AN APPROACH TO MULTIMODAL AND ERGONOMIC NOMADIC SERVICES

A research experience and a vision for the future

Marco Riva and Massimo Legnani
Cefriel, via Fucini 2 - 20133 Milano (Italy), {surname}@cefriel.it
Tel.: +39 0223954 203 - +39 0223954 210
Fax: +39 0223954 403 - +39 0223954 410

Abstract: The technological evolution in the last few years in the field of devices and network infrastructure and the consequent diffusion of mobile devices have produced the need to allow different ways to use a Web hypertext. Since mobile applications are generally used on small devices with reduced capabilities, there is the need for services able to make the most of the resources available. Two possible approaches to this problem are the multimodal and the ergonomic delivery of hypertexts. In this article we propose a solution for these two approaches.

Key words: multichannel; multimodal; context aware; nomadic; ergonomic; hypertext.

1. INTRODUCTION

Over the last few years Web applications have evolved in different ways, becoming richer, more interactive and more usable.

From the interactivity point of view it is possible to see a transition from simple static to transactional Web sites, where users may perform complex e-commerce operations like buying something or paying taxes and on without moving from their homes.

Another, even more important, revolution has been made in the field of devices. At the beginning of the history of the Web it was only possible to navigate the Web using a personal computer with a browser. It is possible to name this phase the "table Web", since users needed a table or a desk for

their pc. As time passed, smaller devices were introduced and the Web became the "handbag" Web and then the "pocket Web". The Web had become movable.

The third factor of innovation is due to telecommunications: newer technologies have been introduced, allowing wider connectivity and higher bit rates. Nowadays devices are able to manage different kinds of networks (GPRS, UMTS, Wi-Fi) being able to stay connected while moving, as the context changes. Not only can the devices connect to networks, but also they are even able to communicate with each other creating personal area networks (PANs). It is then possible to use Web applications on a taxi, on a train, or even while walking. The Web has become "nomadic"[2][1].

With such technological changes it is necessary to allow different ways to use a Web application. Referring to the Web applications made to be used from a PC with a browser, we can observe a growing interest in the field of usability attempting to simplify the user experience of Web applications.

When talking about nomadic applications the problems and the situations involved are very different from the ones reported in the case of a "traditional" Web navigation. Mobile applications are generally used on small devices with different features and different capabilities and in a context that can vary very rapidly. The technological development has driven improvements in everyday user's life but the exasperated trend towards nomadic and mobile device miniaturization has led to many problems to both the uses and the service developers. The problems are mainly due to the interaction modes with the services. Displays, for example, may be very small, and keyboards are not suited to insert long text. Features that are advantages when talking about mobility, weight and smallness, may become drawbacks when the user needs to use the device.

To improve accessibility and usability to hypertext content, but also to simplify the user-service interaction new access channels, like for example voice control or DTMF (dual-tone multifrequency), were introduced. Moreover, new service adaptation technologies were proposed to adapt contents and services to the user's terminal and device characteristics.

Web applications, developed and delivered as hypertext are now accessible every time and everywhere. The Web has become so big that today it is the best way to find services and contents to solve everyday problems, but also for work, leisure and to meet other people connected to the Internet.

[2] Due to the arguments treated, in this article the terms "nomadic" and "mobile" are considered as synonyms.

However, is today the interaction with Web applications natural? Why don't we try to improve the interaction with services and contents, like for example as in a man-to-man communication?

Man to man communication uses different modes to communicate: voice for long text, visual interaction for images, gestures to improve the communication.

Multimodal delivery of hypertexts offers a possible solution to these kinds of problems. The possibility to interact with the device using different modes allows the user to send inputs and receive outputs in the way he feels more natural and adequate to the service he is using, to its preferences, and to the situation in which he is.

Services with multimodal access may then cover a fundamental role in the future development and success of the Web applications intended for heterogeneous devices.

Different market researches [2] have shown that multimodal services may improve everyday life of service users, increase the user satisfaction and become a competitive advantage for the firms proposing them.

To achieve these goals, a new framework, M^3L [3], has been studied and developed. This framework allows "multimodal hypertext" definition and delivery. Since at the moment only text and voice technologies are available, our multimodal solution only deals with these two modalities. The framework has been designed to be "open" and it may be easily extended adding other interaction modalities. Now it is possible to define a hypertext in which we can specify content that has to be spoken, shown, or both. The same is valid for the input and navigation between pages: we can specify which input is to be completed by voice, by text or pointing, or both.

