the advantage of having two output media is constant. Nevertheless, we need not duplicate the information on each device but use at best the different features of the two communication modes.

For the users (young and using computers) the text-to-speech synthesis modality allows quickly to peruse a text because Braille reading is slower. It is then indicated for the text reading itself.
However, Pring and Rusted [20] have demonstrated that we memorise much more easily a fact if we have *"seen"* it tactile rather than only listened. We extend this conclusion to the VIP: they memorise better the things they read in Braille.

These facts lead us to propose that it is useful, maybe absolute necessary to relieve the cognitive load of the user spreading over the Braille display a certain number of information in order to punctuate the text with written marks. These can be structural and/or typographical information on the text read out or messages like the header for example.

On the other hand, words or sentences can be not well pronounced by the speech synthesis. That is the case for acronyms, abbreviations, foreign words, ... [4]. A Braille restitution will take off the ambiguity. The enrichment of this modality co-operation with a third aural dimension the computer buzzer or any other acoustic event [13] allows the VIP to mentally represent a certain number of marks like a change of paragraph or page, ...

All these establishments have lead us to define a model of multimodal output presentation based on the features of aural and tactile modality.

5. Towards a model of multimodal output representation

An interface is said to be multimodal [5] when it allows the use of different input/output communication channels.

Our work on the output representation and generation relies on the remarks and the conclusions of [12], exclusively devoted to the problems of a multimodal interpretation (axe: user to system).

The characteristics of the interfaces including several input devices have defined a reference space for the interpretation, namely the axis: modality usage and association of several knowledge sources in the interpretation mechanism. Our purpose here is to demonstrate the validity of this reference space and its extension for the output representation and generation. This reference space is defined according to the following axis (at the state of the model): modality usage, fission axis, output granularity and abstraction level.

- **The modality usage axis** reveal the temporal disposability of the output devices. The use is sequential, when, at a given time, only one output media can be used for output production. In the opposite case, we talk about parallel or simultaneous usage of media. For example, noises and speech constitute two kinds of aural information perceived by the aural channel. The transmission of the message from the system to the user can be spread over several channels in a simultaneous (parallel) or sequential way.
- **The fission axis** represents the possibility of the distribution of the message from the task to the user under different output representations. The lack of fission characterises the fact that the message has only one output mode. For example, we can consider the transmission of a message for the user in its tactile or aural form. If there is no semantic link between the different forms, they are said to be independent. On the other case, they are said combined.
- **The output granularity axis** takes account of the output discretisation level.
- **The level of output representation and abstraction axis** defines the degree of transformation of the message to send to the user. These abstraction levels depend on the state of the task and of the interaction; for example, a vocal message can be synthesised word by word in its totality relatively to its meaning in the task domain.

We can see here that the output multimodality problem space permits the classification of an output multimodality system. It also allows to account for the use of the output modalities. For example, during the system use, we can observe that a modality is not relevant in a synergetic use and a very great usability in an exclusive usage.

6. Conclusion

A given system can only allow one kind of multimodality. Another one can be hybrid and permits, depending on the **contexts, several types of multimodality.** In that case, the generation model must be able to decide dynamically on the appropriate strategy according to **the state of the task and to the user.** It is obvious that the implementation of this kind of output multimodal generation needs further research on the strategies of the users during their consultation task like in the MUIMS.

