HYPERAUDIO : VOCAL NAVIGATION STRATEGIES IN A HYPERMEDIA ENVIRONMENT

Florence Sèdes*, Nadine Vigouroux, Philippe Truillet**, Bernard Oriola****
IRIT-URA CNRS 1399
118, Route de Narbonne - 31062 Toulouse Cedex - FRANCE

*** "Generalized Information Systems" Team**
**** "Human-Computer Communication by Speech and Text" Team**

Tel : (+33) 61 55 63 14 - Fax : (+33) 61 55 62 58
e-mail: **sedes@irit.fr - vigourou@irit.fr**

ABSTRACT: Hypermedia systems emphasize the paradigm of graphic user interface based on navigation through a two/three dimensional space. The HyperAudio described in this paper, a mainly speech-based hypermedia interface, explores issues of navigation in audio environment without a visual display.
The user interface under development uses spoken commands to browse through the hyperbase. For contextual information, user feedback and hypermedia object reading only technologies of audio output are used.
Navigation strategies in the audio space are more difficult than in the spatial one. The concepts of navigation need to be redefined. Some navigation solutions based on a voice interface will discuss how auditory information is processed. We use audio cues as substitution for visual ones, when reading. These audio cues provide navigational information when the reader moves through a hypermedia base. This paper focuses on audio concepts and issues raised by the need for intelligent navigation tools making a system capable of reading a hypermedia base.

INTRODUCTION

Our approach aims to use hypertext as a medium for integrating different information sources text, image and sound for structuring in a heterogeneous environment. Indeed, the hypermedia paradigm provides access to all the information elements regardless of where they are located and whatever their type is: any inquiry or semantic relation is traceable allowing the user to access by sequential reading or navigation through the network of interrelated retrieved information.

Hypermedia information interfaces are generally graphically oriented to present and to browse through the hyperbase. The HyperAudio described in this paper, a mainly speech based hypermedia interface, explores issues of audio based navigation strategies which will be described in details later. We will discuss the differences between visual and audio strategies with regard to the nature of the modality.

The user interface under development uses spoken commands to browse through the hypermedia structured information. For contextual information, user feedback and hypermedia object reading a multimodal sound output is developed.

Navigation strategies in the audio space is more difficult than in the visual space. The concepts of navigation techniques need to be redefined. Some navigation solutions based on a voice interface will discuss the manner in which auditory information is processed. In this research prototype we use audio cues as substitutes for visual ones, when reading. These audio cues provide navigational information when the reader moves through a hypermedia base. Our aim is to investigate what kinds of information should be communicated to the reader (the listener):

- document contents,
- reader location (at a given time) within a document or within the global structure,
- interaction context (i.e., for which intention and in what circumstances this information is successfully communicated).

AUDIO MODALITY: ADVANTAGES AND LIMITATIONS

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

Speech Input

The usability of Speech Recognition System (SRS) is increasing. Speech will become an important component of the computer interface in dictation, report generation, automated telephone services, commands, For more information, see (Rudnicky et al.1994).
The speech technological advances and the suitable user behavior make it reasonable to contemplate it as the possible interactive mode in various areas of voice applications (ESCA 1993).
There are some of opportunities for using SRS:

- 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 situation —free hand .

CNET experiments (Sorin1993) on using speech input as a means of running input/output voice servers also show that users now tend to prefer voice to keyboard interaction. In HyperAudio paradigm, the SRS is used as commands to control the interface and for specifying constraints-based in natural language like : "Accéder à l'article du projet HyperAudio" (Access to the article of HyperAudio Project").

Sound Output

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

Moreover, 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 on the two main types of techniques to produce speech synthesis:

- restitution so-called coded or digitized speech,
- synthesizing of speech from text.

