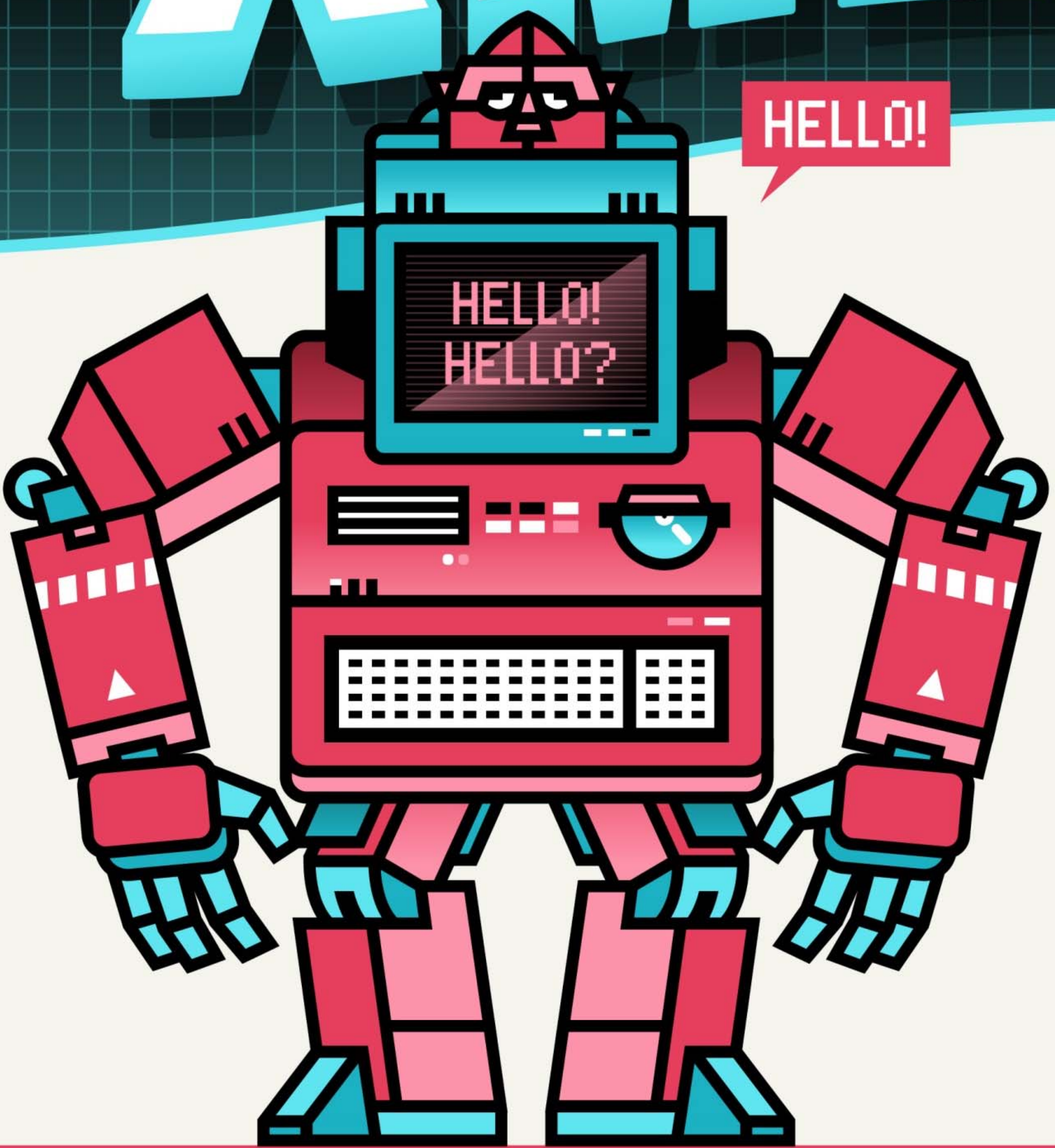
VOICE XML



sas
Specialty Answering Service

WWW.SPECIALTYANSWERINGSERVICE.NET

Contents

1 Introduction

Voice eXtensible Markup Language, aka VoiceXML, is a scripting language that can be thought of as the HTML of the voice web. It is an open standard markup language for creating voice-based applications such as IVR and web services. VoiceXML makes use of the existing web infrastructure for HTML, thus enabling it to create and deploy voice applications in an easy manner by hiding the underlying complexities of the telephony platform from developers. VoiceXML relies on a speech recognition engine and/or DTMF (Dual Tone Multi Frequency) for user input and uses pre-recorded audio or a text-to-speech (TTS) system for output. It provides a uniform standard for developing feature rich and media rich voice applications.

