Voice interfaces in electronic art

Martha Gabriel
University of Sao Paulo
R. Ibaragui Nissui 115 ap 1204 – Sao Paulo – SP – Brazil
martha@martha.com.br

Talking to computers is an old human dream but the advances in the speech synthesis and voice recognition technologies in the past decade have reached enough accuracy and reliability to help making that dream come true. Voice interfaces can be applied in a big range of scenarios, from art to medicine, and one of the most important uses of voice interfaces is related to accessibility and usability, opening options for interactions that bring possibilities where vision cannot be used or does not fit the environment. Regarding the arts, the conjunction of the internet and the open standards of VoiceXML create a new context for exploration and experimentation with voice interfaces. This paper describes the voice interfaces scenario from the point of view of art. From the first known artwork using voice to the state-of-the-art we have today, voice has increasingly being used as interface. We will describe some artworks that are part of that history and briefly present the use of VoiceXML in the construction of the voice interface of the artwork Voice Mosaic.

Voice. Interface. Web. Art. Voice recognition. Speech synthesis. Phone, HCI.

1. INTRODUCTION

Voice interfaces, whether they are speech-only or multi-modal, 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. We could mention, for instance, some television shows and movies like: a) *Star Trek* – the Enterprise's staff talk to the ship systems and androids like commander DATA; b) *Lost in Space* – Will Robinson had in his robot a very loyal and confident friend; c) *Star Wars* – conversations and human interactions with the robots C3PO and R2-D2; d) *Blade Runner* – the androids and voice driven interfaces (Perkowitz, 2004).

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.

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, as said once by Hendrik Willem Van Loon (Loon, 1937), '*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 surrounds us, where convergence and hybridization have become ubiquitous and easy, and 'to talk to computers' is going to become common.

We therefore suggest that we can take advantage of paying attention to artistic expressions in order to get valuable information. In this sense, analyzing voice technologies and the types of voice interactions used in art can be a good way to study their evolution and changes in creative utilizations. The first known electronic artwork dealing with voice technologies emerged at the end of the 20th century (Gabriel, 2006), and since then we have seen several other interesting creations in art using state-of-the-art technologies that utilize voice recognition and synthesis on the web and telephone.

The objective here is to show a brief panorama of artworks related to voice technologies focusing on

the work *Voice Mosaic*, – piece that allows voice interactions on the web through the telephone, dissolving borders and amplifying the pervasiveness. The work was developed on the web using a voice interface with speech synthesis and voice recognition technologies based on VoiceXML. This artwork was exhibited around the world and received several awards. The VoiceXML technology used to its development is the open standard recommended by W3C for voice applications. Therefore, in the end, this paper will present the main aspects of VoiceXML using the Voice Mosaic development as base for that.

2. ELECTRONIC ARTWORKS & VOICE

According to Gabriel (2006), the following artworks draw a brief panorama of interesting relationships between electronic art and voice, contextualizing the artistic development from the beginning (in the 1980s) up to now.

1986 – Synthetic Speech Theatre, by Stephen Wilson
1989 – Barbie Liberation, by Ron Kuivila
1990 – Talking Machine, by Martin Riches
1991 – Inquiry Theater, by Stephen Wilson
1994 – Oh toi qui vis la-bas, by Don Ritter
1994 – Speech Sculptures, by Bruce Cannon
1995 – I Have Never Read the Bible, by Jim Campbell
1996 – Le Pissenlit, by E. Couchot and M. Bret
1996 – Orpheus, by Ken Feingold
1999 – Universal Translator, by David Rokeby
2000 – Giver of Names, by David Rokeby
2000 – Huge Harry, by Arthur Elsenaar, Remko Scha
2000 – Millennium Venus, by Sharon Grace
2000 – n-Cha(n)t, by David Rokeby
2000 – netsong, by Amy Alexader
2000 – Riding the Net, by C. Sommerer and L. Mignonneau
2000 – Talk Nice, by Elizabeth Vander Zaag
2001 – Living Room, by C. Sommerer and L. Mignonneau
2002 – RE:MARK, by G. Levin and Z. Lieberman
2003 – Alert, by Barbara Musil
2003 – Messa di Voce, by Golan Levin et al
2003 – Summoned Voices, by Iain Mott
2003 – Universal Whistling Machine, by Marc Böhlen and JT Rinker
2004 – IP Poetry, by Gustavo Romano
2004 – Voice Mosaic, by Martha Gabriel
2005 – Organum, by Greg Niemeyer et al
2005 – Tampopo, by Kentaro Yamada
2007 – NetAura, by Giselle Beiguelman

These artworks use voice ranging from whistles and blows to speech synthesis and voice recognition. In order to track the involvement of

voice technologies in electronic art we will present some works that are very creative or innovative. The information and images of each work were extracted from its respective website and the URL is listed in the references.

