We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

17

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

Martha Gabriel
University of São Paulo,
Brazil

1. Introduction

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.

2. Voice Interfaces

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 (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.

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.

2.1 Voice interfaces characteristics

Voice interfaces are specialized computational systems that allow that dialogs happen between human beings and computers (other computational systems) in a way that the computational commands be synthesized in voice in order to be understood by humans, and the human speeches be recognized and transformed into computational codes by the computers. In this way, for instance, instead of accessing a visual page on the web via a browser to fill in a form to book a flight, one could do it via a voice interface, talking to the page.
Voice interfaces are exclusive in a way since they are based on the spoken language. The oral communication plays a big role in the human daily life. Since the youngest ages we spend a substantial part of our waken hours in conversations (Cohen 2004: 7).
According to Pinker (2002: 10), language is an instinct – a biological adaptation to convey information. The idea of the language as a kind of instinct was mentioned for the first time by Darwin in 1871 and in the 20th century, the most famous thesis about the language as an instinct was created by Noam Chomsky (Pinked, 2002: 14). Being the language a natural instinct, it is no wonder that since the machines exists in the human imagination, the utilization of the natural language to talk to them is a latent desire. According to Wilson:

> "The ability of producing and understanding the spoken language was identified by anthropologists as one of the main realizations of our species. Other animals, like dolphins and the primates, may have significant capabilities of vocal communications, but they not get close to the human capabilities." (Wilson, 2002: 775)

However, in despite of the fact that to speak is the most natural and human way of interacting, to access the internet via voice interfaces is as different from navigating on the web via a visual browser as to talk on the telephone is different from reading a letter.

Nowadays we are used to 'browse' and 'write' on the web, which is very different from 'talking' to the web.

Thinking about the differences from visual and aural interfaces, we can start listing the particularities of the voice characteristics. The first one is the transiency. As soon as it is spoken or listened to, the voice disappears and demands that we remember what has been said. On the other hand, visual elements are persistent:

> "The voice is an one-dimensional media with zero persistence. The computational monitor is a bi-dimensional media that combines persistence (you can look to them as long as you want) with selective actualization (you can type a value in any field of the screen without changing the rest of it." (Nielsen, 2003).

Adding to that, according to Santaella:

> "The first principle of sonority is in its evanescence, something that the passage of time makes disappear (…), and the first principle of visuality is in the form that makes itself present before our eyes." (Santaella, 2001: 369).

The second particularity of the voice is the invisibility. That makes it more difficult to indicate to the user which options he can execute and what he needs to say in order to execute them. In visual interfaces, we can always have visible menus and instructions that follow each step of the process in execution.

The third particularity is the asymmetry of the voice. The voice can be produced much faster than being understood; an user can speak faster than typing; and an user can listen slower than reading. In visual interfaces, the user has his own interaction pace to synchronize before continuing the process. In voice interfaces, the pace is not always controlled by the user, but by the interface.

We could say that, according to Cohen (2004: 6), there are two possible modalities of voice interfaces regarding the human senses used – 1) purely aural: where all the process occurs only via sounds and the orality of the speech, without using any visual support, and; 2) multimodal: where the process of interfacing via voice is supported by some kind of visual system associated to it. In the first case, pure voice interfaces, we can mention as an example the access to a system via telephone (see, for example, the artwork Voice Mosaic ahead in this chapter). In the second case, we can mention as an example the multimodal browsers like Opera, which allows access to visual information while one interacts simultaneously via voice (see the application *Multimodal Chinese Food*, using the Opera browser at [http://www-306.ibm.com/software/pervasive/multimodal/chinese/], developed by IBM).

The methodologies and principles for voice interface (VUI – Voice User Interface) design overlap substantially with other types of interface design. However, there is an amount of characteristics of voice interfaces that presents unique design challenges and opportunities. Two of these characteristics stand out when the modality of the interface is purely aural and the interaction happens via spoken language (Cohen, 2004: 6).