The text-to-speech cards include a grapheme-to-phoneme transcription module that actualizes technological and linguistic compromises, yielding an acceptable pronunciation for some 90% of the words making up a simple text. But this percentage drops sharply when tackling either technical texts, or small ads, telex's, etc., that involve acronyms, abbreviations, foreign words, proper names, 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.

For summarizing text-to-speech technology allows to synthesize any sentence in orthographic forms. Not only a speech synthesis with a high quality voice is important for text comprehension but also all the spatio-temporal features and typographical characteristics contribute to intelligibility.
It is why sound modality can be released under different forms; each of which is worth of interest and of special importance: sound, music or speech. In this context, (Karshmer et al.1992) have demonstrated the advantage of using systems of notes and chords, in order to make available menus of interactive systems for the visually impaired persons.

In this article, we will present the proper use of speech and other types of sounds in the navigation process based on vocal modality in an aural interface. This usability requires a new interface paradigm rather than a simple improvement of the graphic user interface.

THE CONCEPT OF AURALIZATION IN A HYPERMEDIA SYSTEM

In this research, we define the hypermedia concept as follows (Chrisment et al.1993), (Sèdes et al. 1994): it is based on the structuring of semantic units (nodes) from what are sometimes no more than bits and pieces of multimedia information. The hypermedia data model that we have developed was adapted for the needs of voice interfaces. This structuring is managed by means of relations (links) between the fragments of information. Hypertext systems enable access to a semantic unit through two types of methods: navigation and search through links. All it takes for a user to cross a link is to select a button in a menu, triggering a change of work space. The problem here is the link path within a multimodal sound space.

The problems posed by an exploration in sound

The navigation principle rests upon progressing into the network of nodes (hypergraph or "web") in order to reach information units, strictly via links. When looking at this from a human-machine interaction standpoint, most of the consultation mechanisms are based upon the paradigm of direct consultation mouse / screen inside a two dimensional visual space imaging the hyperbase.
In contrast to these approaches, the herein described HyperAudio Interface explores a strictly aural-based consultation method, enabling the user to consult the hyperbase without resorting to any visual information. This type of exploration raises a number of considerations related to the very nature of aural modality, as well as to the substitution of a multidimensional sound space to a spatio-temporal one.

Let us give now some characteristics that will be taken into account in a multidimensional sound space. Using a sound component as a means of displaying information requires that the following characteristics are taken into account:

- nature of the modality: sound is volatile, as opposed to text data or to static image, both of which are enduring,
- access: the aural information units are received sequentially, whereas visual entities are grasped (semi-)globally; an eye being able to scrutinize at once several elements of a structure,
- perception: sound restitution is partial; as contemporary text-to-speech synthesis cards do not put out a full account of all the semantic information (for more information, see (Cotto 1992)) namely, the morpho-dispositional and typographical properties.

Therefore, within hypermedia systems, the concept of navigation should be redefined: the whole notion of viewing the web from above has to be analyzed a new. As far as structuring is concerned, the problem of information granularity has to be posed.
Few information systems including audio information entities, resort to using the sound component simultaneously in both structuring and consultation processes. Among them, a number of forerunners should be mentioned:

- (Muller et al.1991)'s HyperPhone Prototype uses oral input/output devices for consulting hypermedia documents;
- (Arons 1993)'s HyperSpeech Project; a system enabling the user to navigate, thanks to a vocal input, within a digitized recording base of interviews. Information for system control e.g., spatio-temporal help mechanisms is returned through a synthesis from the text. The particular feature of this system lies with the themewise automatic segmentation process, themewise or in speech acts in the sense of Searle (Searle 1972).

The goals aimed at with our HyperAudio are:

- simultaneous availability of pre-recorded sounds and of textual units; thus, from the point of view of information units, combining the functional advantages of both HyperPhone and HyperSpeech.
- modeling both the concept of sound documents thus defined and an associated consultation methodology.

Our project is particular in that it uses audio input/output devices, to set annotations also called private marking as well as to enable access mechanisms used during consultation. In this article we will present only the menu consultation based on aural modality.