VoiceXML is the World Wide Web Consortium's (W3C) standard XML format for specifying interactive voice dialogues between a user and a computer. Just as HTML allows visual applications to be developed and employed over the web, VoiceXML allows voice applications to be developed and deployed. While a visual web browser interprets HTML pages, VoiceXML files are interpreted by a voice browser. Common deployment architecture for a VoiceXML application consists of a set of voice browsers connected to a Public Switched Telephone Network (PSTN), allowing users to interact with the applications through telephone lines.

Even today, penetration of landline and wireless telephones is higher than that of the Internet. As a result, it is vital that businesses provide customers with access to information over the telephone and not just through the Internet. With the convergence of Internet and telephony, the need for a telephone-based voice interface became evident to the business community. A voice-based interface would use existing investments made in Internet technologies and infrastructure to deliver information and conduct business in a cost-effective and user-friendly manner. VoiceXML was the result of those attempts to address this need.

VoiceXML was proposed by the VoiceXML Forum, and version 1.0 of the standard was published in March 2000. The VoiceXML Forum is a consortium of about 500 companies worldwide, and counting. Today, VoiceXML is the international standard for writing telephony-based voice applications. The standard is currently controlled by the W3C, while the VoiceXML Forum focuses on ensuring conformance to the standard, education, and marketing.

VoiceXML takes advantage of several trends in Internet and speech technologies, such as the growth of the World Wide Web, the advances in computer based speech synthesis and speech recognition, and the spread of mobile Internet.

The World Wide Web has improved in terms of content generation, content presentation, performance, bandwidth, and even the quality of service. We have moved away from static content to dynamic content created with the help of scripts and other server-side and client-side technologies. Web servers also provide access to huge databases of information. Voice XML uses both legacy as well as cutting-edge technologies for its operation.

Voice XML is based on XML, a flexible data representation structure with technologies that make it easy to transform an XML format to another XML or non-XML format. It takes advantage of the improved technologies that aid efficient audio data movement across the web. As new types of web applications and services emerge and web application development tools become more powerful, VoiceXML has become easier to adapt than before, leading to its widespread adoption.

2 Technology

VoiceXML relies heavily on voice-based technologies and the underlying Internet infrastructure. While HTML is based on an assumption of a graphical web browser with a display output and keyboard/mouse input, VoiceXML caters to a voice browser with audio output and audio/keyboard input. The audio input is analyzed by the voice browser's built-in speech recognizer, and audio output is generated either from a pre-existing database of recordings or from artificial speech synthesized by the voice browser's text-to-speech system. A voice browser runs on a specialized voice gateway node that can support several hundred simultaneous callers, and can be accessed through any telephone line. The gateway node is connected to both the Internet and to a PSTN (Public Switched Telephone Network).