Besides the fact that the particularities of the voice affect the purely aural voice interfaces, their operation also differs from the visual interfaces, according the table 1.

According to the Gartner Group (Farber, 2004), in 2015 the interfaces will be invisible and ubiquitous. Although the sensors are the main responsible for the transparent interface of the future, probably the voice interfaces will have their share of responsibility in this process too, since the invisibility is one of the aural characteristics of the pure voice interfaces.

	Visual Interfaces	Pure Aural Voice Interfaces
Based on	Visual pages	Blocks of dialogs
Designed for	Control by the eyes	Control by the ears
User action	Brain/ touch (mouse clicks / typing)	Brain / speech
User control	Multi-task (several windows /screens simultaneously)	Mono-task (one conversation a time)
Interaction control	User (the user controls the visual browser – user in command)	Computer (the server controls the voice browser – the browser controls the process)

Table 1. Comparison between the functioning of visual and pure aural voice interfaces

Kerckhove (2003: 21) says that "apparently in the western art and history during the ancient times and, again, from the Renaissance to the modern times, the dominant sensorial bias has been the vision. (…) Nowadays, thanks to the electricity, the actual dominant bias is challenged by the tactile bias" (2003: 21), since we use mouse, keyboard, etc., during most of our computational interaction processes. If we think about the invisible interface we should remember that invisible is not the same as inexistent. Invisible can be immaterial, but the possibility of projecting visual interfaces on *eyesphones*[1], for example, combines the trend of sensors and invisible computers with the human dominance visual and tactile.

Johnson (2001: 101) argues that "simple words keep playing an enormous role in the interface nowadays. And this role seems fated to become more decisive to our informational space in the next decade." Considering that the text editor affects profoundly our way of creating and writing, and that each modality of interface changes our way of thinking and acting in the world, it is expected that all kind of interfaces co-exist and bring hybridizations of the media and forms, as it happened with the email, which, due its frailty and digital form, created a more casual and colloquial style of writing, a mix of the written letter with the talk on the telephone (Johnson, 2001: 105).

In our actual technological scenario, Wilson states that:

> "Computers have a conceptual background from its historical origins from commercial companies and military. The computer screen and its conventions derive from the long history of the representation in the Western culture, from painting, perspective, photography, cinema, to graphic animations and computer metaphors. Similarly, the computer conventional physical interface with keyboard and mouse has a significant cultural baggage. Its restrictions have limited the imagination in thinking about ways of integrating the digital information to human life. (…) Researchers and artists have started to question how the interface between digital systems and people could extend more widely in the human life. Going beyond keyboard and mouse, how the computers could read the human actions such as movement, gesture, touch, look, speech and interactions with physical objects? The wearable computer can convert the body action into information function." (Wilson: 2002: 729).

[1] *Eyesphones* are small glasses that can be connected to computers that project the screen in front of the eyes.

Voice interfaces are a new option in the actual scenario. According to Wilson (2002: 775), "the extension of speech to machines will mark a significant cultural event that will mobilize the artistic attention."

Considering that to "speak" is not the same as "reading" and "writing", and that these processes co-exist along the human history since the most ancient known references, we could suggest here that the most likely scenario is that all different kinds of interface – visual, oral, sensorial, tactile, gestural, etc. – co-exist in the future to answer to the different human needs, instead of replacing each other. Of course, each new kind of interface brings some benefits that answer to more specific needs, but the human needs are diverse and varies according to the context, culture and convenience.

According to Nielsen (2003), "Voice interface will not replace the screens (visual interfaces) as a matter of choice of most users. (…) Several people have overrated impression about rhe benefits of voice interfaces, probably based on the prominence of the voice operated computers in Star Trek." Nielsen also points out that voice interfaces have their great potential in the following cases:

- Users with disabilities that do not allow them to use mouse and/or keyboard or that cannot see;
- Users in situations with busy eyes or hands, for example when driving a car or fixing a complex equipment;
- Users that do not have access to a keyboard and/or monitor, and therefore could use a telephone.

