Effect of "Sound Fonts" in an Aural Presentation

Philippe Truillet[1], Bernard Oriola[1], Jean-Luc Nespoulous[2], Nadine Vigouroux[1]

[1]IRIT UMR CNRS 5505
118, Route de Narbonne 31062 Toulouse Cedex 4, France
Tel +33 561 556 314

[2]Laboratoire Jacques LORDAT
Maison de la Recherche, 5 Allées Antonio-Machado, 31058 Toulouse Cedex1
Tel +33 05 61 50 46 72
E-mail:{oriola, truillet, vigourou}@irit.fr, nespoulo@univ-tlse2.fr
URL:http://ihmpt.free.fr

Abstract. This paper deals with research for the design of "sound fonts" and development of an evaluation methodology suitable for use with non visual presentation based on the speech modality or on multimodality (speech and tactile).
The work hypothesis of the study presented here relies on the fact that both structure and typographic attributes increase the comprehension process in visual presentation. Based on this constant, the Human Computer Interaction (HCI) question is to find alternative sounds or prosodic variants to display the typographic attributes –bold, italic–, for instance. This question takes part of the paradigm of the information accessibility problems.

Keywords. Aural presentation, speech prosody, "sound fonts", blind people, efficiency of presentation, evaluation methodology

1. INTRODUCTION

The explosion of computer based information applications and the emerging of interaction techniques ? gesture, visual, spoken? introduce new challenges for the assistive technology. This fact is stressed by the widespread use of emerging telematic services - Interactive Voice Response (IVR) systems, spoken dialogue through Internet, etc.- which enable distant/nomadic access to information. These objectives have introduced new problems in the field of the HCI (Human Computer Interaction) research groups mainly concerning the issue of the accessibility of Information Society Technologies for all[1] [Stephanidis 00], [Savidis 00].

On one hand, speech synthesis technologies offer new functions to control the prosodic model which allow to "display" both the layout and the content with parameter values (energy, rate, pitch). On another hand, there are a lot of works about the extension of markup languages specialized in multimedia rendering.

Internet users are accustomed to visual presentation of HTML (Hyper Text Markup Language) documents on a screen. However, this fact is changing with the development of new interaction concepts like WebPhone or WebTV. The user needs to consult his/her mail or some Web pages anywhere and whenever. This phenomenon is accelerated with new

[1] *"all": means people with different cultural context, novice and experienced computer users, the young and the elderly, people with different types of disabilities.*

protocol development such as WAP (Wireless Access Protocol) which allows Internet consultation by means of cellular phones.

We have to conceive new alternatives to interaction and to visual presentation, particularly the improvement of the vocal input/output technology and its low cost makes its use possible [Truillet 97], [Roth 99].

These alternatives are important for the visually impaired and elderly persons. In the last decade, assistive technology entails software and hardware adaptations which facilitate access to the content of the document via filtering procedures and development of specialised input/output techniques. Some technical solutions as screen or browser readers, are available on the market. These screen/browser readers simply "display" the ASCII text after the filtering of graphic objects and/or adaptation of the document structure (loss of the spatial organisation, for instance). They use Text-to-Speech systems (TTS) and/or Braille displays. These solutions lay out on a vocal or a tactile presentation of graphical objects (icon, dialog box, menu, text, etc.).

Moreover, some works [Vivier 97] in the field of cognitive research have pointed out that the document layout is a sense carrier and seems to increase the comprehension and memorisation processes of structured documents for a visual presentation.

We first discuss about the needs to improve information presentation in an aural form with the use of dedicated markup languages. Then, we describe the design of the method and the test we have performed on people's ability to perceive the "sound fonts". The purpose of this experimentation is to measure the effects of these speech presentation forms on the memorisation and comprehension processes on two user's populations: sighted and blind persons.

2. MARKUP LANGUAGES AND NON VISUAL PRESENTATION

2.1. Maturity of Text-To-Speech Technology

In the man-man communication, the prosodic variation conveys the speaker's affective disposition and his feelings. Within the talking machine, the prosodic variation could express the stress, the syntactic and the discourse structures. In an aural presentation system of structured documents, one problem is to find sound or speech equivalence called "sound[2] fonts" in this paper. They aim to represent typographic attributes as well as longer or higher for bold words than others for instance.