THE CONSULTATION INTERFACE

Various approaches to navigation (Conklin 1987) may co-exist; namely,

- by theme: from an initial choice of a theme picked out from a list, manifold itineraries through various previously defined paths,
- by index: same principle but from an initial choice of one or several keywords,
- by position: thanks to a pointing device or graphic browser i.e., a graphic representation of part of a network while visualizing all existing links; accessing information through mere pointing on the corresponding link,
- by abstract: replacing graphic representation by a text (an abstract) supplying all pertinent information on the node; all links being integrated into the text and visualized in reverse video, enhanced brilliance, colors....
- by tour: enabling the author to blend the logical structure of the document into the hypertext; resulting therefore in a linear presentation (chapters, sections, subdivisions, etc.) and in path definition from the outset.

Within the scope of our project, specifics of the sound component have to be considered; e.g., the fact that any visual element (picture, graph, etc.) has to be represented by a description for retrieval, that is associated to a sound component for restitution; making necessary to have a textual or oral description of all merely visual information.

The function of the interface is to make possible a sound consultation of the hyperbase. Given the essentially graphic and visual nature of the helps, proposed within "classical" contexts, the paradigm of sound consultation can be defined thus: how is a spatio-temporal representation of the hyperbase to be accounted for within a sound base ?

The principle of visual enhancement of anchors can be applied: link transparency, expressed through the absence of any marker, supplies the user with a clue as to the nature of the reference, while preserving text continuity which is fundamental, if continuity in the "listener's" discourse is to be preserved .

Therefore, we suggest to use a different type of sound or tone, in order to distinguish an anchor from the rest of the text (an equivalent of enhanced brilliance or reverse video). Taking titles or node names as landmarks is important in resisting "listener" disorientation. In general, a link allows for a connection to be made near the beginning of a node: an abstract may then be stated from the outset, curtailing navigation efforts and sparing user time from irrelevant access. Thence it is possible either to carry on with a consultation through retrieving the whole node content, or to come back in the direction of the previous node.

Making a graphic screen accessible for instance by visually impaired persons can be contemplated by accounting for mouse localization (in general, by adding a scale of sounds identifying the one window which is active at a given time). This process can, indeed, be envisaged though numerous problems remain to be solved: e.g., identifying window location on the screen, memorizing the various tones and differentiation of two adjacent tones (Edwards1989), (Karshmer et al.1992). Yet, the interest of such a retrieval mechanism, simulating multiwindow functions is perhaps questionable: a simultaneous view translated by a simultaneous statement of two texts with different frequencies!
Our experimental results show that this way is not realistic and that an "as is" application of visual techniques can turn out to be disastrous, introducing more "noise" than help to the "listener".

Consultation processes

The classical system offers mainly two alternate consultation modes (Nielsen 1990): menu or navigation (passive or active).

The sound menu consultation mode consists in initially locating oneself on a node, which was found in a table of contents and/or an index of marks. Next, speech input permits movement, and analysis thereof, to settle on a hyperdocument node.

After retrieving the node (possibly, the whole of it, if not its name, title or abstract) a menu of commands is then offered. The user enters a selection, by voice or keyboard interaction, and the sequence is repeated: upon settling on the selected node, the system offers the list of subsequent nodes to be consulted, not necessarily including nodes already looked up. If no further node is available, going back one or several consultation levels is possible (depending on the underlying tree-like structure).

This type of consultation is rendered difficult because of backward movement and vocal repetition of menus. To counter this disadvantage, we would like to suggest that the menus offered by the system should be made fix with a comment specifying invalidated links (so as to avoid circuits within consultation paths): when menus do not list such links, the reader must at the same time conceptually blend in the dynamics of menu-form evolution, while also knowing where to position oneself inside the hyperbase. As already mentioned, far from being negligible, the effort of thus intellect called for results in an increased laboriousness of attention to follow discourse continuity: given the trouble, the thread of the leading idea motivating the consultation to start with might get lost easily.