Therefore, for general applications, we believe that voice interfaces can be a great promise as an additional component to multimodal dialogs, more than as an unique interface channel. However, in the case of users with visual or manual disabilities, voice interfaces can be an important channel for inclusion and accessibility.

A research conducted at University of Mariland, consisting in a functional experience for comparing voice controlled web browsers (in the multimodal mode) with mouse controlled web browsers, showed that the voice control improved the performance time in approximately 50% for some kinds of tasks. Subjective measures of satisfaction indicate that for voice navigation, text links are preferable to numeric links, but yet the mouse navigation is still easier to use for general purpose navigation on the web (Christian, 2000).

We can highlight other possibilities for voice interfaces, such as in situations where users prefer to talk to the computer instead of talking to people, as mentioned by Cohen (2004: 9): when the subject of the conversation can cause some kind of embarrassment to the user (for example, when he wants to know about financing values for longer periods and get uncomfortable to ask about low financial rates or about too many options), the user prefer to talk to a voice interface. Although this factor is not exclusive related to voice interfaces, being present in any impersonal man-computer interface, the fact of being able to use natural language to "talk" to a computer about the embarrassing subjects as if one was talking to another person, can provide a better experience that is attractive and pleasant and at the same time answers to the user needs (Cohen, 2004: 11).

According to Nass & Brave (2005), people are 'activated by voice': we respond to voice technologies as we respond to people and we behave as we were in any social situation, and, therefore, the voice interfaces can really emerge as the next frontier for a efficient and friendly technology.

Considering that telephones – either fixed lines as cell phones – exist in larger number and with more penetration in the planet than computers (some places in Africa, for example, where there are not computers available, have a big probability of having telephones available), we can say that voice interfaces reach wider than visual interfaces.

Thinking about the artistic possibilities that the voice interfaces bring, beyond of talking to computers in natural language itself, we could use several voice characteristics regarding the production of artworks, as aesthetical and informational potential. It has become possible in the speech synthesis the manipulation of tone, gender, volume, speed, intonation and voice stress and it can be used to create different perceptions and reactions in the interactors. This possibility of manipulation of vocal characteristics allows generating a dynamic narrative (even in real-time) for stories allowing the use of different personages, according to different contexts and situations. Besides that, in a given moment, we could use phrasal loopings, in another moment, phrases that overlap with each other creating a tridimensional space – louder in the foreground, softer in the background, associating with other temporalities. The sounds could be used accompanied with visual elements in a way that the tridimensional environment would get sound spatiality, and so on.

In the voice interfaces, and probably in any interface, 'what is said' (content) is the most important question in the interaction functional project, and the most important factor that determines usability, according Nielsen (2003). Therefore, the voice interface do not liberate us from the most substantial problems related to interface design: 1) to select the tasks to be supported; 2) to determine the structure of the dialog; 3) to decide which commands or functionalities will be available; 4) to let the users specify what they want, and; 5) to make that the computer give feedback about its actions. As previously mentioned, according to Cohen, "The methodologies and principles of voice interface design substantially overlap with those used in other kinds of interface design. However, there are a number of characteristics in voice interfaces that pose unique challenges and opportunities" (2004:6). Although the main focus of this text is not the interface project itself, it is important to highlight here, as mentioned in Cohem (2004: 4) that the understanding of basic human capabilities and the user needs and goals are the keys for a successful interface design.

The introduction of intelligent voice technologies in the present scenario increases the sonority complexity when compared with previous computational stages, because besides working like an instrument that allows the extension of hearing capabilities, they also allow the complete digitalization and inscription of the voice into computational language, together and mixed with the verbal language (commands or voice information recognized by intelligent voice interfaces become commands or verbal or textual data). According to Santaella (2001: 371), "The verbal language is the most mixed of all languages, because it absorbs the syntax of the sound domain and the form of the visual domain."