2.1 Artworks that implement non-verbal voice as an input

2.1.1 *Le Pissenlit* (Couchot, 1996) – art installation

Main characteristic – blow in a digital image.

The principle of this work consists of blowing a digital image – a dandelion flower – that dissolves in different ways depending on the way it is blown (see Figure 1). Although this work does not specifically use the voice itself, it involves breath, an integral part of generating vocal sounds. A newer version of this work from 2005 uses not only the blow itself but also the sound of blowing as input.

Figure 1: *Le Pissenlit* (Couchot, 1996)

2.1.2 *Universal Whistling Machine* (Böhlen, 2003) – art installation

Main characteristic – whistle synthesis and recognition.

'Whistling is a communication primitive in most human languages – it is a kind of time travel to a less articulated state.' (Böhlen, 2003). Based on that, this artwork proposes the use of whistling as an alternative way for human-machine interface design by allowing whistle synthesis and recognition. Whistles can be considered non-verbal vocal manifestations and this work is particularly interesting because it investigates new forms of phoneme-less vocal interfaces. 'Whistling is much closer to the phoneme-less signal primitives compatible with digital machinery than the domain of spoken language.' (Böhlen, 2003). Furthermore this work raises questions not only about human-computer interactions but also about human-animal communication (see Figure 2).

Figure 2: *Universal Whistling Machine* (Böhlen, 2003)

2.1.3 Universal Translator (Rockeby, 1999) – art installation

Main characteristic – speech recognition and phoneme analysis.

In this work a microphone records the subject's voice and a live camera records his/her mouth. The voice is processed by a computer, and then audio is output right away according to the phonemic content and vocal intensity of the input sound. A video with the subject's mouth is shown as a secondary aspect of the work (see Figure 3). Since this work focuses on the phonemes and on the non-verbal aspects of the audio input, its results can be interesting to analyze beyond traditional word recognition.

Figure 3: *Universal Translator* video still (Rockeby, 1999)

2.1.4 Talk Nice (Zaag, 2000) – art installation

Main characteristic – speech recognition focused on pitch.

'Talk Nice is an interactive video installation. By exploring the use of upism in young women, the piece explores the way that people, especially young women, frame themselves by speaking up at the end of a sentence. The interactor is coached to talk up in order to interact with the teens (on the video)' (Zaag, 2000). By analyzing the non-verbal aspects of the interaction such as the upism, this work targets a specific social group and the experience of interacting with it is very successful. It also causes awareness about social characteristics related to non-verbal ways of communicating using voice.

2.2 Artworks that employ the verbal dimension of voice – speech synthesis

2.2.1 netsong (Alexander, 2000) – web art

Main characteristic – speech synthesis creating a song based on the web links provided by a search on the web.

This work performs a song based on a search engine robot. 'When provided a search term, the netsong bot will search for this term in a search engine, then choose a page from the search results and begin following links from that page. (...). Gathering text from each page it visits, the netsong bot savours the lyricality and poignant narrative of the web and begins to sing it.' (Alexander, 2000). Besides probably being the first artwork to use speech synthesis on the web, another important thing to notice is its relationship with search engines – one of the most important and influential interfaces of our Digital Era.

2.2.2 Talking Machine (Riches, 1990) – art installation

Main characteristic – speech synthesis created by a physical system that imitates the human vocal system.

'Talking Machine is an acoustic speech synthesizer. The speech sounds are produced using a flow of air and resonators just as in natural speech. The machine has 32 pipes, each one a simplified version of the human vocal tract (see Figure 4). They reproduce the spaces which are formed in the mouth, nose and throat when we speak. (...) The valves which control the flow of air are operated by a computer' (Riches, 1990). This work is particularly interesting due its nature of imitating the human speech system in order to synthesize the voice, using the blow as the agent of the process as it happens in the human body.

Figure 4: *Talking Machine* (Riches, 1990)

2.3 Artworks that employ the verbal dimension of voice – speech recognition

2.3.1 *Mesa di Voce* (Levin, 2003) – art performance

Main characteristic – algorithm that analyses speech transforming it into images.

'*Mesa di Voce* is an artwork concerned with the poetic implications of making the human voice visible. A computer uses a video camera in order to track the locations of the performers' heads, and also analyses the audio signals coming from the performers' microphones. In response, the computer displays various kinds of visualizations on a projection screen behind the performers' (Levin, 2003) (see Figure 5). Besides the beauty and art, another remarkable aspect of this work is that the synthesized visualizations are tightly coupled to the sounds spoken and sung by the performers connecting voice-input (verbal and non-verbal) and images.

