DPRO-110683

Deborah A. Hess,
Russell Ruggiero

Technology Overview
11 October 2002

Historical

VoiceXML and Multimodal Applications: An Introduction

Summary

VoiceXML, along with multimodal access, promises to bring the Web to almost any device imaginable. We explain the technologies and their applications.

Table of Contents

List Of Figures

Technology Basics

VoiceXML

Voice Extensible Markup Language (VoiceXML) is technology that lets Web sites provide interactive voice services to users over telephones and devices that incorporate telephony features. In basic terms, VoiceXML is an XML markup language for creating interactions that can include synthesized speech, digitized audio, recognizing and recording spoken input and input from dual-tone multifrequency (DTMF) (or “touch-tone”) keyed devices, telephony services and mixed-initiative conversations. User input is restricted to voice and DTMF formats.

VoiceXML can collect character and/or spoken input, assign the input to specific elements or tags within an XML document and make decisions that affect how these documents are interpreted. The technology brings the advantages of Web-based development and content delivery to interactive voice response (IVR) applications. The World Wide Web Consortium (W3C) oversees VoiceXML and related specifications.

Input and Output

VoiceXML applications process two types of input from telephones: voice input and DTMF “touch-tones.” Voice input requires speech recognition, which, in turn, must use grammars that help the application differentiate between extraneous sounds and valid input. Touch-tone input requires specialized hardware and software support.

VoiceXML applications generate two types of audible output. The first, speech synthesis, provides automated speech from words in texts and is well-suited for dynamic output applications and for testing and refining voice-based services. The second, audio playback, uses a prerecorded audio file and is best used for static content or situations requiring excellent voice quality.

VoiceXML Services

VoiceXML products provide the following:

- Text-to-speech (TTS) (or synthetic speech) output
- Speech input recognition and recording
- Audio file output
- DTMF input recognition
- Dialog flow control
- Telephony features such as call transfer, hold and disconnect

Basic Concepts

A VoiceXML *session* starts when a user interacts with a VoiceXML system. During the session, the system uses one or more VoiceXML documents. The user, a VoiceXML document or the interpreter can end the session.

Documents and Applications

VoiceXML data may be included in a single document or a set of documents known as an *application*. If a VoiceXML application incorporates multiple documents, one document becomes the application *root* (or parent) document and the rest become application *leaf* (or child) documents.

Each VoiceXML document starts with a *vxml* element at the top. This element identifies the version of VoiceXML in use and the address, or namespace, where the referencing document is located. If a document is a leaf, its *vxml* element must identify the root document.

Dialogs, Forms, Menus and Sub-Dialogs

The *vxml* element is primarily a placeholder for dialogs, which govern the flow of the document. A VXML document or application is actually a type of finite state machine, where each dialog represents a state. The user must always be in one of the states identified in the document.

There are two types of dialogs. *Forms* provide information, collect user input and interpret the meaning of the input. *Menus* provide a way to navigate through a series of alternative choices, prompt the user to make a choice and transition to another dialog in the same or a different document according to what is chosen.

Sub-dialogs transfer control to a new dialog and then return to the original dialog—much like subroutines or function calls. A sub-dialog might confirm a user's action or handle specific tasks. Developers can create libraries of reusable sub-dialogs for performing recurring tasks.

Fields

VoiceXML documents usually contain *fields* for handling input or specifying prompts. A field declares a variable and identifies prompts, grammars (see corresponding section), touch-tone sequences, help messages and other event handlers for obtaining a value. Users must provide a value for each field before going to the next element in the document.

Blocks

Most documents include *block* elements, which contain executable code for applying logic or translating between data, speech or DTMF touch-tones. A document may be linked to other documents through uniform resource identifiers (URIs).