Voice interfaces are a new step and possibility for the human-computer interaction, in a process of dissolving the border line between telephone and the internet, and co-existing with other types of interfaces. It is clear that they find their biggest potential in activities and applications in which the modality is auditory and the interaction happens through the spoken language. Due its own peculiar characteristics, the voice brings new artistic potentialities associated with other limits, specialties and complexities, and allow, through the convergence that the technology permits, a new way and new media for communicating, interacting and creating.

2.2 Hypermedia

It is well known that the internet is formed by several servers and clients, and that the most common types of clients, so far, are based on visual interfaces, like email clients (for example, Outlook Express or Gmail), web browsers (for example, Internet Explorer, Firefox, etc.), telnet clients (as Hyperterminal), etc. On the other hand, voice interfaces add one more kind of client to the network, not affecting its topology in terms of servers, but changing drastically the client.

The voice client, i.e., the voice browser, can be a hardware (like the telephone), a software that emulates the telephone (like VoIP softwares, for example), a multimodal browser (like Opera, for example), or even computational devices (like microphones/speakers). Although in voice interfaces the clients are different from those that use visual interfaces, we can have the same applications using simultaneously these two kinds of interfaces. According to Palazzo (2002), the hypertextual system architecture is divided in three levels: presentation level, abstract machine level and database level. Voice interfaces are in the presentation level and can change the system structure specifically in that level, leaving the other levels intact as shown in the figure 1.

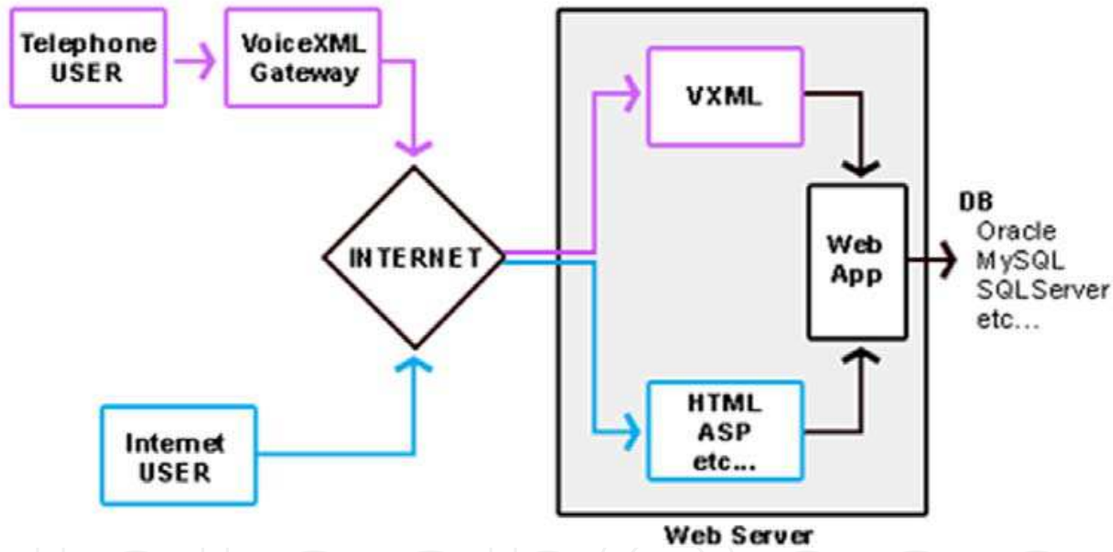

Fig. 1. Diagram showing how voice interfaces created with VoiceXML works in the presentation level, simultaneously with the visual browser.

As can be seen in the figure 1, the line "Telephone USER / VoiceXML Gateway / VXML" represents the application presentation level when accessed by the voice interface, via telephone. The line "Internet USER / HTML" represents the application presentation level when accessed by a visual web browser. The abstract machine level is the box "Web App", which is accessed in the same way by both kinds of interfaces – visual and voice interfaces. From the abstract machine happens the accesses to the databases – the database level – that can also be the same for any kind of interfaces, regardless which technologies are used on the three levels – presentation, abstract machine and database.