Text-To-Speech (TTS) technology including prosodic models is advancing rapidly: they start to offer possibilities for the prosodic[3] generation [Larrey 98]. One way is to use this new functionality to design "sound fonts" in a talking reading machine.
There is a reason to believe that its use will be widespread in the near future of the Information Society for nomadic access to information.

---

[2] *This concept is inspired by the prosodic font concept defined [Rosenberger 99].*

[3] *Prosody is defined as the tune, tone and rhythm of the speaking voice. The prosodic models are different according the communication: IVR, mail consultation, document reading, etc.*

Many studies measuring the performance –perceptual intelligibility, effects of use– of TTS have examined the perception and comprehension of speech technologies by listeners. For a taxonomy of the assessment of synthesis systems, see [Gibbon 97].
However given that this TTS technology will be used soon in the telematic services, the of the comprehension study of documents presented through synthetic speech is still an important and an open question.

Few studies address the listener's preferences for human versus synthetic speech and judgment tests about the "naturalness" of synthetic speech. [Stern 99] have studied the persuasiveness of synthetic speech produced by two versions of TTS systems (DECTalk and Monologue TTS systems) compared to "natural" (e.g. human) voice with 193 participants. They pointed out that many factors affect the degree of persuasion, for example the credibility of the person delivering the message and speech characteristics such as rate and rhythm. As the result, listeners rated human speech as softer, higher pitched than synthetic speech. Nevertheless, little difference was found between human and synthetic voice in degree of persuasiveness even if the human voice was rated more knowledgeable, more truthful and involved but less powerful than TTS voice. Based on these results, the authors suggest that improvements in the intelligibility, naturalness and emotiveness of TTS would not have much effect on how effective TTS speech would be in applications.

A possible solution is the use of markup language describing which speech parameters to apply for sentence. Hence, during the last five years, several studies are in progress to improve the emotion produced by these systems by using the markup languages.

2.2. Our previous work in this field

In 1995, effective methods of interacting and of viewing needed to be developed to enable blind people to use electronic information. The SMART user interface [Truillet 97] included facilities both to navigate and explore an HTML document by use of the multimodal presentation concept.

The SMART user interface interpreted the HTML DTD (Document Type Definition); this structure is mainly used to present information in a visual form. The structure enables the sighted user to have an overview of the document, as a visual parsing of the headers can do the job. We thought that the same "parsing" concept could be applied to the blind people's reading habits by synthesizing the headers.

The SMART system could present structured documents by means of a TTS system. The structure was presented by variations of the values of the three prosodic parameters –speed, intensity and pitch– according to relevant tags. For example, if a part of a text is in bold characters, the user interface translates this attribute by synthesizing this bold text at a slower speed for individual word emphasis. Default values of each prosodic parameter are linked to the various tags.

2.3. Dedicated Markup Languages

The availability of prosodic generation is a new challenge for the markup languages which aims to offer better non visual rendering. With the rapid development of cellular phones and the IVR enabled on Internet, several projects based of markup languages (derived from

SGML) aim to present structured information through TTS systems. [Dardailler 00] describes concepts to XML Accessibility and provides guidelines for new DTD, oriented for multimedia rendering. In [Noonan 00], the author illustrates the fundamental differences between designing visually oriented and speech oriented applications. You agree with the recommendations proposed by the author, due mainly to the visual and auditory modalities.

As examples, we can mention SSML (**S**ynthe**S**is **M**arkup **L**anguage) [SABLE 00] from Bells Laboratory, VoiceXML [VoiceXML 00] (developed in partnership with IBM, ATT, Motorola and Lucent Technologies), W3C Voice Browser Activity and Aural CSS [W3C 00] or Extended Cascading Style Sheets [Truillet 99b] are some projects based on XML language or HTML able to present "enhanced aural information".

```
<!doctype ssml system "SSML.dtd" []>
<ssml>
SSML allows explicit labelling of text.  Just press the
<emph>start</emph> button. Also phrases can be marked
in text.  Even in utterly <phrase> inappropriate places </phrase>

<voice name="male2">
Different voices, as well as different languages may be selected by
another simple tag.

<voice name="male1">
<define word="edinburgh" phonemes="e1 d - i n - b r @">
Also desired pronunciation of words like Edinburgh
can be explicitly given. So the pronunciation is correct
<sound src="bong.au"> and not wrong <sound src="splat.au">
</ssml>
```

