Voice Interfaces in Art – An Experimentation with Web Open Standards as a Model to Increase Web Accessibility and Digital Inclusion

Martha Carrer Cruz Gabriel

University of Sao Paulo, University Anhembi Morumbi, R. Ibaragui Nissui 115 #1204, 04116-200 São Paulo, SP, Brazil
martha@martha.com.br

Abstract. The web has been largely mute and deaf but since the beginning of the 21st century this scenario is changing with the possibility of using intelligent voice interfaces on web systems. In this paper we present the *Voice Mosaic* – a system that allows voice interactions on the web through the telephone. Its voice interface uses speech recognition and synthesis solutions developed with VoiceXML, an open-standard in voice technologies adopted by the W3C. *Voice Mosaic* is an artwork that allows people to get in touch with the possibility of talking to the web, intending to cause awareness about it. Since the technology used in *Voice Mosaic* can be used to improve accessibility (for visual impaired people) and digital inclusion (since the telephone is one of the cheapest devices in the world), dissolving borders and amplifying the pervasiveness, we believe that the concepts presented here can be useful to other developers.

Keywords: voice, web, interface, hybridization, telephone, accessibility, digital inclusion.

1 Introduction

Voice interfaces are a fascinating subject. The human dream of talking to computers in a natural way is not new. Science fiction books and movies that live in our imagination present several examples of this aspiration, as old television and movie series like "Star Trek," where the Enterprise's staff talk to the ship systems and androids like commander DATA; "Lost in Space," where Will Robinson had in his robot a very loyal and confident friend; the conversations and human interactions with the robots C3PO and R2-D2 in "Star Wars"; "Blade Runner" and its androids and voice driven interfaces; among others [3].

Until recently, talking to computers was in the realm of fiction – the web has been largely mute and deaf. However in the beginning of the 21st century talking to computers has become possible and easy due the enormous advances in speech synthesis and voice recognition technologies as well as the open standards adopted by the W3C (such as VoiceXML). The accuracy level reached by voice technologies now has allowed us to use them widely on the web.

C. Baranauskas et al. (Eds.): INTERACT 2007, LNCS 4663, Part II, pp. 669 – 672, 2007.

The potential of using voice interfaces is explosive. From speech-only applications integrated to the whole web, to multi-modal applications combining aural and visual abilities into web browsers, voice interfaces add to the flavor of the web a fundamental spice, which is surely going to impact it.

Tim Berners-Lee said at SpeechTEK 2004, NY- "Speech technology is an important ingredient for the Web to realize its full potential." In fact, voice interfaces on the web bring undeniable resources for several areas, as convenience for mobile users, v-commerce, natural interactions, and usability. Beyond the more obvious utilizations for voice interfaces, the ability to talk to the web also provides an important way to improve web-accessibility – not only by multi-modal applications, but also through speech-only ones. Besides that, speech-only applications liberate users from any client computer device to access the internet – in this case, all they need is any telephone in any place in the world. In this sense, since the telephone is one of the cheapest devices in the world, voice interfaces can help improving digital inclusion. This is the alliance of the widest computing network with the most pervasive communication device on Earth – internet & telephone.

However, talking to computers adds "ears" and "mouths" to the Internet organism, changing the way we interact with it, bringing new possibilities and new challenges as well. We must face the increasing complexity that voice interfaces bring to the web while we also open new channels for digital inclusion, provide more accessibility and increase mobility through voice. All these things affect the human role inside the high-tech social structure we live in, at once causing excitement and fear.

In this context, in 2004, it was created the *Voice Mosaic* – a web-art work that allows voice interactions on the web through the telephone, causing border dissolution between Internet and telephone. As said once by Hendrik Willem Van Loon [1], "*The arts are an even better barometer of what is happening in our world than the stock market or the debates in congress*." and we believe that artworks help people to understand and experience the new emergent techno-social world that surround us, where convergence and hybridization have become ubiquitous and easy, and "to talk to computers or the web" is going to become common.

Since the technologies used in *Voice Mosaic* can be used in other kinds of voice applications on the web, improving accessibility and digital inclusion, we will present next the work and its main aspects, regarding either the art concept or the technological implications. This artwork received several awards and was also presented at SIGGRAPH Art Gallery 2006, in Boston, MA (USA).