When the system (web app) accessed by different kinds of interfaces (visual and aural) is the same, the data captured in the interfaces cannot differ neither in quality nor in quantity, since they are the necessary input for the abstract machine. However, the data input usually

needs completely different treatment between visual and aural interfaces. For example, in a visual form it is possible to present all European countries at once as options to be selected by the user. To create a way for the user to choose an European country in a voice interface, maybe the options must be divided in regions (like North, South, East, West), and then be refined to allow the choice of the country. Once the need data (for example, the choice of the European country chosen by the user) is available, regardless where it came from (visual or aural interface), it goes to the same abstract machine. In those cases, the same application can be used simultaneously by both interfaces – visual and aural.

The fact that we can have multiple interfaces without having to change the abstract machine is a very important factor to allow systems integration and hybridization, since it is necessary that there be some common level in the intermixed systems in order to allow the intermixing process.

Of course, voice interfaces allow specific functionalities and data inputs that are not possible in visual interfaces, like voice recording, for example. Applications that benefits from voice interfaces specific capabilities may eventually need to capture / process / store data in specific ways too, separately from the main application, in order to be able to deal with those specific data. However, this situation of having specific data to treat is not new – even in visual systems, according to their devices and its technological capabilities, it is very common to find diverse characteristics in diverse devices/interfaces that need a different treatment in parallel with the main system. One example of that are the web browsers for smartphones that many times have several limitations and/or differences regarding the type of information they can provide to the main systems, when compared with desktop computational browsers. However, the challenge is to keep the same abstract machine and database levels regardless the kind of interface or device in the presentation level, even if it means to build broader databases that comprise optional specific data according to the interface type. That trend, regulated by the W3C (World Wide Web Consortium) open standards, aims to allow a bigger web integration and interoperability, making it easier the convergence and hybridization.

Voice interfaces, therefore, work like an entrance door or like an initial center for the system that connects the user to a larger hypermidiatic context on the web, in a way that the interactor can 'write' his path on the web in a non-linear way starting from the available options in the voice interface. However, when analyzing the 'reading' process starting from the decisions taken in the voice interface, there is no way to escape from the linearity inherent to orality: the options are presented in a linear order to allow a subsequent non-linear choice by the user.

Considering that the orality presents linear characteristics regarding the user reception, one could initially conclude that voice interfaces would not be hypermidiatic systems and that the complexity level would be small when compared to visual interfaces in the network. However, linearity is just a particular case of non-linearity and voice interfaces are part of bigger hypertextual systems, acting as nodes and transitory centers, connecting the interactor to the further levels in the network. According to Murray (2003:10), the term 'non-linear' should be replaces by 'multisequencial' and 'multiform', as expressions to understand the new narratives forms that: allows the ability to navigate through inter-crossed paths from different points of view, in the first case; and in multiple versions generated from the same fundamental representation, in the second. Besides that, the voice particularities – transience, invisibility, asymmetry, imperfections and limitations – increase

the complexity level of the system and, consequently, the necessity of organizational rigor. The voice asymmetry, for example, requires a rigorous voice interfaces analysis and program in order to adequate the rhythm of voice reproduction/comprehension.

Although the presentation of each level of a voice interface is done through the linear orality, the user navigation between levels follow a hypertextual path, i.e., non-linear, which can even intercross and interconnect with visual and/or hybrid systems in the network, making the complexity even bigger. An example of such hybrid system is the artwork Voice Mosaic (that will be presented ahead in this text), in which the options and fragments of recorded voices made in the system through a voice interface accessed by phone, configure and present visual and aural records in a visual web interface. The signs – visual and aural – stored in the same database are accessed, generated and experienced by two distinct hypermidiatic interfaces – the voice and the visual ones.