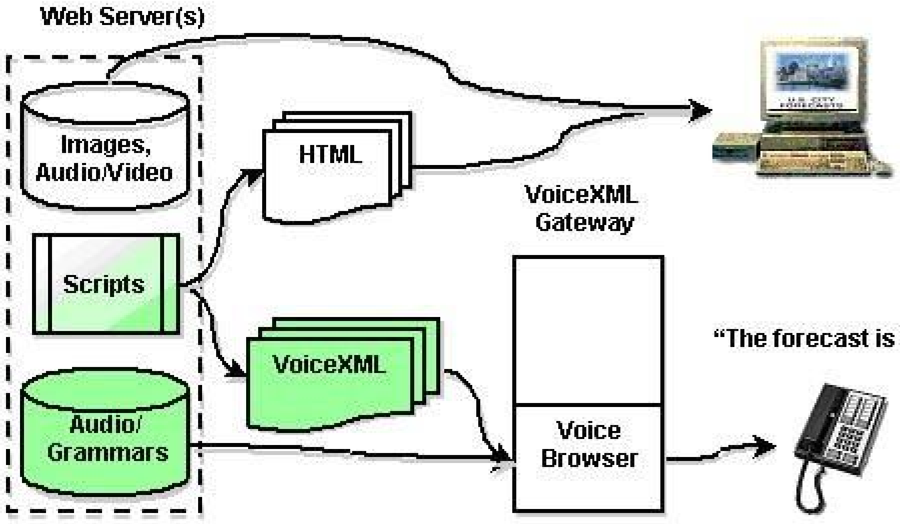


Figure 1: Voice Browser Architecture (Source: http://www.voicexml.org)

VoiceXML has tags that instruct the voice browser to provide various speech services such as speech recognition, speech synthesis, dialog management, and audio playback. The classic Hello World application that can be interpreted by a VoiceXML interpreter to output "Hello World" with artificial speech is given below:

```
<vxml version="2.0" xmlns="http://www.w3.org/2001/vxml">
 <form>
  <block>
   <prompt>
    Hello world!
   </prompt>
  </block>
 </form>
</vxml>
```

HTTP is the most commonly used transport protocol for fetching VoiceXML pages. Applications may use static VoiceXML pages or generate dynamic VoiceXML pages using an application server like Tomcat, Weblogic, IIS, or WebSphere.

2.1 VoiceXML Architecture

VoiceXML connects to a telephone network on one side, and a TCP/IP network and the corresponding application server on the other side.

A typical VoiceXML-based system consists of four main components:

- *Telephone Network*: VoiceXML can work with both a PSTN network as well as a VoIP packet network.
- *Application Server*: The application server is typically a web server that hosts the VoiceXML application and the business logic.
- *VoiceXML Gateway*: A VoiceXML gateway consists of a VoiceXML interpreter, which is integrated with speech resources (automatic speech recognition, text-to-speech, and audio playback) and telephony resources (DTMF, call control). A VoiceXML gateway downloads the applications from the application server and interprets them.
- *TCP/IP Network*: The underlying network of a VoiceXML application can be a LAN, WAN or even the public Internet.

In VoiceXML, the application logic is separated from the Voice Interface leading to two main advantages:

- Businesses can leverage their existing investments in web technologies and infrastructure while migrating to voice-based applications.
- Businesses can retain full control of their application logic (which is often a competitive advantage) and still have the flexibility to outsource the Voice Interface Design and hosting.

In addition, VoiceXML is an internationally accepted standard, allowing businesses to create ‘write once run anywhere’ applications. VoiceXML is a simple scripting language that is platform independent, thus offering the flexibility to choose the Speech and Telephony platform, and develop applications easily without worrying about platform related complexities.

2.2 Related Standards

The W3C's Speech Interface Framework defines several other standards that are closely associated with VoiceXML.

2.2.1 SRGS and SISR

SRGS or Speech Recognition Grammar Specification is used to define the grammar of sentences to the speech recognizer. These sentence patterns are used to determine the most probable sentence that it has heard as input. The semantic meaning of that sentence is then specified via the SISR (Semantic Interpretation for Speech Recognition) standard and fed into the VoiceXML interpreter. The set of ECMAScript assignments that create the semantic structure outputted by the speech recognizer is specified using SISR.

2.2.2 SSML

The Speech Synthesis Markup Language (SSML) is used to add additional information to the textual prompts. These help to add human characteristics to synthetic speech by selecting aspects such as the voice to be used or the volume of speech.

2.2.3 PLS

The Pronunciation Lexicon Specification (PLS) helps to define word pronunciations. This pronunciation information can be used by both speech recognizers and voice synthesizers in voice browsing applications.

