IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Suite 200, Hershey PA 17033, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP5247

This paper appears in *Managing Modern Organizations Through Information Technology*, Proceedings of the 2005 Information Resources Management Association International Conference, edited by Mehdi Khosrow-Pour.

A Framework for Multimodal Web-Based Information Systems on Mobile Agents

Cecilia Wai Shee Chan and Lindsay Smith
School of Multimedia Systems, Monash University, Clyde Rd., Berwick, Australia,
cwcha8@student.monash.edu.au, lindsay.smith@infotech.monash.edu.au

ABSTRACT

The proliferation of ubiquitous computing is changing the way Web-based information is accessed. As wireless computing devices, such as Personal Digital Assistants (PDAs), Smartphones and mobile phones, become smaller and possess sufficient processing power to handle a variety of functions, the conventional unimodal input via minuscule keypads and styluses becomes impractical as a means of interaction. The first objective of this research paper is to assert multimodal interaction as an alternative and effective approach for accessing Web-based information on mobile devices. The second objective is to propose a Multimodal Document Development Framework to guide the development of multimodal Web-based information systems specific to mobile agents. In validating the effectiveness of such interaction and development framework, a prototype called The Multimodal Library Information System is developed and implemented.

INTRODUCTION

Remotely accessing information via ubiquitous computing is fast becoming commonplace as Personal Digital Assistants (PDAs), Smartphones and mobile phones increase in popularity (Weiser 1991). However, these tiny devices entail limitations that impact user experience and satisfaction during Human Computer Interaction (HCI) (Pham, Schneider et al. 2000). The embedded Web browsers in these mobile devices are predominantly Graphical User Interface (GUI) based, offering cumbersome keypads and styluses for input, and small visual display screens for output. Although the Vocal User Interface (VUI) also serves as an output medium, it rarely performs as a command input medium. An important requirement for immersive HCI, and effective access and retrieval of Web-based information in mobile computing devices is to utilize not only the sense of vision (GUI), but also sound (VUI) and touch (Tactile User Interface – TUI). Such an advance raises new challenges in Web-based information systems to both employ and fully exploit the three senses via a Multimodal User Interface (MUI). The shortcomings of unimodal interactions are taken into account when constructing the Multimodal Document Development Framework and creating its development guidelines. As part of validating the proposed framework and guidelines, a prototype called The Multimodal Library Information System was developed paralleling the Monash University Library Voyager catalogue system (Monash University 2004).

UNIMODAL WEB-BASED INFORMATION SYSTEMS

Traditional GUI based unimodal Web applications emerged in the early 1990s using a universally understood markup language HyperText Markup Language (HTML) (HTML Working Group 2004). HTML is composed of standard tags that define fonts, layouts, graphics and links to other documents creating an immense electronic collection of globally distributed text and multimedia documents (Whitehead and Maran 1997). HTML documents are commonly presented via Web browsers on desktop or laptop Personal Computers (PCs). These applications utilize input peripherals such as keyboard and mouse, and graphical output peripherals such as display screens of high resolution.

With the growing mobile-commerce industry, companies are convinced that anyone possessing a mobile phone, two-way pager or PDA is anxious to conduct transactions over the Internet via a wireless network (Schultz 2001). In addressing the need for mobile devices to access Web-based information, the Wireless Application Protocol (WAP) was developed as a communication bridge. WAP supports a set of tags called Wireless Markup Language (WML) that is based on HTML (Piven 2001). Together, WAP and WML were supposedly designed as a graphical modality to optimize Web access on mobile phones by catering for their low bandwidth, small screen size, low latency, and universal connectivity needs. However, it is widely believed that WAP failed to elevate Web-based information access power because it could not compensate for the devices' reduced input/output capabilities, such as the lack of an alphanumeric keyboard, and nature of very small displays (Armaroli, Azzini et al. 2002).

Both HTML and WML do not support the VUI, thus limit themselves to the unimodal GUI domain. Over reliance on unimodality hinders the complete utilization of Web-based information in mobile devices. In overcoming this limitation, the graphical, vocal and tactile modalities need to be extended and employed in a multi-channel and multimodal manner.

MULTIMODAL WEB-BASED INFORMATION SYSTEMS

In addressing the need for multimodal interactions on the Web for mobile agents, the World Wide Web Consortium (W3C) and the WAP forum initiated the *Multimodal Interaction Working Group* in 2002 that visualises "web pages you can speak to and gesture at" (Maes and Saraswat 2003), and works towards defining standards for a new class of device that supports multiple modes of interaction on the Web (Raggett and Froumentin 2004). The *Multimodal Interaction Working Group* initially focused on defining *Multimodal Interaction Use Cases* (Candell and Raggett 2002) and *Multimodal Interaction Requirements* (Maes and Saraswat 2003). The two specifications later led to the publication of the *Multimodal Interaction Framework* on 6 May 2003 (Larson, Raman et al. 2003). The *Multimodal Interaction Framework* defines how markup languages should describe the functions of and data flow between major components in multimodal systems, and provides specifications for how the Web supports multiple modes of interaction. In addition, it ensures that the set of markup languages for presenting multimodal content are built upon existing W3C markup languages along with the W3C Document Object Model (DOM).

Two key computing industry players, Microsoft and IBM, are currently working towards the development of differing multimodal markup languages:

- The Speech Application Language Tags (SALT) forum, headed by Microsoft, initiated the SALT specification (Members of SALT Forum 2004); and,

- The Voice eXtensive Markup Language (VoiceXML) forum, constituted by IBM, initiated the XHTML + Voice Profile (X+V) specification (Axelsson, Cross et al. 2004).

XHTML + VOICE PROFILE OVERVIEW

The Multimodal Library Information System prototype for mobile devices developed for this research project utilizes the following technologies:

- **Multimodal markup language:** *XHTML + Voice Profile (X+V) 1.2* – the latest version of X+V as at 15th September, 2004
- **Multimodal Web browser for embedded devices:** *ACCESS Systems' NetFront Multimodal Browser* (Access 2004) for PocketPC 2003 equipped with the IBM ViaVoice speech technology (IBM Multimodal Technologies 2004b)

X+V is a W3C standards compliant markup language developed by IBM, Motorola and Opera. It involves the integration of XHTML 1.1 (Altheim and McCarron 2001) with VoiceXML 2.0 (Burnett, Carter et al. 2004) using XHTML Modularization (Altheim, Boumphrey et al. 2004). The modes of interaction between XHTML and VoiceXML are then subsequently synchronized using the Document Object Model 2 Events (DOM 2 Events) model (Pixley 2000) via the authoring of eXtensible Markup Language Events (XML Events) (McCarron, Pemberton et al. 2003). In other words, the *tactile* XML Events are responsible for the correlation of *vocal* VoiceXML snippets with the associated *graphical* XHTML elements:

- *graphical* elements include XHTML forms, input text boxes, check boxes, etc,
- *vocal* elements include VoiceXML fields, forms, prompts, etc, and
- *tactile* elements include XML Events actions, mouse clicks, on focus, etc.

XHTML, VoiceXML and XML Events are all official standards for the Web defined by the Internet Engineering Task Force (IETF) (IETF 2004).

A multimodal X+V Web application can be viewed as a composite of three specialized files:

- the parent file (mxml) comprises of the graphical, vocal and tactile components, and
- the child files (jsgf and pbs) contain the vocal grammar and pronunciation definitions in VoiceXML tags (vxml).

The files cooperatively handle multimodal interactions, error recovery, user confirmation, speech recognition, and speech synthesis.

When the user speaks to the vocal interface of the NetFront multimodal Web browser, the in-built IBM ViaVoice speech recognition engine converts the spoken sound-waves (raw data) into meaningful fragments (sensible information). This process involves comparing the spoken utterances with a set of defined words archived in the Java Speech Grammar Format (JSGF) file. JSGF is a platform and vendor independent Java programming convention that adopts traditional grammar notations developed by Sun Microsystems (Sun 1998), and is formally utilized by VoiceXML applications (Burnett, Carter et al. 2004).

In addition to defining sets of grammatical syntax, specifying their pronunciation is also important, particularly when tolerating regional accents and variations in context among words of identical pronunciation. The IBM ViaVoice speech recognition engine supports a customizable pool (PBS) file that adopts the International Phonetic Alphabet (IPA) system for defining pronunciations (IBM Multimodal Technologies 2004a).

Similarly to a typical HTML document, the structure of an X+V document comprises of the **declaration**, **header**, and **body** sections:

- The **declaration** resides in the initial section of the document. It contains the definition of the Document Type Declaration (DOCTYPE) the Document Type Definition (DTD) version, the namespaces specifications for the graphical, vocal and tactile components, and the specification of the speech language (e.g. U.S. English). For X+V 1.2, the graphical component should be compliant with XHTML 1.1, the vocal component with VoiceXML 2.0, the tactile component with XML Events (Axelsson, Cross et al. 2004);
- The graphical component spreads between the **header** (wrapped inside the <head> and </head> tags) and the **body** (wrapped inside the <body> and </body> tags) sections; this contains the visual markup based on XHTML;
- The vocal component resides in the document **header** section; this contains the vocal markup based on VoiceXML;
- The tactile component resides in the **body** section; this contains the tactile markup based on XML Events. XML Events handles calls initiated by XHTML to snippets of VoiceXML.

THE MULTIMODAL DOCUMENT DEVELOPMENT FRAMEWORK FOR MOBILE AGENTS

The proposed Multimodal Document Development Framework in this research paper is based on IBM's multimodal development flowchart published in the *Getting Started Guide for Multimodal Tools v4.1* distributed within the *IBM WebSphere Multimodal Toolkit* (IBM Multimodal Technologies 2004b). In adding value to this framework, this research paper proposes development guidelines with the aim to effectively integrate multiple modalities via the analysis of multimodal interaction patterns feasible on mobile agents.

IBM's Multimodal Development Flowchart

The multimodal development flowchart published by IBM provides descriptive development methods, but focuses on illustrating how input and output fields of graphical XHTML documents (the GUI) are voice-enabled via the association of vocal VoiceXML (the VUI) components, rather than addressing the integration of all three modalities. According to Streit's (1998) multimodal integration strategy for building multimodal systems, supporting gestures (the TUI) should first be integrated with natural language (the VUI), and then this integrated result should be applied to graphical interface (the GUI). The first step towards constructing a more effective Multimodal Document Development Framework is to include the tactile XML Events (TUI) in the flowchart to strategically integrate multiple modalities according to suitable multimodal interaction patterns for mobile devices.

Patterns of Multimodal Interaction

A Multimodal User Interface (MUI) applies various aspects of human communication to computing. In order to develop an effective MUI that offers constructive HCI, multimodal designers must learn "how linguistic notions of conversation can be incorporated into graphical user interfaces" (Sullivan and Tyler 1991). This research paper hypothesizes two types of effective multimodal interaction patterns for mobile devices, namely **serial** and **parallel**.

In a **serial** pattern, the user switches between using available modalities, with only one modality utilized during its period of activation. Oviatt (1999) asserts that this type of condition effectively increases the value of the interaction via mutual disambiguation. During mutual disambiguation, each modality provides partial information that best contributes its strengths, aids the interpretation of other modalities, and compensates for the weaknesses and errors of other modalities. In this respect, the user may adopt less fatiguing and stronger modalities to accommodate changing communication climates surrounding mobile environments.

Under the **parallel** pattern, the user utilizes more than one modality in a simultaneous and collaborated manner. This condition improves efficiency since the user may more quickly access and respond to information by exploiting more than one modality at a time. When dealing with simultaneous multiple inputs, the system is required to

Figure 1. Multimodal Document Development Framework

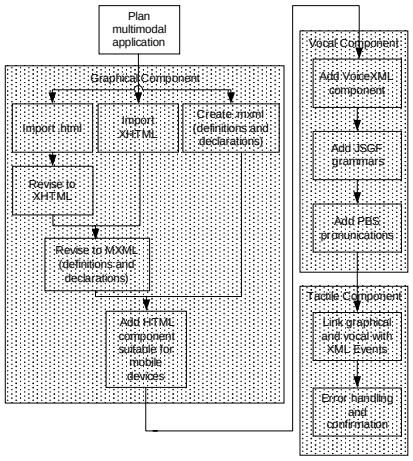
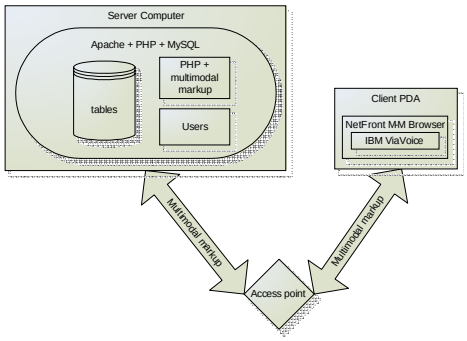


Figure 2. System Architecture of The Multimodal Library Information System



integrate data from each contributing modality and subsequently process them into semantically compatible information.

The Proposed Framework

The proposed Multimodal Document Development Framework is established by bringing together IBM's multimodal development flowchart and the two hypothesized patterns of multimodal interaction (Figure 1).

The Multimodal Document Development Framework aims to guide the designer throughout the course of developing a multimodal application. The framework suggests three possible ways to commence the creation of a multimodal document – from either an existing HTML or XHTML document, or starting from scratch. The subsequent steps illustrate the integration of the graphical, followed by the vocal, and lastly the tactile components. The framework presents a well-structured development model that adheres to the standards and definitions of W3C's *Multimodal Interaction Framework*. Furthermore, the framework strives to facilitate constructive multimodal interactions by ensuring the **serial** condition is satisfied by exploiting strengths of each modality, and the **parallel** condition by linking modalities in a simultaneous manner.

THE MULTIMODAL LIBRARY INFORMATION SYSTEM PROTOTYPE

As part of validating the proposed framework, a Web-based prototype was developed for the Monash University library. Monash University operates several libraries, many of which are equipped with Wireless Fidelity (WiFi) connectivity for academics and students to connect their personal computers, laptops and PDAs. The library's Voyager catalogue system, which is a unimodal Web-based system, was an excellent candidate for conversion to multimodal operation for mobile devices. A scaled down prototype, "The Multimodal Library Information System", was created based on the current Web system and successfully implemented on the NetFront multimodal Web browser running on a Microsoft PocketPC 2003 equipped PDA. There are three components in the prototype, namely a wireless access point for WiFi connectivity, a server computer and a client mobile device. Figure 2 illustrates how these components are integrated.

- The WiFi network access point allows electronic mobile devices equipped with a WiFi connector to access The Multimodal Library Information System.
- The server computer runs the Apache Web server, and is equipped with PHP scripting and the MySQL database. This server hosts The Multimodal Library Information System, where its database is structured comparably and populated with similar data to the actual Voyager catalogue system.
- The wireless PDA client (equipped with embedded microphone and speaker, in conjunction with the NetFront multimodal Web browser) supports full multimodal input and output for the searching of The Multimodal Library Information System hosted on the server.

The prototype development validates the proposed Multimodal Document Development Framework by following the "Import .html" path (see Figure 1). The Voyager catalogue system offers two types of query methods, one of which is "Basic search". The "Basic search" HTML document was amended for prototype development (see Figure 3).

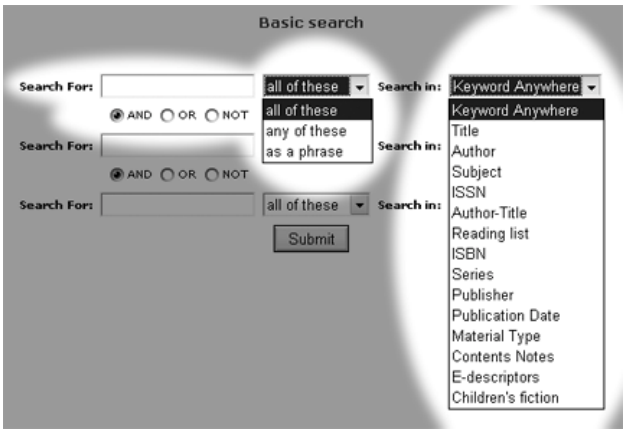
Graphical Component

The revision of an HTML document to an MXML multimodal document involves coding the multimodal X+V 1.2 declarations and definitions into the initial section of the document. Next, the graphical input fields are identified for subsequent vocal association (see Figure 3).

Vocal Component

The association of the vocal component involves defining the grammar (via JSGF) and pronunciation (via PBS) for capturing the users' spoken input. Using the "Search In" dropdown box as an example for illustration

Figure 3. The "Basic Search"



(see Figure 3), the grammar definition in JSGF syntax for capturing valid vocal utterances would include:

```
#JSGF V1.0 iso-8859-1;
grammar search_in;
public <search_in> = Keyword [Anywhere]
                   | Title
                   | (Author | Writer)
```

The annexed PBS pronunciation syntax to this JSGF grammar file, which defines the valid phonetic utterances, would include:

```
Keyword K IY W ER DD
Title T AY DX AX L
Author AO TH AXR
Writer R AY DX AXR
```

Tactile Component

After the graphical XHTML and vocal VoiceXML components have been defined, the tactile XML Events component is then attended to. Figure 4 shows how XML Events links the graphical and the vocal components in the X+V markup language.

The Multimodal Interaction Experience

The following small excerpt demonstrates a scenario where a user interacts with The Multimodal Library Information System in a multimodal manner. The scenario demonstrates serial and parallel multimodal integration patterns during the selection of "Author" from the "Search in" dropdown box (see Figure 5).

(i) The multimodal application requests user input *graphically* (the label "Search in" and associating dropdown box) and *vocally* (speaker prompt: "search in") – **parallel interaction**;
(ii) User inputs *vocally* (speaks "writer" into the microphone) – **serial interaction**;
(iii) Application *validates* the vocal input (recognizes "writer" as synonym of "author" in the JSGF and PBS file) and, if necessary, performs error recovery;

Figure 4. Correlation of HTML with VoiceXML via XML Events in X+V Markup Language

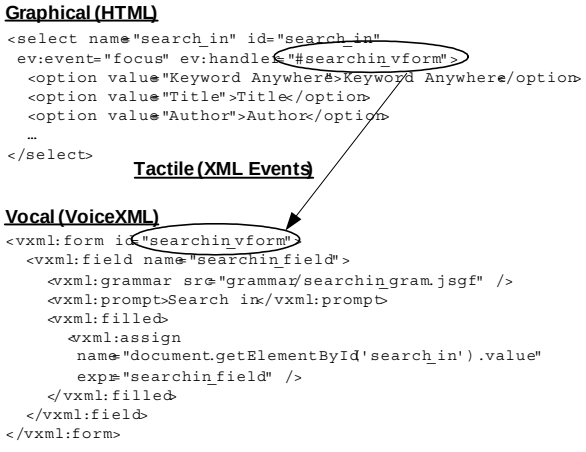
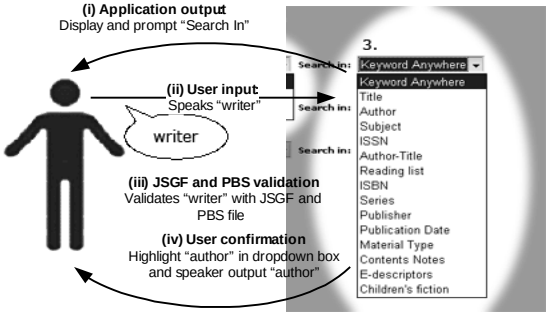


Figure 5. Multimodal Interaction for "Search In" Dropdown Box



(iv) Application *confirms* its interpretation *graphically* (highlights "author" entry in dropdown box) and *vocally* (speaker prompt: "author") – **parallel interaction**.

CONCLUSION AND FUTURE DIRECTIONS

With the proliferation of ubiquitous computing, users now desire access to Web services anywhere, on any device and at anytime. The conventional GUI based Web browsers are inadequate in meeting this demand due to the limited nature of unimodal interaction

This research proposes a Multimodal Document Development Framework that guides the development of multimodal Web-based information systems for mobile agents by ensuring the specifications of both the W3C Multimodal Interaction Framework and IBM X+V markup language are complied with. Together with this proposed framework, effective HCI is demonstrated through the analysis of multimodal interaction patterns. A prototype system for mobile agents, called The Multimodal Library Information System, was developed to illustrate the effectiveness of this framework by validating the integration and synthesis of the graphical, vocal and tactile components. The successful implementation of such MUI on the Monash Voyager catalogue system demonstrates how existing unimodal Web applications may be reengineered in a multimodal manner. In addition, this prototype system will undergo user testing for investigating feasible and suitable multimodal interactive patterns for mobile agents.

The application of MUI to Web applications is a relatively new field of study that is undergoing progressive development and is receiving significant amount of attention from both the W3C and the industry. The proposed development framework and prototype are designed to facilitate research and improve multimodal functionality available in the Web arena for mobile agents.

REFERENCES

Access. 2004, "Access - Embedded Products", (Access), Available: http://www.access-us-inc.com/Prod_NetFront_nf_xhtml.html (Accessed: 15.09.2004).

Altheim, M., F. Boumphrey, et al. 2004, "Modularization of XHTML", (W3C Technical Reports and Publications), Available: http://www.w3.org/TR/xhtml-modularization/ (Accessed: 15.09.2004).