2 Voice Mosaic

The *Voice Mosaic* is a web-art application that combines speech and image, building a visual mosaic on the web with the chosen colors and recorded voices of people who interact with it from any place in the globe. The voice interface, developed with open-standards in speech synthesis and voice recognition technologies (VoiceXML), works through phone calls from any telephone – mobile or not. To participate in English, call in US: (800) 289-5570 or (407) 386-2174 **/** PIN number: 9991421055 (to participate in Portuguese, call in Brazil: (11) 2122-0203 / application code: 1155723602). The mosaic is accessed on the web at www.voicemosaic.com.br.

The application was developed in 2004, in three languages – Portuguese, English and Spanish - in order to encourage global participation. The phone calls form the mosaic on the web, and it happens spontaneously, therefore the mosaic changes as time goes on and its ongoing aesthetics and final result are unpredictable.

In this context, the work causes time-space collapse, and maps in one screen the participations that comes from several different geographical places, in different languages, and different times. Furthermore, using the search field, one can easily locate his/her participation by searching his/her own phone number. Also, one can locate all tiles in the mosaic within the same telephone area, which means to map geographical participations in the visual work.

The work puts together several dualities that do not oppose each other, but complete each other: speech / image, simple / complex, old / new, low-tech / high-tech, time / space, individual / community, passive / active, expected / uncertain, among others, in order to cause reflection and awareness about talking to the web, media convergence and hybridization between the telephone and the web.

2.1 Interfaces and Technology

The work has two interfaces – the voice interface accessed by phone and the web interface. As the web interface uses common and well known technologies – html, data base and Flash --, we will focus here on the voice interface, which is the core of the system.

The voice interface works via phone (mobile or not) interacting with the web. It is developed with VoiceXML, a structured language that offers support to build dialogs. When accessed by phone, the interface uses a Voice Gateway which allows voice recognition and speech synthesis during the conversation.

During the interaction by phone the person talks to the interface, choosing a color and recording a free speech message.

There are seven options available for choosing the color. This number, seven, is due the limit of information that a person can hold in the short-term memory. According to Miller [2] and explained in Zakia [4], “There is a limit to the amount of unrelated information a person can hold in short-term memory (STM), from five to nine items, averaging seven. (…) Since we are limited in the amount of information we can retain correctly in STM, one should be cautious with the amount of information included in a multimedia program if it is going to have some memorable impact”.

The free speech message is limited to 15 seconds because of the web interface where it will be listened – recorded files longer than 15 sec. would generate WAV files larger than 100kb, which is the maximum file size to allow a comfortable user experience while clicking and listening to the mosaic tiles without waiting too long to start playing.

The voice interface was designed using both pre-recorded human voice (in the welcome message) and synthesized text-to-speech voices to instruct the user, in order to cause the experimentation of the differences and similarities between them. Also, it is used touch tone and speech tone interactions in order to put side by side voice recognition (human-like feature) and touch recognition (machine-like feature)

intending to cause reflection about the two ways of interacting by phone – talking and dialing.

In order to allow data visualization either by tracking or by locating the interactions in the visual mosaic, the voice interface records the Caller ID phone number. Due that we can know where the interactions come from in the globe and also locate all the interactions from within a specific area code. This reveals the space collapse in the mosaic on the web.

The phone calls, through the voice interface, are the way the data (and people) enter the *Voice Mosaic* on the web. No data enters the work via its web interface, which is used only for purposes of data visualization, interpretation and reflection.

3 Conclusion

The web and telephone have been the realm for the state of the art in voice technologies.

Voice Mosaic is on the web, and it has received voice participation for more than two years now, summing up about 800 tiles. Although we could realize that people do not know much about the technology they are experiencing in the work, they use it easily and get excited about "talking to the web" and becoming immediately a permanent tile there. We also realized that technical people (IT, engineers, etc.) were more resistant to first experiment with the work than lay people. The kind of messages people create is also interesting – they range from recorded music and people singing to love declarations and creative use of the voice.

From now on we think that it will be possible to provide wider and deeper experimentation with voice interfaces due to the available technologies integrating the web and telephone. We expect it will probably allow us all to break frontiers and go further in human accessibility and digital inclusion developments.

An interactive poster was created based on this paper intending to show the concepts involved in *Voice Mosaic's* development and also to present the conclusion of this work, encouraging the audience to experiment the voice interface via phone and the web interface in order to check the results.

References

1. Loon, H.W.V.: The Arts (1937)
2. Miller, G.: The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. Psychological Review 63, 81–97 (1956)
3. Perkowitz, S.: Digital People: From Bionic Humans to Androids. Joseph Henry Press, Washington (2004)
4. Zakia, R.: Perception and Imaging. Focal Press (1997)