Figure 5: *Mesa di Voce* Performances (Levin, 2003)

2.3.2 *Inquire Theater* (Wilson, 1991) – art installation

Main characteristic – speech recognition interpreted by a virtual navigation system.

'*Inquiry Theater* is an installation where participants could take a virtual walk down Mission Street in San Francisco's ethnic Mission neighborhood. Speech recognition determined direction of movement and virtual entry into the stores.' (Wilson, 1991). Although this work uses verbal speech recognition, in this work the answer is given through navigation. Besides, it was one of the first works to use speech recognition in electronic art.

2.3.3 *Voice Mosaic* (Gabriel, 2004) – web art

Main characteristic – speech recognition and synthesis in an interface by phone integrated in real-time with the web.

The *Voice Mosaic* is a web-art application developed in three languages – Portuguese, English and Spanish – that converges speech and image, building a visual mosaic on the web (see Figure 6) with the chosen colours 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. The mosaic is seen/heard on the web – <http://www.voicemosaic.com.br>.

Figure 6: *Voice Mosaic* screenshot in Jan/2007

As people make phone calls to participate – choosing colours and recording free messages – they form the mosaic spontaneously and it changes as time goes on. The ongoing aesthetics and final result are unpredictable.

In this context, the work causes time-space collapse, and maps in one screen the participations that come 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.

Besides the work being available online for interaction, there are also two videos on Youtube about the Voice Mosaic, explaining the artwork – <http://www.youtube.com/watch?v=YUURctJYckM> & http://www.youtube.com/watch?v=c_5tfTg8NqY.

Next, we will focus on the development of Voice Mosaic using VoiceXML.

3. VOICE INTERFACES

The artwork Voice Mosaic has two interfaces – the voice interface accessed by phone and the visual/aural web interface. As the web interface uses common and well known technologies, we will focus here on the voice interface, which is the core of the system.

The voice interface works via telephone (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. In the image below (Figure 7) we can see the difference between accessing the application via phone and via web:

Figure 7: Accessing the web by phone using VoiceXML

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

There are seven options available for choosing the colour. This number, seven, is due the limit of information that a person can hold in the short-term memory. According to Miller (1956) and explained in Zakia (1997), '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 both. 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 dialling.

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.

4. CONCLUSION

While internet is becoming the most important communication via in the world, generally speaking the web has been practically deaf and mute so far – it seems that it still has not been able to deal properly with speech, be it natural language processing or simple voice commands. On the other hand, voice technologies have only reached enough accuracy and reliability to be used in large scale at the beginning of the 21st century, bringing to the surface the possibility of finally using them on the web.

In this context, the VoiceXML language can be used as an open standard for developing voice interfaces and Voice Mosaic is an artwork that allows people to experiment and reflect about the possibility of 'talking to the web', its new benefits and complexities.

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 artistic/human possibilities and developments.

5. REFERENCES

Alexander, A. (2000) netsong. <http://netsong.org>

Böhlen, M. and Rinker, J. T. (2003) Universal Whistling Machine.
http://www.realtechsupport.org/new_works/uwm.html

Couchot, E. and Michel, B. (1996) Le Pissenlit.
<http://www.artmag.com/techno/landowsky/projet.html>

Gabriel, M. (2004) Voice Mosaic.
<http://www.voicemosaic.com.br>

Gabriel, M. (2006) *Interfaces de Voz em Ambientes Hipermidiáticos*. Master's Degree Dissertation, University of São Paulo, Brazil.

Levin, G., Lieberman, Z., Blonk, J., and Barbara, J. (2003) Mesa di Voce.
<http://tmema.org/messa/messa.html>

Loon, H. W. V. (1937) *The Arts*.

Miller, G. (1956) The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information. *Psychological Review*, 63, pp. 81–97. 1956.

Perkowitz, S. (2004) *Digital People: From Bionic Humans to Androids*. Washington: Joseph Henry Press.

Riches, M. (1990) Talking Machine.
<http://www.floraberlin.de/soundbag/index53.html>

Rockeby, D. (1999) Universal Translator.
<http://homepage.mac.com/davidrokeby/trans.html>

Wilson, S. (1991) Inquire Theater.
<http://userwww.sfsu.edu/%7Eswilson/>

Zaag, E. V. (2000) Talk Nice.
<http://www.canadacouncil.ca/news/releases/2000/un127241852363281250.htm>

Zakia, R. (1997) *Perception and Imaging*. Focal Press.