Figure 1: A SABLE fragment (SSML markup language)

As illustrated in figure 1, with SSML you can specify emphased word (tag `<emph>`) and define appropriate pronunciation of words, etc.
These markup languages rely on a DTD which defines the speech parameters to use.

Even if technical aspects are available to improve "emotion" of TTS systems, few studies measure the effects in terms of memorisation and comprehension processes of such presentation. In fact, is this presentation really efficient? And if the answer is affirmative, in which situation?

2.4. Discussion

It seems that the degree of a system's intelligibility is certainly important in terms of the user's ability to understand the utterances produced by the system. Our hypothesis work relies on that it is possible to find an alternative presentation form for a bold sequence of words in a auditory form. This study is limited to the research of "sounds fonts" for bold attribute.

3. THE EXPERIMENT: METHOD

We address a number of specific questions in conjunction with the general issues of how to bring word into "salience". First, we expect that blind users will better memorise a word in

bold more easily than the other users. Second, we expect that because of an emphase prosody model of the TTS speech for the bold word, listeners will be more persuaded by this version compared to the standard TTS system.
The issue of this study is to define a prosodic model adapted to the presentation of typographic attributes.

3.1. Users

33 graduate students and 15 blind users were selected. These subjects should have a good audition and be non accustomed with vocal technologies. Each third of them was affected to one of the three aural presentation text (see below 3.3).

3.2. Materials

The experimental platform consists in:

- A multimedia computer which allow the presentation through a text-to-speech system [Elan 2000]. Users listened the message through a set of commercial quality speakers;

- A DAT (Digital Audio Tape) with a microphone to record the answers of the subjects

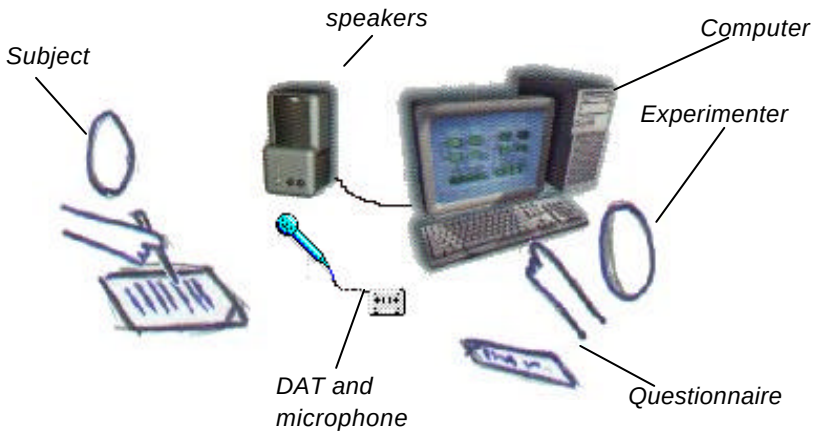

Figure 2: Experimental platform.

3.3. Stimulus material

We used the text "Le vieil homme" ("*The old man*") [Truillet 99] composed with the contribution of neuro-linguists of the "Jacques Lordat" laboratory. This text is composed of 196 words linguistically calibrated as shown in Figure 3.

Un vieil homme acariâtre, qui vivait seul depuis toujours et qui allait avoir soixante quatorze ans en décembre, ne supportait pas les enfants. Il habitait une maison entourée d'un jardin bien entretenu pour son plaisir, et avait à portée de la main, dans son entrée, une canne en bambou dont il menaçait les enfants turbulents de la cité HLM voisine.
Un mardi, alors qu'il venait de détruire un nid de guêpes, il s'est retrouvé coincé sur le toit haut de trois mètres cinquante. Car, en voulant redescendre très vite, il a fait tomber l'échelle en alu qu'il avait posé en équilibre instable contre le mur de l'appentis. Comme l'homme s'est mis à appeler à l'aide d'une voix forte, un gamin courageux qui jouait sagement

aux billes dans la rue, le long de la clôture, a levé la tête, a compris la situation et a replacé l'échelle qui était par terre, à côté d'un rosier. Depuis cette fâcheuse aventure, le dimanche, il invite son sauveur blond dans son jardin et, pour le remercier, lui offre sous les arbres un goûter accompagné de jus de pomme.