Tasks

VoiceXML provides the necessary elements for handling common events that may take place during a typical user session. These include the following:

- Iterations and looping (*if, then, else, elseif*)
- Taking actions (*enumerate, transfer, option, submit, choice*)
- Jumping to another location in the same document (*goto*)
- Connecting a user with another voice application, phone line or other entity (*transfer*)
- Handling exceptions and events (*throw, catch, error, nomatch, noinput, link*)
- Declaring variables (*var*) to hold data and assigning values (*value*)
- Activating prerecorded or synthetic speech prompts (*prompt*)
- Accessing platform-specific functions (*object*)

- Reusing dialogs and creating reusable applications (*dialog*)
- Defining metadata using a schema (*meta, metadata*).
- Collecting and recording audio input from a user (*record*)
- Ending a session with a user (*exit, disconnect*)

Referencing external grammars within local document grammars (*import*)

Grammars

VoiceXML applications use grammars to recognize and process input. A voice grammar consists of a set of valid spoken inputs, procedures for evaluating these inputs and the actions to be taken when each input occurs. The grammar should also indicate the most common ways that users can speak responses to system prompts (that is, “yes,” “yup,” “right,” “yeah,” “OK,” and so forth), and most VoiceXML grammars are already prewritten.

- DTMF grammars are also available for processing touch-tone input. These specify sequences of key presses and assign a semantic interpretation for each sequence.
- Both voice and DTMF grammars can be used within the application or referenced by an external namespace and can be interspersed.

How to Use VoiceXML

Users access VoiceXML-enabled Web sites by dialing a telephone number, which connects to an application running in a voice portal on a VoiceXML server. The voice portal—which provides signal processing and interfaces between the telecommunications system and the VoiceXML server—interacts with the Web application. In many cases, an outside service provider operates a voice portal that can handle hundreds of processing requests from different organizations. The VoiceXML server normally hands off internal application processing by invoking an application server or an integration broker. The Web application, in turn, conducts the interactions with and delivers the services to the user.

Multimodal Technology

On the heels of current advances in voice technology, there is now considerable interest in combining speech with other types of interactions. Multimodal technologies—sometimes referred to as “ubiquitous computing”—are still very immature. Although individual vendors have been defining specifications and feature sets, standards bodies have just begun work to oversee and standardize this work.

The relatively new W3C Working Group known as Multimodal Interaction Activity is spearheading most of this effort with strategies to extend Web interaction—and XML markup technology—to voice as well as more traditional input modalities like keypads, keyboards, styli and mice. These modalities may be available on a single device—such as a speech-recognition-equipped computer that responds to voice commands—or with a combination of several devices working in tandem (that is, speaking into a handheld phone and viewing results on a personal digital assistant [PDA]).

Synchronized Multimedia Integration Language (SMIL) is one of the key technologies enabling multimodal computing. SMIL is an XML language that coordinates multimedia client interfaces, including advanced interfaces simultaneously incorporating both voice and “written” data. (The latter “written” part of the interface appearing on a display is often called digital ink). A multiclient system using XML and Extensible Stylesheet Language (XSL) can easily generate SMIL markup.

The W3C Multimodal Interaction Activity is also creating markup-based mechanisms for synchronizing information across disparate modalities and devices. One idea combines SMIL, display markup and forms markup with speech synthesis and recognition; another allows loose coupling between VoiceXML dialogs and visual interaction; and yet another provides a “digital ink” output component to multimodal Web applications. Voice can be used by itself for data input, while a user might create drawings with a stylus or use a mouse to create math equations or tables.

There is significant interest in using speech interaction with Web-based applications and services. The W3C currently supports a Voice Browser Activity organization, which is creating requirements and specifications for a new W3C Speech Interface Framework.

Operating Requirements

A VoiceXML 2.0 interpreter device must include the following technologies:

- **Document Acquisition**: The VoiceXML interpreter must be capable of receiving input through HTTP, telephony devices, telecommunications networks, voice portals or other VoiceXML documents.
- **Audio Input**: The VoiceXML product must simultaneously detect and report both character and spoken input. This can include typed characters and speech grammar data as defined in the W3C Speech Recognition Grammar Specifications. VoiceXML elements can include speech grammar data or make reference to such data via an external URI. A voice browser—usually operated by an outside source—handles the electromechanical transduction.
- **Audio Output**: VoiceXML products must be capable of working with audio (that is, WAV) files and TTS output. The VoiceXML specification lists audio formats, which the interpreter must support.
- **Audio Recording**: VoiceXML devices or products must be capable of recording audio sent by a user. The recording entity must support the same audio formats as the audio output entity.
- **Transfer**: VoiceXML devices must be capable of making third-party connections via telephone or other communications networks.
- **Processing Engines**: These are XML and XSL proxies and parsers that process requests from the interpreter device and format results for the interpreter to send out over the Web.

Technology Analysis

History

AT&T, IBM, Lucent Technologies and Motorola were the original proponents of VoiceXML version 1.0, which was submitted to the W3C on 17 March 2000. Subsequently, at its 10 to 12 May 2000 meetings in Paris, the W3C’s Voice Browser Working Group agreed to adopt VoiceXML 1.0 as the basis for the development of a W3C dialog markup language. Version 2.0 of VoiceXML was released as a Working Draft by the W3C in October 2001, and it is now in review. A finalized W3C “Recommendation” of version 2.0 is expected by the fourth quarter of 2002.

VoiceXML Architecture: How It Works

VoiceXML’s main objective is to free the authors of interactive voice response applications from low-level programming and resource management. It integrates voice and data services with data services using the familiar client/server paradigm.

A voice service is a sequence of interaction dialogs between a user and an implementation platform. The dialogs are provided by document servers that may be external to the implementation platform. Document servers maintain overall service logic, perform database and legacy system operations and produce dialogs.

A VoiceXML document specifies each interaction dialog to be conducted by a VoiceXML interpreter. User input affects dialog interpretation and is collected into requests submitted to a document server. The document server replies with another VoiceXML document to continue the user's session with other dialogs.

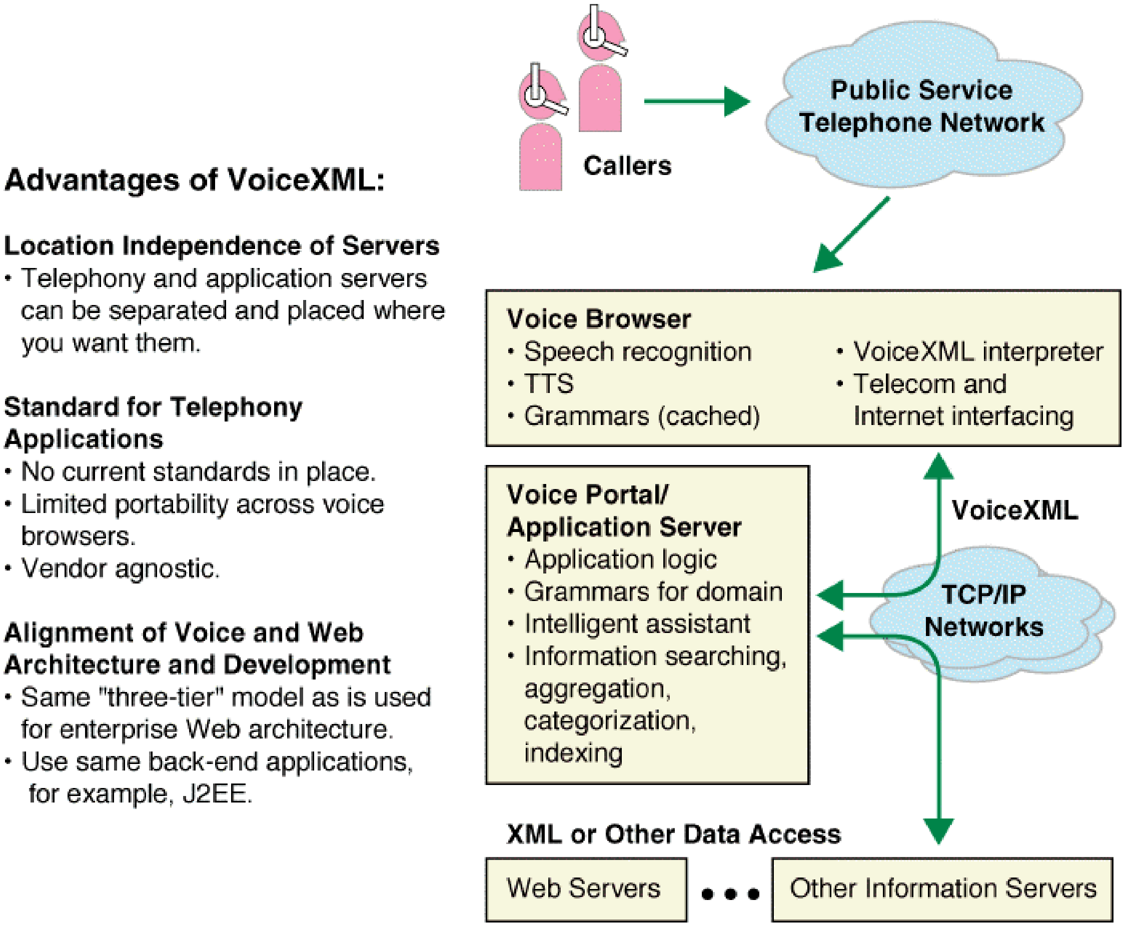
VoiceXML and Voice Portals

Gartner defines a voice portal as a system that provides telephone access to Internet- and intranet-based information, and which relies on speech technologies to improve the user interface. The information used by voice portal applications is stored on back-end Web and database servers.

VoiceXML is emerging as an important standard for enterprise voice portals. Three key advantages of VoiceXML are, first, that it allows the telephony resources to be separated from the application. This wasn't possible with earlier interactive voice recognition implementations. Second, it provides an element of application portability between vendor platforms. Third, it allows the same architecture to be used for Web and telephony services.

Figure "*Enterprise Voice Portal Model*" shows the model for an enterprise voice portal.

Figure 1: Enterprise Voice Portal Model



Advantages of VoiceXML:

Location Independence of Servers
- Telephony and application servers can be separated and placed where you want them.

Standard for Telephony Applications
- No current standards in place.
- Limited portability across voice browsers.
- Vendor agnostic.

Alignment of Voice and Web Architecture and Development
- Same "three-tier" model as is used for enterprise Web architecture.
- Use same back-end applications, for example, J2EE.

VoiceXML and Multimodal Applications: An Introduction

The Need for Multimodal Access

Over the past several years, computing has expanded from relatively large-footprint, stationary computers and laptops to increasingly smaller devices like PDAs, cellular phones, pagers and new products that combine features from several different devices. The newest of these products have built-in or optional Web connectivity. Ever since, there has been tremendous interest in bringing speech processing to the Web for a number of reasons, including the following:

- Stylus input and small-format keypads make text input extremely slow.
- An increasing need for Web connectivity in automobiles or other situations where hands and eyes are unavailable.
- Applications—such as drawings or scientific notations—which require both speech and "digital ink."

Multimodal Specifications

IBM, Motorola and Opera Software have drafted a proposal for combining VoiceXML with Extensible Hypertext Markup Language (XHTML). This proposed relationship leverages the strengths of VoiceXML for building applications that accept voice and DTMF input and produce synthesized/recorded output, and the use of XHTML to build visual applications that accept mouse and keyboard input and produce visual output. The combination of both standards has been termed as a "Multimodal Standard" by IBM.

The W3C has recently accepted this multimodal proposal, and a dedicated W3C Working Group—the W3C Multimodal Interaction Activity Working Group—formed in February 2002 will look into standards and software to access Web applications and services by voice, keyboard, keypad, mobile phones and devices.

Business Use

Voice and Multimodal Applications

VoiceXML is a Web-enabled XML protocol for telephony and speech recognition devices. The technology should eventually be capable of accommodating the vast majority of voice response services. On the other hand, services with extremely stringent requirements may best be served by dedicated applications providing a finer level of control.

The proposed "Multimodal Standard" combines VoiceXML with XHTML (these are HTML commands formatted as XML). There are a number of scenarios where multimodal technology applies:

- **Mobile stock trading:** Using voice to request stock quotes with a mobile computing device, having the quote appear as a chart and submitting a trade by voice.
- **Web-based auctions:** Using a mobile computing device to view a specific item and then bidding via voice commands.
- **Navigational systems in automobiles:** Voice-enabled navigational devices.
- **Mobile dictation:** Users can create documents outside the office, then use the Web to enter the information into a word processor or other productivity application.
- **Computing for visually impaired users.** Blind or visually impaired users can speak to the computer or have the computer read information from the display.
- **Web browsers in automobiles, trains or airplanes:** In the first case, the device automatically shuts off the graphical browser and switches to voice to ensure that the driver is not distracted. In the

remaining two cases, keyboards may be expensive to maintain or available space may make keyboarding impossible.

- **Collaborative processing between mobile and network devices**. A user may want to create voice or touch-tone input via mobile phone or PDA for use in a structured application. The device may not have enough space to store the type of complex VoiceXML or DTMF grammars to interpret the information. When the user connects to a Web-based VoiceXML server, the server invokes a remote recognizer, which provides the necessary grammar for bidirectional translation between audio and data.

Network VoiceXML

VoiceXML can be used in place of network prompting for cost savings. Network VoiceXML offers both hard and soft advantages over traditional network prompting. The primary hard cost advantage is that network prompting is based on per-prompt charges while most Network VoiceXML products allow unlimited prompting and caller interaction. A second hard savings of network VoiceXML is that it allows the enterprise to delay the routing of the call to a site until it is convenient to do so. This lets the enterprise obtain more information from the caller, reducing misdirects and permitting better use of resources.

A soft advantage of VoiceXML solutions is that enterprises retain control over their application and can execute it at their premises. Network prompting application changes can take several weeks to complete while VoiceXML changes can be implemented immediately. Finally, VoiceXML can support speech recognition as an option, unlike network prompting. Because Network VoiceXML doesn’t charge per prompt, the more prompts the greater the savings. Gartner research indicates that based on an average application profile, an enterprise can save $100,000 per million calls containing eight prompts.

Benefits and Risks

Benefits

VoiceXML Separates Interaction and Service Code

VoiceXML separates the user interaction code, written in XML, from the service logic usually formatted as Common Gateway Interface (CGI) scripts. That makes it easy to modify the nature of the interaction and is especially helpful for applications or documents requiring frequent changes. This also promotes service portability across implementation platforms, shields application authors from low-level signal processing details and minimizes device-to-server contact by allowing multiple interactions within a single document or application. In addition, VoiceXML can be used to quickly create simple interactions, but also provides features that support complex dialogs.

VoiceXML Applications Promote Reuse

There are several advantages to multidocument applications, primarily centering on inheritance and code reuse. Root document variables are available to leaf documents, property elements in the root provide default values for the same properties in the leaves, root documents can define default event handling for leaves and certain root document grammars allow users to interact directly with the root from inside a leaf document.

Leverages Established Web Technology

VoiceXML is a markup language based on commonly available Internet technology. This lets developers use their XML expertise, work with familiar Web development tools, integrate easily with back-end systems and publish applications to existing Web servers and many Internet portals. Future multimodal technology specifications are also likely to build on top of existing Web, Internet, I/O and telephony technologies.

Voice Access Is Inexpensive and Widely Available

The normal VoiceXML client is a telephone (or a device that includes telephony services), so there is almost no cost to the end user. In addition, organizations providing services can do so over an extremely accessible and available technology with very little investment in infrastructure. Organizations can arrange with an external provider for VoiceXML interpretation if they do not want to build their own platforms. Companies that currently support proprietary interactive voice response systems can move these to extensible, platform-neutral voice portals, which are easier and less expensive to maintain and upgrade than the voice response systems. Consumers will also benefit from a voice-accessible Internet; there will be many more applications that can be accessed by phone and users will be able to interact with multiple applications or services within a single phone call.

Risks

Very Immature Technologies

At the present time, there are very few specifications governing multimodal technology, and the W3C Multimodal Activity Group has just started to look at technologies and requirements. Although VoiceXML is considerably more advanced than multimodal technology, the relevant specifications are still in Working Draft status at the W3C. Also, there is still work to be done integrating VoiceXML with complementary XML technologies like XSL Transformations (XSLT), XForms, XPath, XPointer and Natural Language Semantics Markup Language.

(Attention Span)—Problems of Mobile Device Users

Unfortunately, speech recognition/response applications do not remove the risks associated with using a mobile phone while driving or engaged in attention-intensive activity. Studies show that speaking and listening require a significance proportion of the brain's capacity, reducing the user's ability to perform other tasks. In some localities, providing applications that might be seen as encouraging driver distraction could result in legal actions or other unintended problems.

Voice Recognition and Device Compatibility Issues

VoiceXML and, to a lesser extent, multimodal technologies are likely to stumble across the same voice recognition and device compatibility difficulties that plagued earlier products. VoiceXML will not improve speech recognition on phones and computing devices; there is still no fast and easy way to train VoiceXML interpreters, and sound-to-data transduction remains complex. In addition, these devices will still have to confront issues with multiple telephony standards (that is, Personal Conferencing Specification [PCS], Global System for Mobile Communications [GSM], and so forth) and the many proprietary—and incompatible—specifications for mobile and wireless hardware and software.

VoiceXML Application Development Complexity

Building a VoiceXML application is extremely difficult. The application must be created with VoiceXML-compatible tools, the client and server endpoints can require a lot of adjustment to maximize voice quality and speech recognition and the application itself has to be trained to recognize spoken input and correctly articulate spoken output. In addition, VoiceXML developers must address the usual Web application issues of scalability, availability, quality assurance and responsiveness. This makes voice applications much more complex than common Web applications.

Storage Limitations of Mobile Devices

An individual mobile device, such as a PDA, normally cannot recognize more than a few hundred spoken commands, nor does it have sufficient storage for prerecorded speech prompts that do not sound “robotic.” These devices can presently handle only very simple voice dialogs; rich voices and extensive interaction would have to be moved to a more powerful platform if available.

Standards

VoiceXML Documents and Committees

The W3C manages VoiceXML activity from its Voice Browser Working Group. The latest VoiceXML specification is VoiceXML 2.0, which is a W3C Working Draft.

Multimodal Documents and Committees

Dedicated multimodal technology standards efforts are still at a very early stage. Most of these target voice, graphical display and character I/O for portable devices like PDAs, cell phones, pagers and the like.

The W3C formed a Multimodal Interaction Activity Working Group in February 2002. At present the group has not published any drafts or requirements documents. There is a variety of related initiatives, mostly relating to voice I/O. These include the following:

- **Multimodal Access Position Paper** from Siemens, 26 November 2001. This describes a multimodal interaction scenario using a visual client and VoiceXML interpreter.

- **XHTML and Voice** from Motorola, IBM and Opera Software, which describes multimodal interaction markup using VoiceXML and XHTML. The trio submitted a paper on this technology to the W3C on 30 November 2001.

- **Speech Application Language Tags (SALT)** 1.0 from Microsoft is a proposal for speech standards using HTML and XHTML developed using .NET technologies.

- **Session Initiation Protocol (SIP)**: This is a text protocol, sponsored by a SIP Working Group, for managing interactive communication sessions. These sessions may include voice, video, chat, interactive games or virtual reality. The protocol is still in draft format. SIP can work with VoiceXML to provide infrastructure solutions for enterprise contact and call centers.

Technology Leaders

The VoiceXML Forum

The VoiceXML Forum, through its 600-plus members, is developing products and deploying applications built on the VoiceXML standard. The natural extension of the VoiceXML standard to support multimodal applications will hasten the expansion of combined voice and data applications. Forum members and supporters of VoiceXML will contribute to the W3C’s efforts to develop multimodal standards.

The W3C

The W3C was founded in 1994 to lead the Web to its full potential by developing common protocols that promote its development and ensure its interoperability. The W3C is an international industry consortium, which is jointly hosted by the Massachusetts Institute of Technology Laboratory for Computer Science (MIT/LCS) in the U.S., the Institut National de Recherche en Informatique et Automatique (INRIA) in Europe and the Keio University Shonan Fujisawa Campus in Japan.

The Consortium provides numerous services that include a repository of information regarding the Web for developers and users, reference code implementations to embody and promote standards and various prototype and sample applications to demonstrate the use of new technology. Membership in the W3C currently totals approximately 500 and includes industry leaders such as Cisco, IBM, Microsoft and Sun.

Other Players

Other leaders in the VoiceXML and multimodal space include these:

- BeVocal
- Call Interactive
- Convergis
- Genesys/Alcatel
- Tellme
- VoiceGenie
- Voxeo

Technology Alternatives

SALT 1.0

Microsoft and several partners support this initiative, which uses XML tags to mark up speech input. Microsoft offers a SALT development toolkit for .NET, and organizations looking into .NET for contact centers may be interested in the technology's progress. There is a SALT Forum managing the technology, and it recently submitted a SALT 1.0 draft to the W3C. W3C may well intend to combine SALT with other multimodal-related proposals into a single standard. At this point, there are no enterprise-class products using SALT.

Wireless Application Protocol (WAP) and Wireless Markup Language (WML)

WAP, which governs how wireless networks handle information delivery and telephony services on mobile phones, predates VoiceXML and multimodal computing by several years. Its oversight body is called the WAP Forum, which is a consortium of mobile phone vendors. WAP uses WML to tag information for display on mobile phones. WAP held great promise when first announced, but its incapability to overcome many of the difficulties of wireless phones such as incompatible technologies and formats, security, issues with WML and usability has generated less than stellar success. WAP 2.0 provides many improvements, such as compatibility with XHTML Basic, and appears to have better prospects than its predecessors.

Insight

As mobile, hands-free, wireless devices have proliferated, so has interest in using voice input and output to access the Web. VoiceXML, a markup language for telephone access to Web sites, is one of the most promising technologies in this area, especially since many of these devices offer styli or tiny keypads for data entry. VoiceXML is a voice application solution and provides three major advantages: First, telephony resources can be physically separated from the application. This useful because it lets organizations pick best-of-breed Voice Browser vendors and makes it less expensive to reuse telephony resources. Second, it provides, for the first time, the ability to use a standard interface to telephony resources. Third, organizations can leverage Internet resources, people and infrastructure, with telephony applications, resources and people. VoiceXML can be used by itself or for combining "digital ink" with verbal or touch-tone input and output. This has paved the way for a new concept, multimodal interaction, which allows users to work with the Web using a combination of speech, keyboards, touch-tones and styli. The W3C handles the VoiceXML specification, which is beginning to interest commercial software and hardware organizations, and has created a new W3C Multimodal Interaction Activity Working Group—which will propose standards and software to access Web applications and services using multiple technologies.