2.2.4 CCXML

The Call Control eXtensible Markup Language (CCXML) is yet another W3C standard that can be used in VoiceXML platforms as well as for non-VoiceXML applications. It is usually used to handle initial call setup between the caller and the browser, and to provide services such as call transfer and disconnecting the voice browser.

2.2.5 MSML, MSCML, MediaCTRL

In order to address the deficiencies of VoiceXML in handling media server applications such as multi-party conferences, several companies have come up with their own scripting languages. These include Convedia's Media Server Markup Language (MSML) and Snowshore's Media Server Control Markup Language (MSCML). The languages are designed to help call legs to interact with each other simultaneously by providing 'hooks' for external scripts (such as VoiceXML) to run on legs where IVR is required. An IETF working group called MediaCTRL ("media control") is currently working on unifying these scripting systems to develop an open and widely adopted standard.

3 Relevance of VoiceXML

There are several reasons why the VoiceXML approach is important for voice-based businesses:

3.1 High Penetration of Telephone Networks

The primary reason that voice-based businesses thrive is the widespread usage of phones compared to Internet-enabled computers. In the last decade, mobile cellular connections have grown at a much faster pace than Internet connections. As of 2013, there was a 76.23% saturation for mobile connections as opposed to 30.08% for Internet usage. There are more than 1.5 billion phones (mobile and landline) in use today – much higher than the number of Internet-enabled computers. In addition, telephone networks are still more reliable than data networks. Telephones are simple to operate and use the most natural form of communication, the human voice.

The proliferation of the wireless phone has made access to the telephone network even better. Mobile phones have the advantages of portability, cost-effectiveness and long battery lives compared to desktop or notebook computers. Mobile phones also support multi-tasking – it is far easier to use a mobile phone while walking or driving than it is to use a computer. While most smartphones today have WAP/XHTML browsers, voice is still the preferred mode of interaction. This is because small screens and keypads make typing and traditional browsing difficult on the phone.

3.2 Business Focus on Customer Service

Businesses are increasingly looking for better ways to communicate and conduct business with customers through widely used and easy to operate channels. Companies are investing significantly to build and improve their service infrastructure with the aim of offering convenient, easy-to-use, and cost effective assistance to all potential customers, making it a means of competitive advantage.

3.3 Demanding Customers

At the same time, customers have become more demanding than ever and want anytime, anywhere access to information and services through mobile devices. This has given rise to an opportunity for a telephone-based user interface, rather than a computer-based interface, that is an easy and cost-effective means of delivering information and conducting business transactions.

Voice XML allows businesses a means to leverage the capabilities of the Internet and telephone networks to interact with customers efficiently and effectively.

3.4 Growth of the Internet

Although the diffusion of the Internet is not as significant as the telephone network, it has grown tremendously in the recent past. The Internet has provided a convenient channel for businesses to communicate with their clients. Most businesses have made massive investments in their Internet infrastructure in order to enable fast delivery of information. The Internet is also a great way to route telephone calls at low costs using VoIP (Voice over Internet Protocol) protocol.

3.5 Advances in Speech Technology

Over the past years, speech technology based systems such as artificial speech synthesizers and speech recognition systems have improved drastically due to better algorithms and acoustic models, as well as cheaper and easier access to computing power and electronic storage.

Earlier text-to-speech systems generated speech completely from scratch resulting in poor naturalness of speech and reduced intelligibility. Today, with the use of waveform concatenation techniques, speech synthesis has progressed to a level where lifelike speech can be generated from pre-recorded waveform libraries.

Speech recognition applications can run on home computers with an over the counter microphone and minimal training with the speaker's voice. On the other hand, for a business application using speech recognition, it is easier to 'teach' the system a limited set of grammar rules and vocabulary (commonly encountered words and phrases in that domain), to achieve optimum performance. With technologies such as distributed speech recognition, the accuracy rates have improved even for speech recognition over mobile devices and in noisy environments. In distributed speech recognition, the first level of analysis is done in the device itself and only the output of the initial analysis is given to the application or the server as the case may be.