Figure 3: Text "Le vieil Homme"

The previous text was presented through the TTS. The Robert's male voice and the default values for speech output were used. To measure the effect of the aural presentation of the bold attribute, the previous text is presented through different modalities of aural presentations:

- **A "neutral" version** called A with the default values;
- **A verbalised version** called E: Each salient word is presented by addition of a verbal description of the typographic attribute ("in bold" with a decrease of 15% of the current pitch is said before the salient word);
- **And a prosodic version** called E': Each salient word is pronounced with an increase of 13% of the default pitch.

Several preliminary trials with various values of pitch, energy and speech rate were proposed to listeners. The pitch variation was retained as the more discriminant parameter.

3.4. Dependent measures

To measure the identification of salient words, through the three versions, two understanding exercises were designed. These types of exercises are usually used to evaluate the comprehension process of the users. The objective of the first exercise (*free recall*) is to measure the impact of the salience in the comprehension/memorization processes. The objective of the second exercise (*indexed recall*) is the same but allow users to recall words not spontaneously recalled in the previous exercise.

First, we defined ten words on which salience is applied. These words are spatially well-distributed in the text. Each word is mono or bi-syllabic and may be not relevant for the global comprehension of the text according to the [Cadilhac 97] study.
Here is the list of these terms: *maison* (home), *cité* (estate), *toit* (roof), *sauveur* (saviour), *billes* (marbles), *mur* (wall), *entrée* (entrance), *jardin* (garden), *alu* (aluminium) and *pomme* (apple).

3.5. Procedure

Participants were told that the experiment concerned the topic of comprehension during aural presentation. Four different tasks complete the procedure:

Task 1 (Dictation): A dictation was used to train users with TTS systems and control the auditory performance before the following experimental tasks.

Task 2 (Experimental task): The assigned text (A, E, or E' version) was presented once with the TTS.

Task 3 (Free recall protocol): 5 minutes after hearing the assigned text, subjects were invited to tell the story which was recorded on a DAT device. Each story is analysed to identify the salient words named.

Task 4 (Indexed recall protocol): subjects were asked to answer 20 questions in order to determine their memorisation degree.
These questions are based on the 10 words put in salience and on 10 words not relevant for the global comprehension according to the [Cadilhac 97] study. At each word corresponds a question on a precise element of the text.

4. RESULTS AND OBSERVATIONS

The primary assessment for this study was the comparison between recalls for each version (Table 1.)

		Free recall			**Indexed recall**	
	Count	*Mean*	*Std. Dev.*		*Mean*	*Std. Dev.*
Blind, v. A	5	3.000	2.000		4.200	2.168
Blind, v. E	5	2.400	1.817		5.200	3.701
Blind, v. E'	5	3.200	1.304		5.800	1.643
Sighted, v. A	11	2.273	1.348		4.455	1.508
Sighted, v. E	11	2.455	1.572		4.364	1.912
Sighted, v. E'	11	3.364	2.203		4.909	1.868

Table 1: Means and standard deviations for recall according to the type of subjects and versions of text.

Versions: Table 1 shows that there is no significant difference between neutral and enriched versions (mean difference between version A and version E=-0.250, p=0.7334; mean difference between version A and version E'=-0.813, p=0.4447 respectively). Nevertheless, we note an important deviation in recalls between subjects.

Blind/sighted: It appears that there is no significant difference between recalls by blind and sighted (mean difference: 0.170, p=0.7560). However, Figure 4 and Figure 5 show some difference in percentage for recalls between them.

First, for blind, percentages of indexed recall show that enriched versions seem to be better than neutral version (52 and 58% versus 42% for version E, E' and A respectively).