The user can interact with the service using the best mode for the specific part of the content, as decided at design-time by the hypertext designer.

However, may the designer foretell the best interaction mode at design-time? This is true for standard web applications, where users access the Web contents and services using a PC in their office or at home. But this is not true for a nomadic or mobile user [4]. This kind of user may be in different situations, where one mode is better than another; this is not predictable at design-time, but the best time to make this selection is at run-time. For a nomadic user not only the best interactive modes need to be selected at run-time but also content, hypertext navigation [5] and presentation need to be adapted at run-time to the situation the user is in. An hypertext, able to be adapted at run-time, depending on the delivery environment properties surrounding the user, its preferences and the device being used to improve usability and service's agreeable, is called "ergonomic hypertext".

To adapt a service to the delivery environment we need to know the context information. New technologies for context capturing are available and in the future more a more are going to be available.

Context aware services are sensitive to the state of the context in which they are used. Context awareness is a possible solution to improve the delivery of hypertext on small mobile and nomadic devices. Context-awareness is also one of the most important factors enabling the creation of ergonomic services.

According to what we said, an ergonomic service is also able to select the best interaction mode depending on the situation in which it is delivered. This property is very important for Web services and content created for a nomadic user.

The approach to multimodality and to ergonomic service followed in this work is focused on the following main assertions:

- The service must be written once and automatically deployed and supplied
- The service model has to be simple and easy to translate into XHTML
- The adaptation of contents, presentation, and hypermedia structure must be made server-side so that no particular requirement must be imposed on the client devices. This approach is different from multimodal client-side like SALT [13] or X+V[14] [15] in which the client has to support specific and often proprietary technologies
- The proposed solution must work with today available standard technologies (XHTML, voiceXML, HTML, WML, etc) and devices (PDA, Mobile Phone, Laptop, Smart Phone, etc.)
- The user may use one or more devices by different access modes. This is not possible in client side solution like x+v or SALT.
- The framework must be an open framework, so that it is possible to improve the framework with the new technologies that will be available.

This document describes the results of our research activities and the architecture of the prototype we built to demonstrate the validity of the approach.

This paper is organized as follows: in the first section we describe the M^3L language, that we have defined for the writing of multimodal services and the multimodal delivery framework that we have developed to deliver the multimodal hypertexts written with M^3L. We then describe the ergonomic delivery platform and the eML language, for the delivery of ergonomic services. It must be noted that the main aim of this paper is to describe the multimodal solution while the ergonomic aspects of the problem are mainly introduced to show how the proposed solution may be simply used in the nomadic and ergonomic contexts. The final section of the paper contains our conclusions.

2. MULTIMODAL DELIVERY

There are different definitions of the term "multimodal", but in this paper we adopt a W3C (World Wide Web Consortium) [6] derived one:

"Multimodal interaction will enable the user to speak, write and type, as well as hear and see using a more natural user interface than today's single mode browsers"

With the term "multimodal" then, we mean the possibility to interact with a service using different modes. As a consequence, a multimodal service must be able to support different input and output modes.

In our work, as input modes, we refer to the voice and to the keyboard of a mobile device (PDA or smartphone); voice, audio, written text and images have been chosen to be the output modes. This choice has been made considering the capabilities and characteristics of the devices available on the market. Actually, the modes supported by the devices are the ones listed above.

Anyway, when richer devices will be available, it could be possible to consider other interaction modes, like gestures and haptics, for example.

The solution that we propose for the multimodal delivery of hypertexts allows writing services once. The service is then automatically delivered to the user allowing him the use of different interaction modes.

The solution that we propose for the multimodal delivery of hypertexts is thought to be used in the situations where the hypertext is the same independently from the delivery channels, so that it is possible to write a service once. This solution is not suited to write services where the contents or the structure must be different on the different channels.

The integration of our solution with a methodology (and related tools) for the design of hypertexts [7] that considers different site views for different channels may anyway solve the problem. An example of such a methodology is WebML [8] (and WebRatio).

The main goals of our multimodal delivery platform are:

- The user must be able to interact with the service using the most natural interaction mode.
- Different channels must be used to offer a service to the wider set of users.
- The hypertext creator must define an intrinsic multimodal service to improve service usability and service pleasantness

The framework built allows:

- The access to the services using market available devices.
- The use of currently available delivery technologies. No specific software is needed onboard the device.

2.1 Approach

Our approach to multimodality starts from the experience with multichannel frameworks.

In a multichannel environment, the user must be able to use the same application using different channels (but only one at a time). It is then possible, for example, to use the same service both from a Web browser running on a desktop PC and from the microbrowser of a mobile phone and from a telephone connected to a voice gateway.

A new multichannel markup language [9] was defined. This language is used to write new services (existing ones may be anyway simply manually translated). Every multichannel document written in this language carries extra information about the objects contained, like their essentiality for the overall document comprehension.

The multimodal solution described here starts from our multichannel work, introducing the possibility to use in a coherent and synchronized way the different interaction modes supplied by different channels. It is possible to use a single device (supporting different modes) or even different devices each supporting a subset of the interaction modes.

The same device (or different devices) may then access the same multimodal service using one or more different channels that must be synchronized and coherent while with the multichannel approach only a channel at a time is used.

A new multimodal markup language, called M^3L, has been defined. This language is aimed at writing hypertexts.

The M^3L language allows writing a service that can be delivered using at the same time multiple interaction modes. A prototype shows the integrated "vocal" and "visual" interaction modes being used to deliver web hypertexts. The language offers to the developer a set of attributes and elements that allow to:

- Select the better interaction mode to present the different contents
- Define the modes that can be used to interact with the service
- Force the user to input data only availing of specific mode(s). For example, it is possible to force the user to insert his password only by keyword.

M^3L has been derived from XHTML, adding new elements and attributes. Those elements are used to manage the synchronization of the inputs from the user (essentially during forms definition) and to allow the developer to choose the best interaction mode to deliver the outputs.

2.2 Implementation

Figure 1 shows the architecture of the multimodal delivery platform built: the two modes supported (vocal and visual) are managed using two different channels that are simultaneously open.

The framework proposed here allows a user to interact with both a graphic interface (visual mode) and a vocal one. Those interfaces may be accessed by adequate browsers, able to interpret HTML and VoiceXML documents.

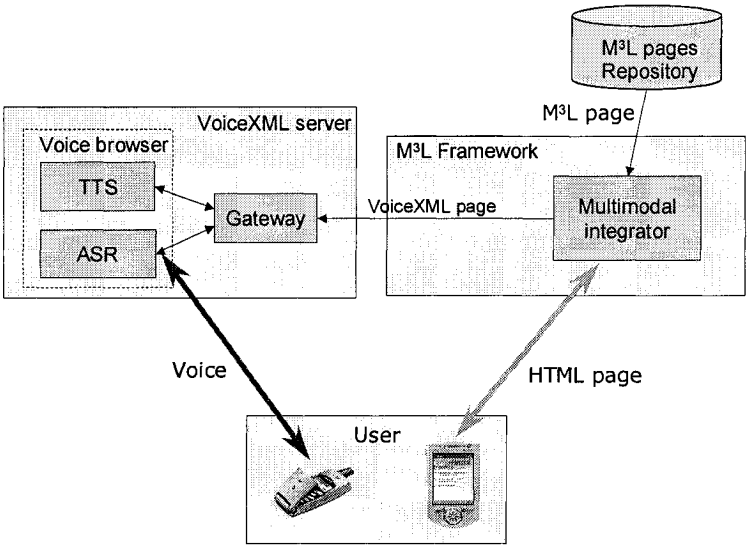

Figure 1. Proposed architecture

The *multimodal integrator* is the core of the multimodal framework we built because it manages the overall operation logic of the system and integrates the inputs coming from the different channels and modes connected. The integrator determines the outputs to send to the user and manages the synchronization between the different channels.

The multimodal integrator is completely written in the Java language and can be installed on any platform supporting the J2EE technology.

The process followed by the integrator to deliver contents to the user is here briefly described:

- At the first request, the multimodal integrator creates a new user's session and sends back to the voice server the VoiceXML initial document. It then creates an applicative session that is necessary to synchronize the different modes used.

- When the user connects to the system with an HTML browser (or with a phone to the voice server) the multimodal integrator asks to the user to login; then takes the current session and finds the M^{3}L document to be delivered. After that it applies to the M^{3}L document a set of XSLT transformations thus obtaining the VoiceXML and HTML pages. The multimodal integrator then creates a finite state machine from the M^{3}L page requested by the user. This machine synchronizes the inputs and outputs coming from and directed to the user on the different channels.
- The pages just created (or part of this) are sent to the browsers that will interpret them.
- The multimodal integrator waits for requests coming from one of the channels in use. These requests may be updating requests, requests of a new M^{3}L page, requests carrying data inserted by the user or commands.
- Depending on the requested operation, the multimodal integrator sends the updated page or the new page requested or the page resulting by the execution of the command or update requested.

The *M^{3}L repository* is the container of the multimodal services and contents.

The *voice server* is the component that allows the vocal communication between the user and the service. It receives the VoiceXML documents generated by the multimodal integrator for the communication with phones, interprets them and manages the vocal interaction with the user. A TTS (text-to-speech) is used to generate the speech to the user, while an ASR (Automatic Speech Recognition) is used to manage the speech of the user and then to collect input data. The IBM WebSphere Voice Server SDK was used, but every VoiceXML compliant voice browser may be used. The voice server enables the vocal interaction allowing the transmission of the voice over ordinary PSTN or GSM networks. As an alternative, it is possible to directly send the VoiceXML file to the client as long as it has an adequate vocal browser.

2.2.1 The M^{3}L language

The M^{3}L language was defined as a set of XHTML modules [10]. New tags have been inserted and among them the most important ones are:

- <label> that associates a label to the different input tags available in the language (as in HTML 4.01 specification);
- <menu> that allows to create link groups, that is to say menus. Grouping anchors is important relating to the voice channel since the reading of a long list of links would be very boring for the user.

The new language specifies even a set of new attributes associated to the elements of a page. Those attributes allow the service developers to select the preferred (or compulsory) delivery or input mode.

The *out* attribute allows specifying which modes can be used to deliver to the user the content of an element. This attribute is made available to any tag that contains information to be presented to the user.

The *mode* attribute, instead, specifies the modes that a user can use to input data. It is then associated to form fields and, relating to our base framework, may have three possible values: "text" to indicate that the user can use a keyboard, "voice" to indicate that the user may use his voice and "all", to say that every known input mode may be used (this values may be extended to support new interaction modality).

The multimodal integrator analyzes these attributes during the transformation from M^3L into the two VoiceXML and HTML documents.

If, for example, the <p> tag has the "out" attribute set to "visual", the text contained will be delivered only with a visual mode (a screen, for example).

2.2.2 Problems solved by the proposed framework

The design of a multimodal delivery platform for hypertexts requires the resolution of different problems. These problems are essentially due to the use of different channels and modes at the same time.

The need to line up the different channels is the most important problem in the multimodal delivery, since it is necessary to offer to the user the sensation to dialog with the same service even when using different devices. Synchronization has the objective to align the data flows sent or received by the different channels used at the same time.

It is possible to separately deal input (from the user to the service) and output (from the service to the user) synchronization problems. The data inserted by the user in the form fields, the navigation commands and the page change requests are managed by a component in the multimodal integrator that generates and maintains the information about the data inserted by the user.

The founding idea of the synchronization mechanism implemented in our framework consists in the extrapolation of a set of tasks from the M^3L document being processed. The sequence of tasks reproduces the steps that must be performed to complete the service distribution. For any input the user must give to the platform, the system creates a task that will be completed by the user with the insertion of the requested information. Only the available/allowed modes may be used. Once a task is completed, the system processes the next one requesting again the insertion of the needed

information. The insertion may be performed only with the modes that are allowed in the M^{3}L document (with the *mode* attribute).

The multimodal integrator manages this process extracting the flow of tasks directly from the M^{3}L document and then generating a finite state machine memorized on the server.

Depending on the collected data and on the reached state, the multimodal integrator determines both if the data insertion is completed and the correct VoiceXML dialog to send.

3. ERGONOMIC MULTIMODAL DELIVERY

According to the definition from IEA (International Ergonomics Association) [11] ergonomics is the scientific discipline concerned with the understanding of interactions among humans and other elements of a system, and the profession that applies theory, principles, data and methods to design in order to optimize human well being and overall system performance.

We consider ergonomics to be the science that studies the relationships between man, machine and environment to obtain the best mutual adaptation.

Therefore, in our case, an ergonomic service must be able to adapt its presentation, contents and navigation to the status of the context that is to the delivery environment, user preferences and device features.

Being our ergonomic solution an evolution of the multimodal research activities, the channels and the modes considered are the same described for the multimodal delivery.

Ergonomic hypertexts are enabled by a framework that has been built to develop and deliver applications able to adapt themselves to the context, thus becoming more appealing as the situation changes. An ergonomic service can automatically propose different interaction modes, different contents and adequate graphical layouts as the context changes.

3.1 Approach

As described above, M^{3}L allows specifying the modes to be used to deliver every single part of the contents of an application. With the same approach it is possible to specify which modes can be used to input data, from the user to the service itself.

The choice of the modes that best fit the user's needs in M^{3}L is made at "design time", when the service is created, but since it is possible to use the service in very different situations, the "design time" forecasts may be wrong. If, for example, the vocal mode is chosen to input the information for

a search engine, the service cannot be used when the user is in an extremely noisy environment.

There is the need to design a solution for the adaptation of the service and of the interaction modes to the state of the environment in which the user may be. This can be achieved selecting at "run-time" the best interaction mode(s) between the user and the service itself.

The next step is to introduce ergonomic features in the multimodal platform.

This platform may be very useful in every application requesting a great amount of interaction between the user and the machine, like information kiosks (at the airport, at the station, in a public building). Typical goals of the platform are:

- Simplifying the creation and use of applications usable [12] by people with disability.
- Reaching a wider set of users
- Increase the usability of the service
- Increase user satisfaction

There are two main concerns that have to be solved in order to achieve these objectives and then allow the ergonomic creation of services:

- Define a model for the description of context information. This model must be abstract enough to be independent from the measurement (capturing) systems (like environmental sensors, body sensors and so on).
- Define the model of an "ergonomic service". This model must describe a service that can perceive the context information available.

The approach followed to make an ergonomic service is to support a static service, placing side by side a set of active ECA (Event Condition Action) rules. These rules can make the service reactive to the available context information since allow to specify:

- An "event" part, used to catch changes in the context status
- A "condition" part, where a set of predicates allow to choose the activation of the rule, on the basis of the information related to the event occurred
- An "action" part, where the actions to be taken are listed. These actions must be done only if the rule is activated.

We designed a new XML language aimed at writing "adaptable" services. This language was named eML (ergonomic Markup Language) and it is thought to write Web hypertexts that can react to context changes by ECA rules.

eML allows describing the parts that compose a service, the elements that make up the parts of the service and the actions that may be applied to the elements.

The language is composed of a passive and a reactive section. The passive section describes the non-ergonomic aspects of the service and is composed of:

- Contents: the information exchanged between the user and the service
- Navigation: navigation paths available to reach the contents
- Style: styles applied to the page
- Layout: the structure of the page

On the other side, the reactive section of the language is used to express the rules that make the service reactive to changes in the context (events).

The general idea of our approach is shown in figure 2.

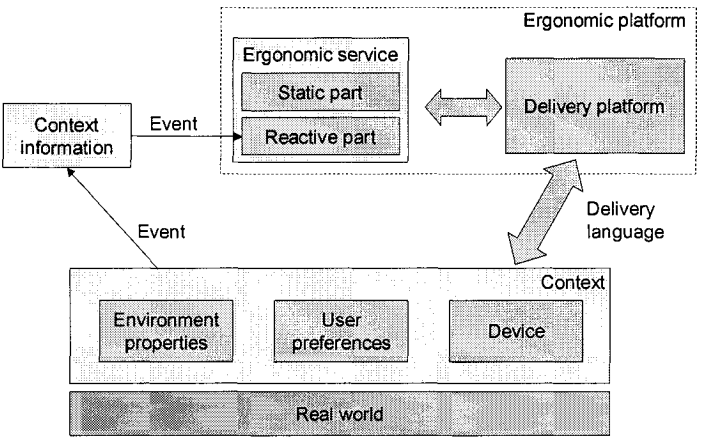

Figure 2. the approach to ergonomic services

4. RELATED WORK

To improve the results of our recent research activities, we are working on different approaches to improve user-service interaction and hypertext delivery mechanism.

The first activity is to try to define "design methodologies" and "development tools" able to create ergonomic nomadic services, accessible by different "channels" also simultaneously (multimodal ergonomic services for Nomadic and mobile users).

The second activity is to improve the multimodal technologies available, putting together different approaches available for multimodal delivery of content. In this activity we propose a service model and a set of related tools to select the multimodal system available to the user in the specific situations

the user may be and using available devices. If the user, for example, has a SALT capable browser (client side approach to multimodality) we can deliver the modeled services as a "SALT" service. On the contrary, if the user has two devices (for example a mobile phone and a video-text guide), we provide the services in M^3L like mode, delivering the voice part to the mobile phone and the visual part on the visual terminal.

Another active work is the MAIS project. We are trying to put together a data intensive hypertext methodology (WebML and a WebML run time, like WebRatio) with the multimodal delivery ergonomic system presented here. The primary goal is to improve methodologies, and provide tools to simplify the design and implementation of multimodal adaptive services.

Finally we are working on the Nomadic Media project, that aims at integrating advanced interaction modes like gesture and haptics with classic interaction modes (voice, video, text, etc) to provide the best available value added service to a Nomadic User.

5. CONCLUSIONS

In this paper we described the M^3L framework and language that support the creation and the delivery of multimodal hypertexts. In addition we mentioned the eML language for the delivery of ergonomic hypertexts.

M^3L provides an easy way to write and delivery newly made multimodal services while eML allows writing and deploying ergonomic services.

Starting from our previous experience, we demonstrated that, with simple extensions (a few attributes and tags) to the XHTML language, it is possible to create very powerful and automated multimodal and ergonomic applications, only using currently available technologies (devices, standard, etc.).

We are now working on the extension of the data intensive methodology WebML, which is our proposal for adaptive hypetext modeling; eML is the language defined to write ergonomic services. The language defined for the final (multimodal) delivery is M^3L.

It is then possible to model a data-intensive hypertext application using WebML* and then make it ergonomic and context-aware specifying its implementation and a set of active rules using the eML language. An eML service may be delivered to the final user by the M^3L multimodal framework proposed here.

This solution is suited even for the multimodal and ergonomic delivery of contents to persons with disabilities.

The typical problems of disabilities, in fact, may be solved with the eML & M^3L solution when considering that all the people have, more or less,

disabilities, depending on the specific situation they are. A person with disabilities has specific needs, as anyone else has, and thus, considering the disability as a part of the context, it is possible to define a set of adaptive rules with eML to adapt the service to the specific needs of a person with disabilities like it is possible for the person that hasn't disabilities.

Even if this approach does not allow solving all the accessibility problems of all the people with disabilities (like, for example, cognitive disability where contents have to be different), the union of the multimodal ergonomic delivery with the "site view" approach of WebML may be suited for most cases.

REFERENCES

[1] Leonard Kleinrock - Nomadic Computing - an opportunity - January 1995.

[2] Comverse - Comverse And The Yankee Group Announce Preliminary Results From User Research Into Multimodality - 19 February 2002 (http://63.64.185.200/news/news_big.asp?cat=65&newsid=247)

[3] Marco Riva, Massimo Legnani, Maurizio Brioschi - Multimodalità nella fruizione dei servizi - 07 October 2003. MAIS report 7.1.1

[4] M. Weiser - The Computer for the 21st Century - November 1991

[5] Andrew Dillon - Designing usable electronic text: ergonomic aspects of information usage - 1994

[6] W3C Activity - Multimodal Interaction Activities (http://www.w3.org/2002/mmi/) – 2002

[7] Niels Erik Wille - Hypertext concepts: A historical Perspective - November 2000.

[8] Stefano Ceri, Piero Fraternali, Aldo Bongio – Web Modeling Language (WebML): a modeling language for design Web sites – WWW9 Conference May 2000 - http://www.webml.org/webml/upload/ent5/1/www9.pdf

[9] Massimo Legnani, thesis "CDI-ML: Channel and Device Independent Markup Language", October 2000

[10] W3C Recommendation - Modularization of XHTML - 10 April 2001. Available online at http://www.w3.org/TR/xhtml-modularization/

[11] IEA (International Ergonomics Association), "The Discipline of Ergonomics", http://www.iea.cc/ergonomics/, August 2000

[12] Maurizio Boscarol - Che cos'è l'usabilità dei siti web – novembre 2000 - http://www.usabile.it/012000.htm

[13] SALT – Salt forum (http://www.saltforum.org) – 2002

[14] W3C Note - X+V – XHTML + Voice Profile 1.0 (http://www.w3.org/TR/xhtml+voice/) - 21 December 2001

[15] VoiceXML forum - X+V – XHTML + Voice Profile 1.2 http://www.voicexml.org/specs/multimodal/x+v/12/spec.html - 16 March 2004