Due the voice transience, the paths and options presented during the speech in a voice interface need to be kept in our short memory. In order to be accessed in a comfortable way to be used, it is necessary that the amount of those paths and options be much more limited than in a visual interface, where they can be presented on a computer screen not needing to be transferred to our short memory. As studied by Miller (1956) and explained by Zakia (1997: 82), we can see that we are limited regarding the amount of information we can keep correctly in our short memory:

> "All of our senses are connected in memory. We have memories not only for visual experiences but also of experiences involving sound, smell, taste, touch, movement, and balance. The memories we remember for a long time are called long-term memories (LTM) and are contrasted with short-term memories (STM) that we remember just long enough to use and then forget. (…) There is a limit to the amount of unrelated information a person can hold in STM, from five to nine items, averaging seven."

Therefore, our capability for using spoken options is smaller than our capabilities for using them in the visual mode, where besides of not requiring that the options be all memorized (since they are persistent in the visual and not transient like the voice), it also has a larger associated memory to it.

2.3 Interactivity

Although voice interfaces provide a more natural and human way of interaction between man/computer, they present several differences from visual interfaces.

The first difference is related to how visual screen browsers (like Firefox or internet Explorer, for example) and voice browsers (like the telephone) works. In the first case, the user has a much bigger control over the process because he dominates the time and space when using a visual browser. In the second case, it is the computer that determines the rhythm of the voice browser, by phone (or any equivalent system, such as VoIP) and controls the time/space of the process.

Besides that, in the case of visual browsers, the simultaneous windows and processing allow the multiplication of the user identities in the cyberspace through the simultaneous persistence of several windows and processing. In the case of oral processes, even if we opt to follow a link to an option and then come back to the previous context, there is no way to keep both oral contexts simultaneously. In a moment we are in one context, in the other moment we are in another. Different oral contexts cannot persist simultaneously due their dependence of the time – the voice transience. Therefore, although voice interfaces allow the

hypermidiatic access to a bigger context, they limit some aspects of the interaction that are usually possible through visual interfaces.
Still due to the transience, invisibility and asymmetry inherent to oral processes in voice interfaces, the limitation in the information processed, and that determines the possibilities of interactions, also differ from the traditional computational voice interfaces. For example, the search for a keyword is perfectly possible in a voice interface as much as in a visual one. However, what is the limit of information that we can analyze via an oral result? According to Kerchhove (2003: 20), the difficulty of closing the process is bigger in the oral case. Therefore, voice interfaces make it harder to process large amount of information due the peculiarities of orality.
In this context, the balance between control/pleasure and frustration of the user stays in a fragile zone. According to Murray (2003: 127), "When the things we do bring tangible results, we experience the second pleasure typical of electronic environments – the sense of agency. Agency is the rewarding capability of realizing significant actions and seeing the results of our decisions and choices". When the volume of information and rhythm of voice interfaces allow the closing and control of the process by the user, the agency pleasure really happens. However, a slight deviation that may prevent the control by the user in the agency process and its consequent pleasure can cause frustration and even abandonment of the process. The challenges are big, but no more than the possibilities that rise in the horizon of voice interfaces.
The experimentation of those limits and the combination of possibilities bring additional options to applications that can be explored to deliver richer user interfaces, improving the user experience and increasing the accessibility level.

2.4 Art as tool for experimenting voice interfaces

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 (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 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).

3. Voice mosaic

The *Voice Mosaic* (figure 2) 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. The mosaic is accessed on the web at www.voicemosaic.com.br.

Fig. 2. Screeshot of the artwork Voice Mosaic showing the tiles

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.

3.1 Interfaces and technology

The work has two interfaces (see figure 1) – 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. As mentioned previously in this text, 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 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.

4. 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.

The same kind of VoiceXML based voice interface created for the artwork Voice Mosaic can be used for any kind of application on the web, allowing people to "talk" to the web instead of only seeing it. This ability of dialoging with the web provides a better experience for users with visual disabilities while navigating online.