References

1. Y. Bellik, D. Teil: A Multimodal Dialogue Controller For Multimodal User Interface Management System, Application: A Multimodal Window Manager, In Proceedings of InterCHI'93 Proceedings, Amsterdam, (1993).
2. D. Burger: La multimodalité: un moyen d'améliorer l'accessibilité des systèmes informatiques pour les personnes handicapées. Dans ERGO.IA '92, pp. 262-277, (1992).
3. J.M. Carrol: Creating a Design Science of Human-Computer Interaction, in Interaction with Computers N° 5, 1, pp.3-12, (1993).
4. D. Cotto: Traitement automatique des textes en vue de la synthèse vocale", Thèse d'Université Toulouse III, (1992).
5. J. Coutaz, S. Balbo: Applications: a dimension space for user interface management systems. In CHI'91, ACM Publication, pp. 27-32, (1991).
6. A. Dufresne: L'importance d'un accès multimodal aux ordinateurs pour le non-voyant. Dans Le Curseur, vol II, n°1, pp. 1-4, Canada, (1993).
7. Elan Informatique: Vocalix Dos. Dans Aides techniques pour les déficients visuels, 1991.
8. J. Engelen, B. Bauwens: Large Scale Distribution of Text Information: The Harmonisation and Standardisation Efforts of the TIDE-CAPS Consortium. In Proc. of the European Conference On The Advancement of Rehabilitation Technology Ecart 2, pp. 27.1, Stockholm, (1993).
9. Joint ESCA-NATO/RSG10-Tutorial and Workshop - Applications of Speech Technology, Lautrach, Bavaria, 16-17 September, (1993).
10. Y. Hatwell: Images and non-visual spatial representations in the blind. In Non-Visual Human-Computer Interactions, Eds D. Burger, J.C. Sperandio, Colloque INSERM / John Libbey Eurotext Ltd, Vol. 228, pp.13-35, (1993).
11. IBM, Screen Reader: Système de synthèse vocale au service des non-voyants et des malvoyants. Dans Document Publicitaire, (1991).
12. Compte rendu des ateliers: Interface Multimodale et Architecture Logicielle. TELECOM Paris 93 S 004, pp. 9-45, (1992).
13. A.I. Karsmer, R.T. Hartley, K. Paap: Using Sound and Sound Spaces to Provide High Bandwidth Computer Interfaces to the Visually Handicapped. In SIGCAPH Newsletter ACM Press, N° 44, pp. 1-10, (1992).
14. D. Kochanek: A Hypertext System for the Blind Newspaper Reader. In Proc. of the 3rd Int. Conf. on Computers fro Handicapped Persons, pp. 285-293, (1992).

15. L. Nigay, J. Coutaz: A Design Space for Multimodal Systems: Concurrent Processing and Data Fusion. In Proceedings of InterCHI'93, pp. 172-78, (1993).

16. C. Lu: State of the Art:Publish It Electronically. In Byte, pp. 94-109 (1993).

17. B. Oriola, D. Cotto, G. Pérennou, N. Vigouroux:Consultation de documents électroniques pour les personnes handicapées de la vue. Dans Conférence "L'interface des mondes réels et virtuels", Montpellier EC2, pp. 489-500, (1992).

18. B. Oriola, G. Pérennou and N. Vigouroux: An Oral Input-Output Interface for the Reading of Electronic Newspapers by the Visually Impaired. In 1st TIDE Congress, Brussels, (1993).

19. E. Pascual, J. Virbel: Connaissances linguistiques et morpho-dispositionnelles pour le contrôle de la décomposition structurelle des documents. Dans Colloque national sur l'écrit et le document, CNED, Actes dans Bigre, 80, pp. 217-224, (1992).

20. L. Pring, J. Rusted: Pictures for the blind: an investigation of the influence of pictures on recall of text by blind children. In British Journal of Developmental Psychology, 3, pp. 41-45, (1985).

21. A. Rudnicky, G. Hauptmann, K.F. Lee: Survey of Current Speech Technology. In Communication of ACM, Vol. 37, N°3, pp. 52-57, (1994).

22. J.R. Searle: Les actes du langage. Hermann, Paris (1972).

23. F.P. Seiler, N. Vigouroux, M. Truquet, P. Bazex, C. Decoret: Access To electronic Information for Visually Impaired Persons. Seminar Day, 4th ICCHP Vienna, (1994).

24. C. Sorin: Synthèse de la parole à partir du texte: état des recherches & des applications. Dans Deuxièmes journées nationales du GRECO-PRC, CHM, Toulouse, pp. 131-146, (1991).

25. C. Sorin, D. Jouvet, M. Toularrhoat, C. Gagnoulet: Commande vocale et synthèse à partir de texte pour les services vocaux: l'expérience du CNET. Dans Conférence "L'interface des mondes réels et virtuels", Informatique 93, Montpellier EC2, pp. 305-310, (1993).

26. U. Strempel: Presentation of different types of Talking Newspapers with Special emphasis on ETAB. Stiftung Blindenanstalt Frankfurt am Main, Allemagne, (1990).

27. B. Thon, J.C. Marque, P. Maury: Le texte, l'Image et leurs Traitements Cognitifs. Dans Colloque Interdisciplinaire du CNRS, "Images et Langages" Multimodalité et Modélisation Cognitive, pp. 29-39, Paris, (1993).