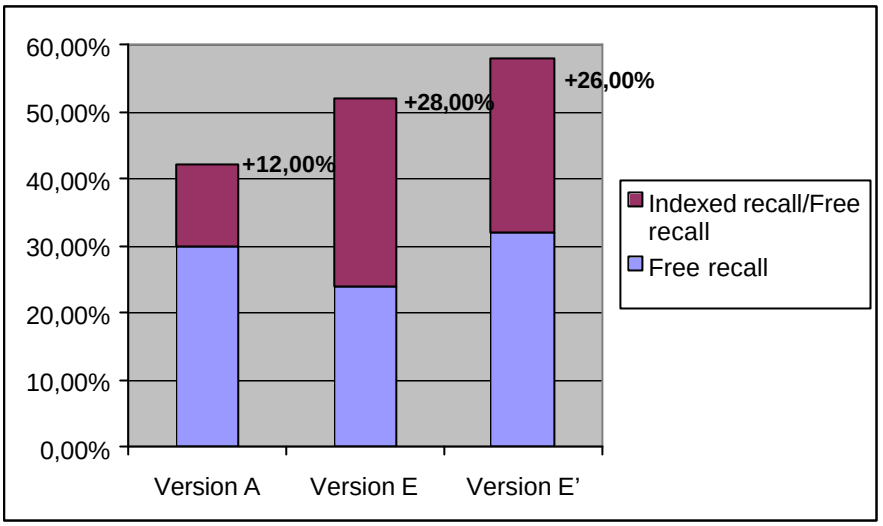

Figure 4: Mean of recalls for blind.

Then, for sighted, percentages of indexed recall show than the prosodic version seems to be better than both verbalized and neutral version (49.09% versus 43.64 and 44.55% for version E', E and A respectively).

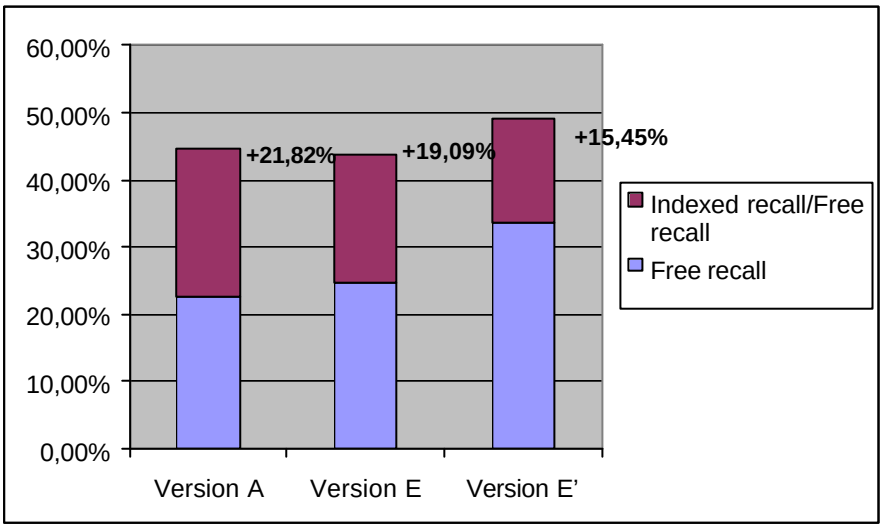

Figure 5: Mean of recalls for sighted.

Even if there are no significant results yet between versions, we think that some differences exist between recalls of the prosodic and the neutral version; standard deviations show that results are perhaps simply too "noisy".
As an encouraging result, means show us that both verbalized and prosodic versions don't damage recalls in comparison with neutral version.

5. CONCLUSION

The experiment presented in this paper is an initial effort in evaluating "sound fonts. We are following up this research with a more focused search for significant difference between neutral version and enriched versions of TTS". This study must be also extended to compare

the results obtained for oral presentation to those obtained for visual presentation. Under the hypothesis, that there is no degradation during the text comprehension, the use of "sound fonts" is a new challenge for all application based on non visual interaction. New investigations will be pursued along several axes: 1) observation to a great many observed users to have better traceable results, 2) observation to another population of users as elderly persons on WebTV, 3) for tactile presentation on Braille displays.