These advances in speech technology have contributed to more and more applications being developed around voice, and consequently the use of VoiceXML has spread. However, it is important to remember that VoiceXML can be used even in environments that do not use speech technology. For example, VoiceXML applications can take inputs from a keypad and generate audio output by selecting the right response from a set of pre-recorded words and phrases. While speech technology can improve the 'human quotient' in an application, VoiceXML acts as the bridge that connects the advances in web development and deployment to older computer telephony applications.

3.6 Web-Enabled Smart Devices

Today, the Internet extends beyond personal computers and smartphones. Web-enabled devices range from personal organizers with wireless data connections, to 3G enabled digital cameras, to vending machines that can self-reorder when stock goes below preset re-order levels, to smart home electronic devices such as web-based televisions and wall display units.

Voice-based applications and consequently VoiceXML are relevant for web-based devices beyond the telephone. For example, a voice-based 'home remote' can be used to control all home devices like telephones, stereo systems, refrigerators, washing machines and so on using the human voice. These remotes have an on-board voice browser and can even offer content-based value enhancements using VoiceXML applications, such as the ability to query the television programs currently running or the status of the current cycle in the washing machine.

Speech technology is a very natural interface for web-enabled devices. Unlike keyboards/keypads and screens required for visual browsing, microphones and speakers necessary for voice browsers can be built in into smaller devices. In the future, VoiceXML is also likely to find applications in embedded software for devices that have on-board speech recognition facilities.

4 History

The history of VoiceXML can be traced back to 1995 when Dave Ladd, Chris Ramming, Ken Rehor, and Curt Tuckey of AT&T Research brainstormed about various ideas on how the Internet would impact telephony applications. Their discussions on developing a gateway system that could run a voice browser that interprets a voice dialog markup language and delivers web content and services to ordinary telephones led to the AT&T Phone Web project. When Lucent was spun off from AT&T, Ken moved to Lucent and a parallel Phone Web project was started there. Dave and Curt moved to Motorola where they worked on the same ideas.

By early 1999, AT&T and Lucent had their own version of the Phone Markup Language, Motorola had developed VoxML, IBM had SpeechML, and not one of them was compatible with the others. Soon the players in the field realized the need for designing a standard language for the voice web. Although the original group of researchers worked for three different companies, they were still close friends and thus the VoiceXML Forum was born with AT&T, Lucent, and Motorola as the founding members. IBM soon joined the ranks and from March to August 1999, the Forum technologists worked on creating a common new language, VoiceXML 0.9, that had the best of all worlds. It also had some cool additional features such as DTMF support and mixed-initiative dialogue. Once VoiceXML 0.9 was published, there were a lot of comments from the user community. This resulted in an improved and feature rich version 1.0 that came out in March 2000 with client-side scripting, sub dialogs, and properties. This resulted in nearly twenty different implementations being rolled out.

Meanwhile the VoiceXML version 1.0 was submitted to the World Wide Web, which 'accepted' VoiceXML in May 2000 resulting in widespread media coverage. The W3C's Voice Browser Working Group then set about earnestly to create a revised version. W3C focused on a consensus-based approach to ensure that the result was a strong and widely accepted standard. However, the process of achieving the consensus among the participating companies was daunting and time consuming resulting in inordinate delays and the first Working Draft of VoiceXML 2.0 was published to the public only in October 2001. VoiceXML 2.0 became a Candidate Recommendation in January 2003 and the final 'recommendation stage' was reached in March 2004. There were no drastic changes between VoiceXML 1.0 and 2.0, and most changes were fairly conservative. Most of the time was spent on detailing out the expected behaviors and correcting a few errors in the specification. A significant effort was also spent in incorporating new standards for speech recognition syntax and text-to-speech markup language. There were a few extensions such as the new element, but Version 2.0 retained most of the characteristics of Version 1.0.

Historically, VoiceXML platform vendors have implemented the standard in different ways, and added their own proprietary features. However, the VoiceXML 2.0 standard clarified most areas of difference. Today, the VoiceXML Forum, provides a conformance testing process that certifies vendors' implementations as conformant to the standard.

VoiceXML 2.1 incorporated some additional features to VoiceXML 2.0 based on user feedback, and retained backward compatibility with version 2.0. It reached W3C Recommendation status in June 2007.