The navigational consultation mode, on the other hand, consists in using soundmarks (also called "audicons") embedded in the audio sequences and corresponding to link activation anchors. If no mark activation occurs, reading goes on sequentially. On the opposite, whenever mark activation

occurs, the reader ("listener") may do either one of two things: take the proposed path or proceed with the reading, if the vocal commentary associated to the mark is not especially of interest.
This kind of navigation is characterized by an introduction of numerous links. Typification of these links happens according to the choice made among the activated dialog box of the navigation system.

Dialog boxes are used by the system to pose a list of questions. The reader supplies all or a fraction of the answers reflecting his/her motivation for the search process. An analysis of these answers gives suggestion to a path, enriched with an explanatory comment.
A more ambitious phase, though not yet experimented, consists in implementing an architecture that backs up topical-similar search, including a sound component music and/or sounds,.. A simple, though interesting case, is that of semantically typified marks through tone differentiation. From an index compiling all the different types of soundmarks, searching through the hyperbase while dynamically structuring a consultation path can be contemplated.

THE FRONT-END SYSTEM

HyperAudio uses commercial input/outputs devices. The prototype consists of:

- (a) a hyperbase built automatically on the proceeding of a French congress in linguistics in French language on a UNIX workstation,
- (b) a speaker dependent continuous speech recognition and system for 500 words (VECSYS 1992),
- (c) a text-to-speech system for French language based on a PSOLA technology (ELAN 1991)
- (d) and the audio tools of the SUN-4 Workstation.

Command	Action
Play	Displays a node
Stop	Stop of audio displaying (time-compression speech, text-to-speech synthesis)
Parameter Modification	Modification of speech synthesis parameters (speech, volume, prosody)
Undo	Enables the previous command
Repeat	Do again the same command
Next	Next Node
Previous	Previous Node
Where-am-I	Displays the current position
Save	Saves the path

The above table shows the main spoken commands used for the navigation. The speech recognition system was assessed in a lab by several members of the team. The performance obtained after a fine training phase gained toward 99% of accuracy. It is why after each spoken command it was decided to cancel the feedback of the recognized command: users of the system prefer HyperAudio if the command echoing is turned off. Until now there is no repair dialog.

We have proceeded towards an intelligibility evaluation of the text-to-speech synthesis used for the textual node displaying of the hyperbase. we found that the three major parameters leading to difficulties in the comprehension in text-to-speech synthesis are:

- the pronunciation errors (1),
- the missing of the semantic structure (2) of the text
- and the prosody poorness(3).

In order to overcome limitations (1) , an environment for the pre-processing of linguistic texts has been developed: TEXOR (Cotto1992), a system operating on the orthographic string to be synthesized to be used as input for TELEVOX-PSOLA (ELAN 1991). This system uses both lexical and phonological knowledge, a set of morpho-orthographic rules and a bi-class grammar. We have defined this adaptation and lead an evaluation: the rate of mispronounced words was reduced to around 12 %.

In order to cover the limitation (2) a multimodal presentation system is been developed to interpret and to "display" the structure of a document by using the ICADD-DTDs extension of the CAPSNews DTD (Document Type Definition) (Kruger 1992) of ICADD (International Committee for Accessible Document Design) compatible with the model of the hyperbase.

CONCLUSION

The HyperAudio concept described in this paper demonstrates the possibilities to use auralization concepts in navigation strategies.
The hyperbase concepts and structure have been enriched to take into account the specialties of sound modality for acquisition and restitution; hyperbase organization and representation must render the corresponding specifications. The foregoing experiments lead us to the beginning of a definition of voice interface for hyperbase while some levels of nodes and links have been generated by the hypertext model. Many adaptations seem to be necessary with regard to the granularity and the type of the information and the user's behavior.