REFERENCES

[Cadilhac 97] C. Cadilhac, *Des structures textuelles à leur traitement : compréhension et mémorisation d'un récit par déments de type Alzheimer et sujets normaux âgés*, Thèse de doctorat de l'université Toulouse II, Décembre 1997.

[Dardailler 00] D. Dardailler, XML and Accessibility, in *CSUN 2000 Proceedings*, `http://www.csun.edu/cod/conf2000/proceedings/0103Dardailler.html`

[Elan 00] Elan Informatique, `http://www.elantts.com`

[Gibbon 97] D. Gibbon, R. Moore, R. Winski., *Handbook of Standards and Resources for Spoken Language System*, Edited by Gibbon D., Moore R., Winski R., Mouton de Gruyter, 1997.

[Larrey 98] P. Larrey, N. Vigouroux, G. Pérennou, Generating Spoken Utterances from Concepts and Prosodic Schemes, *Proceedings of the 1st Workshop on Text, Speech, Dialogue*, TSD'98, Brno, September 1998, pp. 269-274.

[Noonan 00] T. Noonan, A Strategy and Information on Design User-Friendly Automated Telephone Services Which incorporate Text-to-Speech, *in CSUN 2000 Proceedings*, `http://www.csun.edu/cod/conf2000/proceedings/0190Noonan.html`

[Rosenberger 99] T. Rosenberger, R. L. Mac-Neil, Prosodic Font:Translating Speech into graphics, *Proceedings of CHI'99 Extended Abstracts*, Pittsburgh, May 1999, pp. 252-253.

[Roth 99] P. Roth, L. Petrucci, Th. Pun, A. Assimacopoulos, Auditory browsers for blind and visually impaired persons, *Proceedings of CHI'99 Extended Abstracts*, Pittsburgh, May 1999, pp. 218-219.

[SABLE 00] SABLE Synthesis Markup Language `http://www.research.att.com/~rws/Sable.v1_0.htm`

[Savidis 00] A. Savidis, C. Stepahnidis, *Development requirements for Implementing Unified Interfaces*, In C. Stephanidis (Ed.), User Interfaces for All (Chapter 22), Lawrence Erlbaum Associates, Mahwah, NJ, ISBN 0-8058-2967-9, 850 pages.

[Stephanidis 00] C. Stephanidis, Universal Access Through Unified User Interfaces, *in CSUN 2000 Proceedings*, `http://www.csun.edu/cod/conf2000/proceedings/0243Stephanidis.html`

[Stern 99] St. Stern, J. Mullenix, C. Dyson, St. Wilson, The persuasiveness of Synthetic Speech versus Human Speech, *in Human Factors*, vol. 41, December 1999, pp. 588-595.

[Truillet 97] Ph. Truillet, B. Oriola, N. Vigouroux, Multimodal Presentation as a Solution to Access a Structured Document, *Electronic proceedings of 6th World-Wide-Web Conference*, Santa-Clara, April 1997.

[Truillet 99] Ph. Truillet, N. Vigouroux, Étude des effets de la mise en forme matérielle de textes par des présentations non visuelles, *Proceedings of CIDE'99*, Damas, July 1999, pp. 263-277.

[Truillet 99b] Ph. Truillet, N. Vigouroux, Non Visual Presentation of HTML documents for Disabled and Elderly Using Extended Cascading Style Sheets, *5th ERCIM UI4ALL Workshop*, UI4ALL Workshop Dagstuhl, 29 November - 1st December 1999, pp. 159-166

[Vivier 97] J. Vivier, M. Mojahid, Adaptation de la mise en forme matérielle au lecteur : collaboration pluridisciplinaire à une modélisation différenciée selon les objectifs de lecture, Atelier "Texte et Communication" – *Le texte procédural : langage, action et cognition*, PRESCOT, E. Pascual, J-L. Nespoulous, J. Virbel Eds, Mai 1997

[VoiceXML 00] VoiceXML, `http://www.voicexml.org`

[W3C 00] W3C Voice Browser Activity, `http://w3c.org/Voice`