When VoiceXML 2.1 was released, requests for improvements covered two main categories: extensibility and new functionality. In order to accommodate both, the Voice Browser Working Group first developed the detailed semantic descriptions of VoiceXML that earlier versions did not have. This helped to create the foundation for new functionality as well as to restructure the language syntactically to improve extensibility. Detailed semantic descriptions also helped to improve portability within VoiceXML. The first working draft of VoiceXML 3.0 was released in December 2008.

4.1 VoiceXML's Future

As part of completing the Implementation Report, the W3C conducted several interoperability tests to ensure that the VoiceXML standard was implementable, and that different implementations of VoiceXML could execute the same content in the same way.

Some of the improvements suggested in Version 3.0 included:

- Using the proposed W3C Natural Language Semantics Markup Language to represent recognition results
- Defining a new high-level task-oriented dialog construct parallel to <form> and <menu>
- Defining a new low-level procedural dialog construct parallel to <form> and <menu> to provide additional control to application developers
- Ability to centrally define grammar and audio resources and then reference it by the ‘id’ attributes elsewhere
- Providing standardized audio playback controls (analogous to CD player controls) for changing the speed and volume of the audio, and for moving back and forward in the audio stream
- Providing additional security features such as speaker identification and verification; other features include video capture and replay, and a more powerful prompt queue
- Support for new multimodal markup standards by modularization of VoiceXML; enables XHTML to be used as a container for mode-specific markup (XHTML for visual, VoiceXML for voice, InkXML for ink, etc.), so that the modes’ interaction with each other can be defined using XML Events

5 Practical Applications

Telephony based technologies such as IVRs are not new to businesses such as contact centers, telemarketing companies, and helpline services that rely on phone communication with customers as their primary means of conducting business. However, most traditional telephony technologies are proprietary in nature and are built using high level programming languages, making it difficult to customize. In addition, user input is collected using the DTMF key pads, which limits the applications they can support.

Since VoiceXML has an open architecture, telecom companies, contact centers, and small and medium businesses employ it to provide value-added services using existing technologies and network infrastructure. This allows end users to access and manage stored content and information from any website or database using a conventional telephone or mobile phone. VoiceXML has effectively made the use of special WAP, WEB or touch-tone alternatives redundant.

Today, VoiceXML is widely used in several commercial applications where telephone calls are the primary means of conducting commerce. These are spread across diverse industries such as contact centers, logistics (package tracking), transportation (flight enquiry), web browsing (voice enabled email, audio magazines, etc.) and telecom (voice dialing, directory assistance and enquiry). VoiceXML based applications can be broadly classified into a) productivity applications like sales force automation, unified messaging applications and contact center applications; and b) e-commerce applications or revenue generating applications like financial and banking transactions, billing applications and travel reservations.

Some of the common voice applications that are best suited for VoiceXML are discussed below:

5.1 Information Retrieval or Content Services

Information Retrieval (IR) applications lend themselves easily to VoiceXML. In a typical IR application, the audio output will be selected from a set of pre-recorded information based on a voice input which is again a highly limited vocabulary (consisting of a few commands and limited data inputs). In some cases though, the voice input can be quite complex (for example, driving directions based on a street address). Content service applications using VoiceXML include news, horoscopes, stock quotes, sports scores, movie listings, train and flight timings, and weather reporting service.

On the voice-browsing front, a typical IR application is one where a pre-designed voice newsletter is retrieved through the browser using voice commands. This can also be monetized based on a subscription model or using advertisements. Today, these voice newsletters range from general news to sports, weather information, or even specialized company specific news.

Automated directory assistance applications are another key area of VoiceXML usage. In the case of AT&T's VoiceXML based toll-free directory assistance service, the service was so effective that the rate of automation improved from 8% to 55%, saving AT&T $20 million a year. What is more, along with the increased automation and reduced cost, customer satisfaction levels also improved by more than 33%.

5.2 E-commerce

VoiceXML applications can be used for financial services such as carrying out banking transactions, providing stock quotes, and portfolio management. They can also be used for customer service applications such as package tracking, account status queries, standard support requests, and catalog-based phone orders. In other words, VoiceXML can replace human agents in companies' customer service departments.