Two strategies were developed to get from one specific node to another. The interface is goal directed: the speech input allows direct addressing what in opposite is difficult to capture in other modes of communication.

The first evaluation of the system shows that speech output is a good communication medium but the levels of meaning must be embedded in intonation to reproduce the interactive conversational aspects of the reading function.

There are some technological and human-computer interaction aspects that were presented and must be solved for the HyperAudio system prototype. The HyperAudio principle can be a first approach for a new medium of interaction.

REFERENCES

Arons (1993) B. ARONS, "HyperSpeech", in *Proceedings of InterCHI'93* April 1993, pp. 524.
Buxton et al. (1991) W. BUXTON , B. GAVER, S. BLY, "The Use of Non-Speech Audio at the Interface", in *ACM SIGCHI*, 1990, Tutorial Notes.
Chrisment et al. (1993) C. CHRISMENT , C. COMPAROT, C. JULIEN, P.Y. LAMBOLEZ, F. SEDES, "Hyperdocument Structuration, Management and Exchange in an Object Oriented Database" in Proceedings of the 2nd Symposium on Information Retrieval and Document Analysis, Las Vegas (U.S.A.), April 1993, pp. 1-21.
Conklin (1987) J. CONKLIN, "Survey of Hypertext", in *Microelectronics and Computer technology Corporation*, pp. 356-86, 1987.
Cotto (1992) D. COTTO, *Traitement automatique des textes en vue de la synthèse vocale*, Thèse d'Université Toulouse III, Décembre 1992.
ESCA (1993) Joint ESCA-NATO/RSG 10 Tutorial and Workshop Applications of Speech Technology, Lautrach, Bavaria, 16-17 September, 1993.
Edwards (1989) A.D.N. EDWARDS, "Soundtrack : an Auditory Interface for Blind Users", in *Human Computer Interaction, 4,* (1) pp. 45-66.
ELAN (1991) ELAN INFORMATIQUE, *CARTE TELEVOX PSOLA Version 2.00*, 1991; Toulouse, France.
Gaver (1991) W.W. GAVER, "The Sonic Finder : An Interface That Uses Auditory Icons", in *Human Computer Interactions*, Vol. 4, N¡ 1, pp. 67-94.
Kruger (1992) M. Kruger: CAPSNews DTD, Version 1.0, October 1992.
Hypertext (1989) *Hypertext'89*, ACM Conference, University of York, November 1989.
Karshmer et al. (1992) A.I. KARSHMER, 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*, January 1992, Number 44, pp. 1-10.
Muller et al. (1990) M. J. MULLER, J.E. DANIEL, "Toward a Definition of Voice Document", in *Proceedings of COIS'90*, 1990, pp. 174-183.
Nielsen (1990) J. NIELSEN, "The Art of Navigation through Hypertext" in *CACM 90*, Vol. 33, N¡ 3, March 1990, pp. 296-310.
Rudnicky et al. (1994) A. RUDNICKY, A. G. HAUPTMANN, K.F. LEE, "Survey of Current Speech Technology", in Communication of ACM, Vol. 37, N¡ 3, pp. 52-7.
Sèdes et al. (1994) F. SEDES, L. BONNARD, B. ORIOLA, N. VIGOUROUX, "The HyperAudio project:êHow to extend hypermedia to Audio ? ", in *Proceedings of BIWIT'94*, Basque International Workshop on Information Technologies, Biarritz, France, February 1994, pp. 27-38.
Searle (1972) J.R. SEARLE, "Les actes du langage", HERMANN, Paris (1972).
Sorin (1993) C. SORIN, C. GAGNOULET, "CNET Speech recognition and text-to-speech for telecommunications applications", in ESCA Workshop on Applications of Speech Technology, Lautrach, Germany, 16-17 September 1993, pp. 31-4.
Vecsys (1992) Guide de l'utilisateur: LOgiciel DATAVOX Version 2.82, 1992.