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.

5. References

Cohen, Michael; Giangola, James; Balogh, Jennifer. Voice User Interface Design. Boston, Addison-Wesley. (2004).

Christian, Kevin; Kules, Bill; Youssef; Adel. A Comparison of Voice Controlled and Mouse Controlled Web Browsing. (2000). Available at [http://otal.umd.edu/SHORE2000/voicebrowse/]. Access on 27.sept.2005.

Farber, D. 2014: Magic Software, Free Hardware, In ZDNet.com. (2004). Available at [http://techupdate.zdnet.com/techupdate/stories/main/Gates_gives_magical_software_tour.html].

Johnson, Steven. Cultura da Interface: como o computador transforma nossa maneira de criar e comunicar (2001). Tradução: Maria Luiza X. de A. Borges. Rio de Janeiro, J. Zahar.

Kerckhove, Derrick. A Arquitetura da Inteligência: Interfaces do corpo, da mente e do mundo, In: DOMINGUES, Diana (org.). *Arte e Vida no Século XXI*. São Paulo, Editora Unesp, p. 15-26.(2003).

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

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

Murray, Janet H. Hamlet no Holodeck: o futuro da narrativa no ciberespaço. (2003). Tradução Elissa Khoury Daher e Marcelo Fernandez Cuzziol. São Paulo, Itaú Cultural.

Nass, C. & Brave, S. Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship. The MIT Press. (2005).

Nielsen, J. Voice Interfaces: Assessing the Potential. Available in [http://www.useit.com/alertbox/20030127.html]. (2003). Access on aug.30th.2005.

Palazzo, Luiz. Sistemas de Hipermídia Adaptativa: Fundamentos, Tecnologias e Aplicações (2002). Available at [http://gpia.ucpel.tche.br/~lpalazzo/sha/]. Access on 23.aug.2004.

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

Pinker, Steven. (2002). O Instinto da Linguagem – Como a mente cria a linguagem. São Paulo, Martins Fontes.(2002).

Santaella, Lucia. Matrizes da Linguagem e Pensamento – Sonora, Visual, Verbal. São Paulo, Iluminuras. (2001).

Wilson, S. Information Arts. Boston, MIT Press. (2002).

Wilson, S. Intersections of Art, Technology, Science & Culture – Links. Available at [http://userwww.sfsu.edu/~infoarts/links/wilson.artlinks2.html]. (2005). Access on jan.10th.2006.

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

Speech and Language Technologies
Edited by Prof. Ivo Ipsic

ISBN 978-953-307-322-4
Hard cover, 344 pages
Publisher InTech
Published online 21, June, 2011
Published in print edition June, 2011

This book addresses state-of-the-art systems and achievements in various topics in the research field of speech and language technologies. Book chapters are organized in different sections covering diverse problems, which have to be solved in speech recognition and language understanding systems. In the first section machine translation systems based on large parallel corpora using rule-based and statistical-based translation methods are presented. The third chapter presents work on real time two way speech-to-speech translation systems. In the second section two papers explore the use of speech technologies in language learning. The third section presents a work on language modeling used for speech recognition. The chapters in section Text-to-speech systems and emotional speech describe corpus-based speech synthesis and highlight the importance of speech prosody in speech recognition. In the fifth section the problem of speaker diarization is addressed. The last section presents various topics in speech technology applications like audio-visual speech recognition and lip reading systems.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Martha Gabriel (2011). Voice Interfaces in Art – an Experimentation with Web Open Standards as a Model to Increase Web Accessibility and Digital Inclusion, Speech and Language Technologies, Prof. Ivo Ipsic (Ed.), ISBN: 978-953-307-322-4, InTech, Available from: http://www.intechopen.com/books/speech-and-language-technologies/voice-interfaces-in-art-an-experimentation-with-web-open-standards-as-a-model-to-increase-web-access

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen