VoiceXML for Business Applications: A Survey

Sanwar Ali, Mohamed Albohali, and Kustim Wibowo
Indiana University of Pennsylvania
Indiana, PA 15705-1087

ABSTRACT

For almost a decade, voice recognition techniques enhanced business in many ways through telephone communication, especially in the area of customer services. Until recently, the advance of World Wide Web (WWW) has promoted business communication, both in business-to-business and business-to-customer relationships through computer displays that respond to keyboard and other input devices. VoiceXML promises a better business communication without any input devices except a microphone. This technique will help business applications to be more intuitively user-friendly which will enhance customer service in business applications. This paper presents the VoiceXML techniques, its advantages, limitations, and its current experimental implementation in certain business applications by some companies, along with some perspectives on how it could be incorporated into the area of business communication in general.

INTRODUCTION

More people have access to telephones than to computers with an internet connection. Also, the sales of cell phones are booming, and more and more handheld devices with varied browsing and display capabilities are emerging into the market place and into the hands of millions of consumers. This trend of mobile phone technology usage along with the current state of the Wireless Networking frenzy, offers clear potential and a proud range of possibilities for voice browser applications to accessing Web based services from anywhere and at any time. A voice browser also offers a great fit for the next generation of call centers, which will become voice web portals to the company’s services and related websites, whether accessed via the telephone network or via the Internet (www.w3.org/voice).

Voice browsers allow people to access the web using speech synthesis, pre-recorded audio, and speech recognition. Voice browsers may also be offered as an adjunct to conventional desktop browsers with high resolution displays, providing an accessible alternative to using keyboard or screen, for instance in automobiles where hand/eyes free condition is essential (www.w3.org/voice and www.xml.com).

SPEECH RECOGNITION

A spoken language is effective for human-to-human interaction but often has severe limitations when applied to human-computer interaction. The differences between human-human and human-computer interaction may be attributed to the rich emotional content conveyed by prosody language, or the pacing, intonation, and amplitude in spoken language. The emotion aspects of prosody are potent for human-human interaction but may be disruptive for human-computer interaction. The syntactic
aspects of prosody, such as rising tone for questions, are important for a system's recognition and generation of sentences. (Schneiderman, 2000).

**A GLANCE AT VOICEXML**

Historically, different speech companies created their own voice markup languages with different names. As companies integrated languages together, new names were given to the integrated languages. Some companies formed the VoiceXML forum and created VoiceXML (briefly known as VXML). The World Wide World Consortium Voice Browser Working Group has specified the Dialog MarkUp Language known as VoiceXML as a model.

The design and standardization process has followed from the Speech Synthesis Markup Requirements for Voice Markup Languages published in December 23, 1999 by the W3C Voice Browser Working Group. The following items were the key design criteria:

♦ **Consistency**: provide predictable control of voice across platforms and across speech synthesis implementations;
♦ **Interoperability**: support use along with other W3C specifications including (but not limited to) the Dialog Markup Language, Audio Cascading Style Sheets and SMIL (Synchronized Multimedia Integration Language);
♦ **Generality**: support speech output for a wide range of applications with varied speech content;
♦ **Internationalization**: enable speech output in large numbers of languages within or across documents;
♦ **Generalization and Readability**: support automatic generation and hand authoring of documents, which should be human-readable;

♦ **Implementation**: the specification should be compatible with existing, generally available technology and the number of optional features should be minimal (www.w3.org/tr/speech-synthesis <http://www.w3.org/tr/speech-synthesis>).

Figure 1: GUI and Voice-Based Internet Applications

In addition, Speech Synthesis Markup Language specification for voice browsers is designed to provide a rich, XML-based markup language for assisting the generation of synthetic speech in web and other applications. The essential role of the markup language is to provide a standard way to control aspects of speech such as document structure (paragraph, sentence, say-as, phoneme), prosody and style (voice, emphasis, break, prosody elements), and other speech elements (audio, mark, other miscellaneous relevant features). As shown in Figure 1, a wide range of applications can then be integrated into the web-voice accessible information resource center.

**WEB DELAYS AND BUSINESS COMMUNICATION OPPORTUNITIES**

In business-customer interactions customers expect their pages to pop-up immediately, and if they do not see them soon enough, they will waste no time searching elsewhere. Every frustrated surfer who cuts to another site is another potential customer lost (Pugh, 1999). One method to keep the page refreshed quickly is by so called CDNs (Content Delivery Networks) - installing servers on ISP networks all over the world so that content is as close to the customer as possible. On top of that, they
place probes and agents all along the net to gather information on congestion.

By nature, spoken input is slower than point and click mechanisms and naturally more intuitive. For example, in a single web page which is full of menus and other selections, a question coming out of the computer sound system such as: "What are you going to do today?" will naturally be more appealing to the user than scrolling through a long list of selections. Next, if the user’s response is: "I want check my Email!", the computer may respond by asking: "Did you say you want to check your email?" or "You have new email this morning. Do you want to read it yourself or should I read it for you?"

The above example is a transactional (two-way) communication process that actually can be broken into six phases. In short, these phases are: The sender has an idea, the sender transforms the idea into a message, the sender transmits the message through channel or medium, the receiver gets the message, the receiver interprets the message, and the receiver reacts and sends feedback to the sender through the same channel (Bovee and Thill, 1998). Each phase consumes a tiny slice of time. Depending on the conversation content and context, the whole process will take some time to complete. These phases loop back and forth interactively until both parties agree on what they are communicating. Therefore, the human-computer spoken interactive communication will enable the computer with language understanding software enough time to grasp the meaning of spoken word, analyze them and generate the appropriate feedback.

It is predicted that the next generation of speech-based interfaces will enable people to communicate with computers in much the same way that they communicate with other people (Zue 1999). Also, this (time) opportunity will provide the possibility to integrate the spoken instructions into the company’s Enterprise Resource Planning (Davenport 1998).

**IMPLEMENTATIONS**

Several companies have implemented VoiceXML 1.0 and they are extending their implementations to conform with the markup languages in the W3C Speech Interface Framework. Tellme Studio (<http://studio.tellme.com>) allows anyone to develop their own voice applications and access them over the phone. Motorola has the Mobile Application Development Toolkit (MADK), free downloadable software development kits that support VoiceXML (<http://www.motorola.com/>). The IBM Voice Server SDK Beta Program is available at <http://www.alphaworks.ibm.com/>. VoiceGenie is sponsoring a developer challenge in association with VoiceXML Central, a VoiceXML virtual community and search engine, see <http://developer.voicegenie.com>. And, there are many other companies that support spoken language interactions, such as AOL Instant Messenger.

IBM has several patents on its natural language understanding engine which uses probability to guess what people mean if the words are unclear. Philips offers Speech Pearl with a 200,000 word vocabulary, and Speechwave offers text to speech in 35 languages. Nuance, works with around 1,500 application companies and 1,500 platform companies. VocalPoint converts regular HTML text into a voice platform and claims
that they need only a day or about 10 hours of programming to set up a voice-web.

LIMITATIONS, ADVANTAGES, AND BUSINESS OPPORTUNITIES

Like any other new technology, VoiceXML has some limitations and offers some business opportunities. Although the Internet offers many new opportunities for e-commerce and e-business, it also presents a series of challenges. These challenges include unproven business models, business process change requirements, channel conflicts, technology hurdles, legal issues, security and privacy (Laudon and Laudon, 2000). Since this technology supports the communication conducted through the Internet, then it is bound to Internet limitations. The technology constraint is especially true on the network bandwidth, where, in this case no matter how high the capacity of the bandwidth it will always be filled up as soon as the devices can use it. In the wireless world, a phone is for two purposes: to relay data using sound and to relay data visually. Getting the two modes simultaneously is difficult. However, the future of voice browsing will be visual with the objective of hand-free interaction. Also, the implementation of VoiceXML needs additional computer resources. To be effective, it requires appropriate voice input-output devices. In addition, the quality and the location of the devices are very important for this type of communication. Here, location refers to the human-computer portion of communication and voice browsing. It is obvious that the parameters that are associated with the location of the communication itself such as the degree of isolation, uncontrolled crowd or noises, special spaces or room design, location of the devices relative to the user, are some of many factors that may play an important role in determining the quality and the success of any voice browsing session.

Innovation in accessing information via the auditory channels raises access challenges for users with hearing or speaking impairments. For a hearing-impaired user, synthesized text should be displayed visually. For a speaking-impaired user, verbal responses may be entered via a keyboard. The positive impact is on the user group facing the most challenges on the visual WWW - namely blind and low vision users.

The predicted growth of the speech technology market is so great that it makes the actual growth of the voice technology industry anyone’s guess. Cahner In-Stat Group predicts a $1.2 billion voice portal market by 2005. The Kelsey Group estimates the voice browsing market will be worth $6.5 billion by 2005 and generate $5 billion in e-commerce. UK based OVUM predicts a world market of $26 billion by 2005. And lastly, Allied Business Intelligence predicts 56 million mobile voice portal users in North America alone, with 250,000 voice sites and a $50 billion v-commerce market (Figures 2-4.)

Figure 2: Internet Use in the U.S.

Figure 3: How many people have phones

Figure 4: Last year’s users of WAP and Voice Portal

CONCLUSION

Speech is natural; we know how to speak before we know how to read and write. Speech is also efficient; most people can speak about five times faster than they
can type and almost 10 times faster than they can write. And, speech is flexible; we do not need to touch or see anything to participate in conversation. People visit websites and receive visual information. The voice web has voice sites where the information is conveyed through speech. Financially, voice browsing makes great sense for companies. It costs around a dollar a minute for human operators to interact with customers but only around 10 cents a minute for an automated system. In the United States, over 40% of homes are connected to the web via personal computers (Figure 2). In Europe, personal computer penetration rates are much lower and people are purchasing internet-ready phones to PC at a rate of 5 to 1. In Asia people are buying phones over PC's at 9 to 1 rate. Since almost everybody has and will have a phone, then everybody will have the ability to get on the web and will be able to browse websites that offer voice portals. Therefore, in the near future, the voice technology and v-commerce are predicted to grow in an exponential rate.

REFERENCES:


Figure 1: GUI and Voice-Based Internet Applications

Figure 2: Internet Use in the U.S.

Figure 3: How many people have phones
Figure 4: Last year’s users of WAP and Voice Portal

![Bar chart showing total number of users in 2000 (U.S.) for WAP and Voice Portal. Source: U.S. Wireless Carriers.](chart.png)