Altheim, M. and S. McCarron. 2001, "XHTML 1.1 - Module-based XHTML", (W3C Technical Reports and Publications), Available: http://www.w3.org/TR/xhtml11/ (Accessed: 15.09.2004).

Armaroli, C., I. Azzini, et al. (2002). 'An architecture of a multi-modal Web Browser'. ICSLP, September 16-20, Denver, USA.

Axelsson, J., C. Cross, et al. 2004, "XHTML+Voice Profile 1.2", Available: http://www.voicexml.org/specs/multimodal/x+v/12/ (Accessed: 29.05.2004).

Burnett, D. C., J. Carter, et al. 2004, "Voice Extensible Markup Language (VoiceXML) Version 2.0", (W3C Technical Reports and

Publications), Available: http://www.w3.org/TR/voicexml20/ (Accessed: 15.09.2004).
Candell, E. and D. Raggett. 2002, "Multimodal Interaction Use Cases", (World Wide Web Consortium), Available: http://www.w3.org/TR/2002/NOTE-mmi-use-cases-20021204/ (Accessed: 21.09.2004).
HTML Working Group. 2004, "W3C HTML Home Page", (W3C), Available: http://www.w3.org/MarkUp/ (Accessed: 27.05.2004).
IBM Multimodal Technologies (2004a). Getting Started Guide for Multimodal Tools V4.1, IBM. 2004.
IBM Multimodal Technologies. 2004b, "Why IBM - Leadership in multimodal", (Pervasive Computing: IBM Wireless, Voice and Mobile Software Products), Available: http://www-306.ibm.com/software/pervasive/multimodal/ (Accessed: 21.09.2004).
IETF. 2004, "IETF Home Page", (IETF Home Page), Available: http://www.ietf.org/ (Accessed: 15.09.2004).
Larson, J. A., T. V. Raman, et al. 2003, "W3C Multimodal Interaction Framework", (World Wide Web Consortium), Available: http://www.w3.org/TR/2003/NOTE-mmi-framework-20030506/ (Accessed: 21.09.2004).
Maes, S. H. and V. Saraswat. 2003, "Multimodal Interaction Requirements", (W3C), Available: http://www.w3.org/TR/mmi-reqs/ (Accessed: 08.06.2004).
McCarron, S., S. Pemberton, et al. 2003, "XML Events", (W3C Technical Reports and Publications), Available: http://www.w3.org/TR/xml-events/ (Accessed: 15.09.2004).
Members of SALT Forum. 2004, "SALT Forum", Available: http://www.saltforum.org (Accessed: 26.04.2004).
Monash University. 2004, "Monash Voyager Catalogue", (Monash University), Available: http://library.monash.edu.au/ (Accessed: 29.09.2004).
Oviatt, S. (1999). "Mutual Disambiguation of Recognition Errors in a Multimodal Architecture."
Pham, T., G. Schneider, et al. (2000). 'A situated computing framework for mobile and ubiquitous multimedia access using small screen and composite devices'. Proceedings of the eighth ACM international conference on Multimedia, Marina del Rey, California, United States, pp.323 - 331.
Piven, J. (2001). "Latest Bad Rap Zaps WAP." Computer Technology Review vol.21, no.2, Feb 2001, pp.1.
Pixley, T. 2000, "Document Object Model (DOM) Level 2 Events Specification", (W3C Technical Reports and Publications), Available: http://www.w3.org/TR/DOM-Level-2-Events/ (Accessed: 15.09.2004).
Raggett, D. and M. Froumentin. 2004, "W3C Multimodal Interaction Activity", Available: http://www.w3.org/2002/mmi/ (Accessed: 29.04.2004).
Schultz, B. (2001). "The m-commerce fallacy." Network World vol.18, no.9, 26/02/2001, pp.77.
Streit, M. (1998). Why Are Multimodal Systems so Difficult to Build? - About the Difference between Deictic Gestures and Direct Manipulation. London, UK, Springer-Verlag.
Sullivan, J. W. and S. W. Tyler (1991). Intelligent User Interfaces. New York, USA, ACM Press.
Sun. 1998, "Grammar Format Specification", (Java Technology), Available: http://java.sun.com/products/java-media/speech/forDevelopers/JSGF/ (Accessed: 15.09.2004).
Weiser, M. (1991). "The Computer for the Twenty-first Century." Scientific American, September, 1991, pp.94-10.
Whitehead, P. and R. Maran (1997). Internet and World Wide Web: Simplified, John Wiley & Sons.