5.3 Telephone Services

VoiceXML applications offer telecom companies voice-enabled services such as personal voice dialing, appointment reminders, voicemail management and teleconferencing, all of which can be additional sources of revenue for the company. Since the voice web already has standard web security features, VoiceXML can also be used for developing intranet applications for supply chain management, human resources management, and even to manage business portals.

5.4 Unified Messaging

Unified messaging applications leverage voice with the help of VoiceXML. Some typical applications include reading out e-mail messages over phone, recording of outgoing email messages, voice oriented address book management and synchronization.

VoiceXML can find practical applications in any area where voice services can be used. The possibilities are limitless and range from checking the status of bids at an electronic auction site, authorizing bill payments, scheduling pickups of charitable donations, to ordering a wakeup call at a hotel.

5.5 Advantages of VoiceXML Architecture

There are several compelling reasons why contact centers and businesses must consider a migration to VoiceXML Architecture.

5.5.1 Enhanced Revenue

A VoiceXML Architecture offers additional revenue opportunities for enterprises. Telecom companies and ISPs can use VoiceXML technology to provide innovative and personalized services such as information retrieval services, content-based value-added services, and transaction services to generate additional revenue either from the additional use of a telecom network or through a subscription based model.

5.5.2 Additional Channels for Customer Care

As call centers evolve into contact centers using multiple communication channels such as email, chat and web in addition to the conventional telephone for customer interaction, VoiceXML plays a key role in maintaining cost-effectiveness.

5.5.3 Cost Reduction

VoiceXML based applications enable cost reduction through reduced staffing needs, and a reduction in operating and maintenance costs. In a typical contact center, the largest expense is agents' salary. Effective use of voice-based applications can automate several standard transactions using speech recognition and voice-based IVR systems, which frees agents to handle more complex and revenue generating (upselling and cross-selling) calls. Typically, a call handled by a VoiceXML platform can generate almost 90% cost savings when compared to an agent handled call.

5.5.4 24/7 Self-Service Applications

Apart from cost reduction, voice-based self-service applications have other advantages as well. The key among these is the consistency of responses that an automated agent can provide. Unlike a human agent, who can give different answers to the same query, an automated agent can be pre-programmed to ensure consistency of responses, leading to a perception of higher quality of service among callers. In addition, an automated agent can work 24/7 and still answer queries accurately as they are not affected by fatigue or the need for work breaks.

5.5.5 Leverage Existing IT Investments

In a traditional contact center environment using a legacy IVR system, the applications will be closely tied to the base system using proprietary technologies, making it difficult to add any new functionality or to introduce specific customizations. On the other hand, VoiceXML based IVRs can not only leverage the existing web technologies (EJB, JSP, Java beans) and networks, but also provide flexibility in application design and content delivery. VoiceXML applications are easy to develop and maintain, and the existing IT support staff in a contact center can quickly acquire those skills.

The fact that VoiceXML separates business logic from Voice User Interface has led to the development of a new line of business – Voice Service Provider (VSP). Customers can develop and deploy applications on their premises (they also have the option to be hosted by the VSP), which will be accessed through a phone number assigned by the VSP. As VoiceXML technology catches up, there will be a huge demand for such services, because small to medium sized businesses cannot afford to deploy and maintain their own solutions.

5.5.6 Improved Customer Satisfaction and Customer Retention

Use of VoiceXML applications can lead to increased customer satisfaction resulting in better customer retention. For example, a speech recognition system can help to reduce customer wait time and even reduce the need to navigate through complex DTMF menus. Customers are happy to receive consistent and accurate resolution to their queries in less time than previously possible, with the added bonus of being able to communicate with the system using the most natural form of communication – voice.

6 References

1. http://en.wikipedia.org/wiki/VoiceXML

2. http://www.voicexml.org/voicexml-tutorials/introduction

3. http://www.phonologies.com/pdfs/whyvoicexml.pdf

4. http://cafe.bevocal.com/docs/tutorial/tutorial.pdf