Speech-Enabled Mobile Field Applications

Kondratova, Irina

National Research Council Canada
Conseil national de recherches Canada

Institute for Information Technology
Institut de technologie de l'information

NRC · CNRC

Speech-Enabled Mobile Field Applications *

Kondratova, I.
August 2004

* published in the Proceedings of the International Association of Science and Technology (IASTED) International Conference on Internet and Multimedia Systems (IMSA). Hawaii, USA. August 16-18, 2004. NRC 47156.

Canada

SPEECH-ENABLED MOBILE FIELD APPLICATIONS

Irina Kondratova
NRC IIT e-Business
46 Dineen Drive
Fredericton, NB, Canada E3B 9W4
Irina.Kondratova@nrc-cnrc.gc.ca

ABSTRACT

This paper discusses the advantages and challenges of using multimodal and voice technologies for field data collection. Current developments in the multimodal, mobile, field data communications are discussed. VoiceXML technology for speech-enabled information retrieval and input using natural speech over the phone is described. Two field prototype applications developed using this technology are also discussed, including speech-based inventory and time management services for field users. The author presents the results of performance studies for the prototypes and discusses the pilot usability study conducted on a group of potential users.

KEY WORDS

Field data collection, multimodal interaction, mobile devices, speech recognition, voice user interface.

1. Introduction

The potential of using mobile handheld devices in the field is limited by antiquated and cumbersome interfaces. A small screen size and the need to use a pen to enter data and commands, present a great inconvenience for field users - especially if their hands are busy using other equipment, or instruments. Speech processing is one of the key technologies to simplifying and expanding the use of handheld devices by mobile workers [1, 2].

The availability of real-time, complete, field information exchange with the project information repository is critical for decision making in the field of construction-site inspection, as information frequently has to be transmitted to and received from the project repository on-site without the engineer making a costly additional field trip [3]. In some cases, when the security and safety of people and infrastructure are at stake, the importance of real-time communication of field data becomes paramount [4].

2. Multimodal Field Data Collection

Multimodal interaction can be described as the integration of visual and voice interfaces through the delivery of combined graphics and speech, on handheld devices [5]. This technology enables more complete information communication and supports effective decision-making. It also helps to overcome the limitations imposed by the small screen of mobile devices.

2.1 Multimodal Interaction Technology

There are different models for implementing multimodal interaction on mobile devices. The fat client model employs embedded speech recognition on the mobile device and allows conducting speech processing locally. The thin client model involves speech processing on a portal server and is suitable for mobile phones.

The migration of speech recognition to smaller devices, such as handhelds and smart phones was enabled by the introduction of efficient speech recognition engines that can better handle noise and variations in speech, as well as, by the development of greater computing power for small devices, and faster processors for speech engines. It was projected by the Kelsey Group that software licenses from embedded speech applications would grow from US $ 8 million in 2002 to $227 million in 2006. This would make the embedded speech recognition industry one of the fastest-growing segments of the speech market [6].

Currently there are two markup languages proposed for creating applications that use voice input (speech recognition) and output (speech synthesis) and support multimodal interaction. Speech Application Language Tags language (SALT) is a lightweight set of extensions to existing markup languages, in particular HTML and XHTML (XHTML is essentially HTML 4.0 adjusted to comply with the rules of XML) that enable multimodal and telephony access to information, applications and Web services from PCs, telephones, tablet PCs and handheld devices. SALT applications can be implemented

using the thin client model with speech processing done on the speech server [7].

Another markup language that is currently proposed for developing multimodal Web applications is VoiceXML + XHTML (X+V) [8]. It combines XHTML and a subset of VoiceXML (Voice Extensible Markup Language). Currently VoiceXML is the major W3C standards effort for voice-based services [9]. VoiceXML provides an easy, standardized format for building speech-based applications. Together, XHTML and VoiceXML (X+V) enable Web developers to add voice input and output to traditional, graphically based Web pages. This allows the development of multimodal applications for mobile devices based on the fat client model that includes a multimodal browser and embedded speech recognition on a mobile device, and a Web application server (Figure 1).

While both X+V and SALT use W3C standards for grammar and speech synthesis, only X+V is based entirely on standardized languages. X+V's modular architecture makes it very simple to separate an X+V application into different components. As a result, X+V applications can be coded in parts, with experts in voice programming developing voice elements and experts in visual programming developing visual ones. X+V's modularity also makes it adaptable to stand-alone voice application development. Another feature of X+V is that it leverages open industry APIs like the W3C DOM to create interoperable web content that can be deployed across a variety of end-user devices [10].

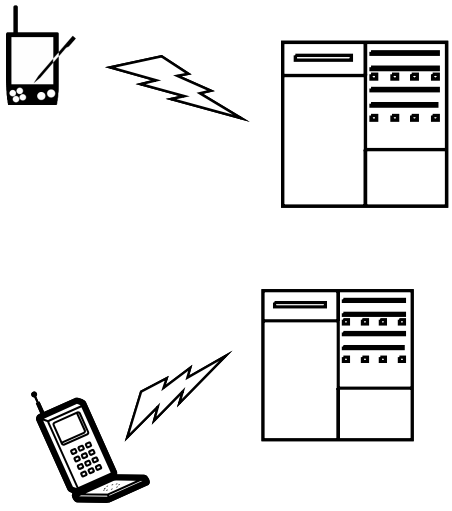


Figure 1. X+V architecture for multiple devices

At the same time, SALT's reliance on the containing environment makes it very difficult to separate out its coding functions, and makes the language insufficient for the task of stand-alone application development. As a result, the application developer must generate different versions of the application for each execution environment (for example, mobile phones or PDAs from different manufacturers).

2.2 Field Data Collection

To facilitate speedy field data collection and timely decision making, especially in the case of field inspection, where information is largely based on visual observations, it would be highly beneficial to use multimodal wireless handheld devices capable of delivering, voice, text, graphics and even video. For example, "hands free" voice input can be used by an engineer to request information using a hybrid phone-enabled PDA and a wireless, Bluetooth technology enabled headset piece. Then, the requested information can be delivered as text, picture, CAD drawing, or video, if needed, directly to the PDA screen [11].

By combining a multimodal mobile handheld device with a GPS receiver and a Pocket GIS system, the gathered inspection information could be automatically linked to its exact geographical location. For example, a handheld computer with a GPS receiver was used for construction damage assessment after the September 11 terrorist attack [4]. In addition, other environmental sensors, such as temperature and moisture sensors, or accelerometers could also be connected to a handheld device, if needed. Some examples of the location-referenced field applications include field data collection forms, control of environmental sampling during site inspection, on-site training, etc. [12].

Our current project on wireless, field quality control data collection is based on concepts of multimodal and voice field data collection. In this project, a field concrete testing technician will be able to enter field testing results using variable interaction modes such as speech, stylus and keyboard on the handheld device or using speech on the mobile phone. Our multimodal data collection application includes fat wireless client on a Pocket PC that has a multimodal browser and embedded speech recognition, and is based on X+V technology. The voice-only data collection application is based on the VoiceXML technology that allows data retrieval and input using natural speech on the mobile phone. Our previous work on the speech-based data retrieval and input applications for mobile workers is described below.

3. Speech-Enabled Field Data Collection

In spite of recent progress made in the introduction of information technology, the telephone is still the most widely used and the most effective information

communication tool in the construction industry [13]. Thus, a speech processing technology, that can utilize the ease and convenience of mobile phone use and enable real-time communication of field data and access to the wealth of information stored in the corporate data repositories, deserves special attention.

3.1 VoiceXML Technology

VoiceXML technology follows the same model as the HTML and Web browser technologies. Similar to HTML, a VoiceXML application does not contain any platform specific knowledge for processing the content; it also does not have platform specific processing capability. This ability is provided through the Voice XML Gateway that incorporates Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) engines. The VoiceXML Gateway architecture model is shown in Figure 2.

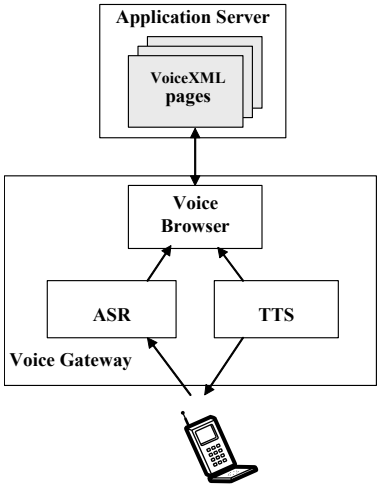


Figure 2. VoiceXML architecture model

VoiceXML allows providers to deliver Web services using voice user interfaces (VUIs). Developers can use VoiceXML to create audio dialogues that feature synthesized speech, digitized audio, recognition of spoken and touchtone key input (DMTF), recording of spoken input, telephony, and mixed-initiative conversations [14]. The words or phrases that a VoiceXML application must recognize are included in a grammar. Large grammars can cause application problems because they can result in recognition errors. Small grammars can cause VUI problems because they require prescriptive prompts that limit the use of natural language dialog. However, small grammars could be used successfully in designing applications for industrial users that are trained in using the application [15].

The advantage of using the VoiceXML language to build voice-enabled services is that companies can build automated voice services using the same technology they use to create visual Web sites, significantly reducing the cost of construction of corporate voice portals. A voice portal provides telephone users, including mobile phone users, with a speech interface to access and retrieve Web content.

3.2 VoiceXML for Field Applications

There are quite a few existing speech recognition applications that use VoiceXML technology to provide voice-enabled services to customers. Most frequently VoiceXML is used for building information services. These services, offered by wireless service providers, allow wireless customers to access news, email, weather, tourist and entertainment information, and business directories.

However, there are only a handful of existing applications that utilize the potential of speech-based information retrieval for industrial purposes. For example, Florida USA Power and Light Co. is using a VoiceXML based system for field restoration crews [16]. Using mobile phones, restoration crews can find out about storm-damaged equipment, and report back to the system on the status of the job.

Considering the widespread use of the mobile phone in industrial field applications, there is an opportunity to apply VoiceXML technology for field applications in construction, manufacturing, power and resource industries. These industries can benefit from voice-enabling their operations. The ongoing NRC research program on Voice and Multimodal communications specifically targets industrial field applications of Voice Web technologies.

3.3 Voice Inventory Management System

The Voice Inventory Management System (VIMS) prototype, developed at NRC IIT e-Business, allows a mobile worker to easily retrieve product and warehouse information out of the Web-based warehouse database, in real-time, using a regular, mobile phone, or phone-enabled handheld device and natural speech dialog [15].

The VIMS application keeps track of a series of products and warehouses in a database. All products in the database are entered into the VIMS speech recognition grammar, so that the grammar is updated dynamically with the information on current products in the database. Each product and warehouse has a number of attributes. Each product has a price, product number and description and is associated with the warehouses that product is located in. Each warehouse has an address, and the contents of that warehouse. The system also keeps track

of product types, represented by a tree that links particular types of products together.

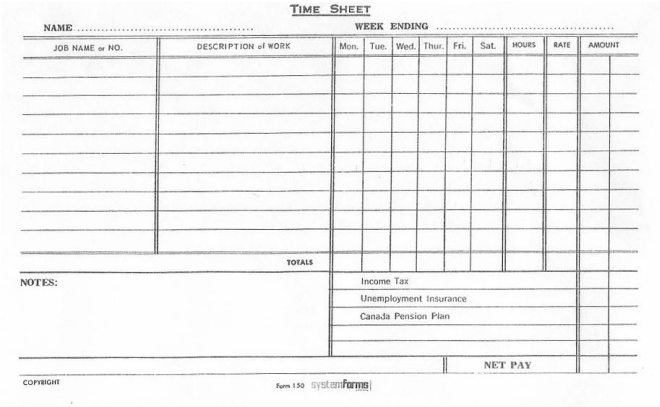
The system has an authentication menu to only allow authorized access to the system. To enter the system, users must say their name and a personal identification number. This number could be spoken or entered using a touchtone keypad. In the future it is planned to augment the authentication option with the voice recognition module to allow for voice print authentication in order to improve access security.

The authorized user has the option of browsing through a hierarchical menu of product types, which is useful if the name of a product is not known. The user of the system can also ask directly about a particular product or warehouse and request information regarding that product or warehouse. A natural dialog is used. For example, a user looking for information on aluminum window frames might ask, "What is the price of an aluminum window frame?" or "What warehouses are the aluminum window frames located in?"

In addition, an extensive listing system incorporated into the program allows the user to browse an alphabetical listing of all products or warehouses stored in the system. As searching through an alphabetical list of dozens of items can be time consuming, the user also has the ability to skip to items beginning with a certain letter.

3.4. Voice Time Management System

The Voice Time Management System (VTMS) prototype was developed to allow field crewmembers to enter their time in the time management application, on the corporate server, using a mobile phone. This technology, if implemented, could potentially allow timely billing of the client for the field services provided and bring substantial savings to field service providers. The system was designed to fill information as required by the standard timesheet for a construction project (Figure 3).


TIME SHEET

NAME ..... WEEK ENDING .....

JOB NAME or NO.	DESCRIPTION of WORK	Mon.	Tue.	Wed.	Thur.	Fri.	Sat.	HOURS	RATE	AMOUNT
	TOTALS									

NOTES:

Income Tax

Unemployment Insurance

Canada Pension Plan

NET PAY

COPYRIGHT

Form 150 systemforms

Figure 3. A timesheet used by construction workers

VTMS is designed to keep track of the user's hours for each job, work number and day and store this information in the corporate database. Each job and work type has a number of attributes. Each job has a job number and a description and is associated with a unique user. Each work has a work number, a job number, a description and days of the week. The job number attribute, from the work field, is what links the work field to the jobs that they are under. The control flow diagram for the VTMS application is shown in Figure 4.

Using voice commands with VTMS, a field service worker can retrieve and input information from a timesheet that is unique to the caller. A user can retrieve information such as hours worked on a particular day for a unique work number. A user can also find out the total hours worked for the week and the total hours worked on a particular work number for a job and update time information such as hours worked on a particular day.

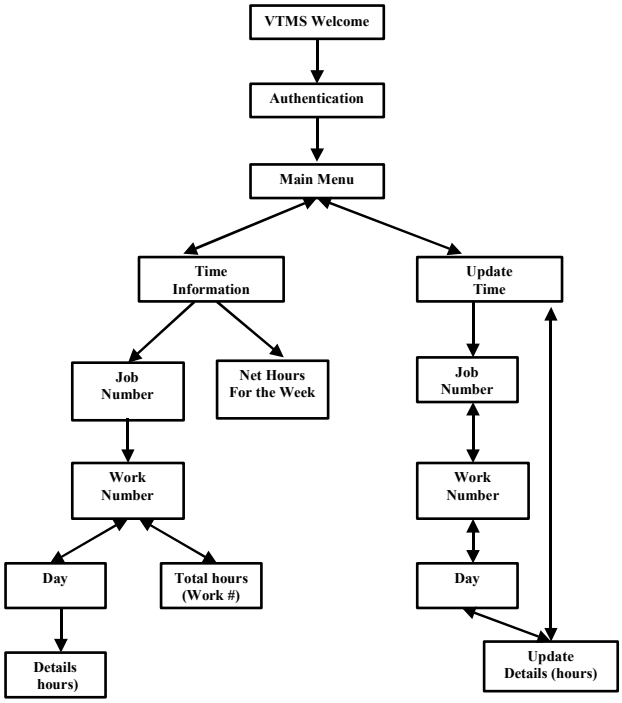


Figure 4. Control flow diagram for the VTMS system

3.5. Voice User Interface Design

In spite of the current improvements in speech recognition technology, it is still best suited for producing niche market solutions for people that have their hands busy in some activity or people that cannot use their hands [16]. Speech technology brings its own unique set of human factor issues. These issues include handling unpredictable errors produced by the speech recognition system; managing high human expectations of computers and speech by maintaining a realistic outlook on what a

computer can understand and respond to, and teaching users to use constrained speech to get correct results; minimizing the memory load while using the speech-only interface by limiting the number of menu choices that the user has to memorize; as well as, providing feedback to the user during processing delays in speech applications [18].

To uncover the problems with the design of speech applications, it is important to conduct usability studies in the environments in which real users will use the application. To design an effective usability study it is customary to conduct a pilot usability study that helps to refine the procedures and the design of user questionnaires. The pilot usability study conducted for the prototype VoiceXML application will be discussed later in this paper.

4. Performance and Usability Testing

Performance and usability tests were performed on prototypes to investigate the accuracy of speech recognition performance in a noisy environment and to conduct a pilot usability testing.

4.1 Performance Testing

The results of the speech recognition accuracy testing for the VIMS prototype showed that a native English speaker using a VIMS prototype system, even in a noisy industrial environment, retrieves the required product information 95 times out of 100. Subsequent testing, conducted using the computer product database containing more than 1000 items, indicated no decrease in the accuracy of speech recognition for a larger database. However, the processing time increased with the increase in the size of the database due to increases in the size of the grammar. This in some cases led to an increase in system response time. In our pilot testing program, we were not able to control system response time due to the use of the online development environment provided by BeVocal Café. Thus, it was difficult to contribute time delays to one particular cause. A new testing program, utilising in-house VoiceXML Gateway by VoiceGenie, will allow us to better address possible system response time issues.

It is possible to create sub-grammars in a VoiceXML application that shorten grammar-matching time, but make the navigation through the voice interface more complex. The grammar and the VUI design are very important components of a VoiceXML application and will require additional research work and testing.

VIMS and VTMS applications are designed to be used in a variety of locations with different mobile telephony equipment, including cellular phones and phone-enabled handheld devices. Testing on mobile devices such as the RIM BlackBerry Wireless and the HandSpring Treo has been done for the VIMS prototype, using the Voice Genie VoiceXML platform. This test was primarily done to see what impact noise had on the operations of the application. All devices performed well under "industrial noise" conditions in laboratory testing. The "noisy" conditions (about 60-70 dB) were created by broadcasting pre-recorded industrial noise in a soundproof room used for testing. Further testing of the prototype, with different mobile devices, needs to be done in the field.

4. 2 Pilot Usability Study

A pilot, usability testing study for a VIMS system, that provides voice access to a database containing entertainment products such as videos and books, was carried out on a group of Web-proficient, university student users. Study participants included two female and ten male students of 18-21 years old, representing a wide range of Faculties (Computer Science, Engineering, Arts etc.). Two out of twelve participants never used an automated telephone system before.

At the beginning of the test participants were given a short demonstration of the VIMS application and learned about the navigation choices available, without the researcher showing the diagram of the navigational structure. After the introduction, they were given a simple task of ordering a product and finding out the address of the warehouse where the product was located. They were also asked to sketch the navigation structure of the VIMS application and answer a questionnaire. This testing showed that, with minimal training, the users could easily navigate through the voice interface, draw a sketch of the VIMS navigational structure and retrieve required information on the products and on warehouses where these products were located.

All study participants were satisfied with the quality of the speech recognition, half of the participants found the quality of speech recognition to be "fair" and the rest found it to be "good" or "very good". Most negative comments were about the system's response time that was sometimes too long because the host BeVocal Café server was overloaded. A further, wider scale, usability testing program will be carried out on a group of field users to evaluate the feasibility of using VoiceXML technology in a carefully controlled laboratory environment and in the field.

5. Conclusions

The advantages afforded by the field use of VoiceXML technology to retrieve corporate and project information could be substantial. The software application developed using

VoiceXML allows "hands free" information retrieval and thus can help to overcome the limitations of the user interface of a field handheld computer [19]. However, VoiceXML technology is limited to only one form of input and output - human voice. Investigation of current multimodal technologies, that may help to overcome these limitations, including X+V and SALT, is one of the goals of the ongoing research program. The future usability testing program will be carried out using an in-house dedicated VoceXML Gateway server that should eliminate system response delays due to high server load. It would be also useful to conduct usability testing for applications with different grammar and navigation options such as small grammar modules, short response time, long navigation time vs. large grammar, longer response time, but shorter navigation time: to see what would be a preferable choice for the users. In addition to this, more prolonged studies are needed to evaluate the quality of speech recognition for an industrial user, after prolonged usage of the system during the workday.

5. Acknowledgments

The author would like to acknowledge the support provided for the project by the National Research Council Canada, valuable input from my colleagues at NRC on the VIMS application and the study design, and the hard work and dedication of the University of New Brunswick and Acadia University Computer Science students that participated in the development of prototypes and conducted the pilot usability study.

References:

[1] Burkhahardt, J., Henn, H., Hepper, S. and Rintdorff, K., *Pervasive Computing* (Boston, NJ: Addison –Wesley, 2002).

[2] IDC Viewpoint. *Five Segments Will Lead Software Out of the Complexity Crisis*, by A.C. Picardi, December 2002, Doc #VWP000148, 2002.

[3] Meissner A., Mathes I., Baxavanaki L., Dore G. and Branki C. The COSMOS integrated IT solution at railway and motorway construction sites - a case study, *Proc. of the Conference on eWork and eBusiness in AEC* (Turk and Scherer, editors), Swets & Zietilinger, Lisse, 2002, 623-626.

[4] Bacheldor, B. 2002. Handheld system assesses damage to see how buildings survived, *Information Week*, March 18, 2002.

[5] Hjelm, J. 2000. Research Applications in the Mobile Environment, in *Wireless Information Services* (New York: John Wiley & Sons, 2000.

[6] Kumagai, J. 2002. Talk to the machine, *IEEE Spectrum Magazine Special R&D Report on Leading Edge Technologies*, September 2002, 60-64.

[7] Moraes. 2002. VoiceXML, CCXML, SALT. Architectural tools for enabling speech applications, *XML Journal*, Sept. 2002, 30-25.

[8] W3C. Multimodal Activity, X+V, http://www.w3.org/TR/xhtml+voice/.

[9] W3C. Voice Browser Activity - Voice enabling the Web!, http: www.w3org/Voice.

[10] VoiceXMLForum. XHTML+Voice Profile 1.2.

http://www.voicexml.org/specs/multimodal/x+v/12/.

[11] Rankin, J. 2002. Information mobility for the construction industry, in Integrated Technologies for Construction, *Canadian Civil Engineer*, spring issue, 2002.

[12] Giroux, S., Moulin, C., Sanna, R. & Pintus, A. 2002. Mobile Lessons: Lessons based on geo-referenced information, *Proc. of E-Learn 2002*, Montreal, Canada, 331-338.

[13] Egbu C.O. and Boterill K. Information technologies for knowledge management: their usage and effectiveness, *ITcon*, Vol.7, 2000, 125-136, http://www.itcon.org/

[14] Beasly, R., Farley, M., O'Reilly, J. & Squire, L. *Voice Application Development with VoiceXML* (SAM Publishing: 2002).

[15] Kondratova, I., Voice and Multimodal Access to AEC Project Information, Mobile Computing in Architectural, Engineering and Construction, *10th ISPE International Conference On Concurrent Engineering*, J. Cha et al. (eds), Swets & Zeitlinger, Lisse, Portugal, 2003, 755-760.

[16] Datria.2001.http://www.datria.com/company/press/floridapowerandlight.htm

[17] Weinschenk, S. & Barker, D.T., *Designing Effective Speech Interfaces*, 2000.

[18] Lai, J. & Yankelovich, N. Conversational Speech Interfaces, in *The Human Computer Interaction Handbook*, (Lawrence Erlbaum Associates Publishers: 2000), 698-713.

[19] Srinivasan, S. & Brown, E. 2002. Is speech recognition becoming mainstream? *Computer Magazine*, April 2002, 38-41.