VoiceXML – Surfing on the Internet Using Voice

a report by

Luciano Regruto

Co-founder and Chairman, VoiceXML Italian User Group.

Luciano Regruto is Co-founder and Chairman of the VoiceXML Italian User Group, a non-profit organisation whose aim is the promotion of the application of voice recognition and VoiceXML technology in Italy and Europe. His interest in voice and speech recognition technology spans many years and he is currently a VoiceXML application developer and author of VoiceXML tutorials and technical articles. Mr Regruto is also involved in co-operation with the International Webmasters Association Italy (IWA-Italy) as the organisation's website domain dedicated to VoiceXML and voice technologies.

Voice eXtensible Mark-up Language (VoiceXML) is a new language based on XML (eXtensible Mark-up Language), developed to create websites that are navigable by voice.

The possibility of dialogue with Web applications and services will help those not readily conversant or familiar with computers to use, in a profitable way, a high-tech network such as the Internet.

From a commercial perspective, VoiceXML will open up the market to hundreds of millions of new consumers. In past years different companies have defined various languages for this same purpose (e.g. Motorola VoXML and IBM SpeechML), all of which are valid solutions and the outcome of great research efforts, but with a common drawback – they were 'owner languages'.

After these early attempts, AT&T, Lucent Technologies, Motorola and IBM began to co-operate and to combine their own experiences in the production of a common standard – VoiceXML. This article analyses some aspects of this new and promising technology.

The Interface

One of the most important factors for the success of a service is always its ease of use. Access must be direct and simple. It should also be noted that many devices (e.g. mobile phones, personal digital assistants (PDAs)) offer the possibility of using voice commands. VoiceXML allows the development of services by means of the 'natural' interface of the voice and without the use of such peripherals as mouse, keyboard, monitor or other interfaces.

The innovation surrounding the interaction and navigation model proposed by the Web, or regarding 'pocket technologies' – WAP, for example – is the possibility of using services in an extremely simple and natural way.

The Navigation Model

The scope for which VoiceXML has been defined is clear. It will supply vocal access to Web applications, either by means of a telephone (fixed or mobile) and a PDA or by means of a standard personal computer (PC) equipped with speakers and a microphone.

The navigation model consists of the customer/user being connected to a website via a uniform resource locator (URL) telephone number. When the server responds, it executes a recorded audio application with which it presents the service options. The customer will be able to receive vocal information through synthesis of speech records, execute navigation commands and insert and carry out selections through voice or through the keys of the mobile handset (dual-tone multi-frequency (DTMF)). During the navigation process, the system responds by reproducing audio recorded files or is dynamically generated using text-to-speech (TTS) engines.

VoiceXML Technology

Implementing a system providing vocal access to the Internet leads to new problems regarding 'traditional' methods of access. In the case of access by means of a Web browser, a Web server containing all of the voice application will be sufficient. If access is provided via WAP, it is necessary to add the WAP gateway, a module that acts as a 'bridge' between the world of the mobile Internet (in particular GSM™) and the Internet.

For vocal access, two modules are necessary: one that takes care of the vocal acknowledgment (speech or voice recognition) and the other that is able to realise the translation of the text in voice (TTS). These two modules transform one 'vocal' choice of the user into a comprehensible 'binary' choice for the Web server and one binary answer (or option) of the Web server into a vocal answer (or option) for the user.

VoiceXML allows the developer to focus completely on the application, because it supplies a great deal of high-level structure to interact with vocal devices and their drivers.

Generally, these technical difficulties constitute a great barrier to the development of advanced solutions, but not in the case of VoiceXML.

> *VoiceXML allows the developer to focus completely on the application, because it supplies a great deal of high-level structure to interact with vocal devices and their drivers.*

Applications

The first application of this new technology will probably be a voice-enabled browser. It will be possible for the user to follow links or be connected to new sites through the simple pronunciation of the connections/links, without the use of a mouse and keyboard. It will be possible to build systems without monitors or displays and provide navigation through the telephone (either fixed or mobile).

These functions would be most appreciated by users who are not fully conversant or familiar with computer use and the new vocal systems will be used to create portals that offer every kind of information to anyone that would prefer to listen to rather than read them.

Currently, vocal portals offer such services as lists of theatre performances and cinema screenings and daily news and e-mail (read and write e-mail using voice). A promising class of applications is that of the so-called ‘location-based application’, which are based on the position of the user at the time he/she uses the service.

As an example, the Tellme portal automatically provides the customer with a list of films that are screening at cinemas in the vicinity of the customer’s location.[1] In the immediate future, it will be possible to use voice technology with global positioning system where the system can communicate direction and destination. Access to database business (intranet via voice) and many other services will be available using voice recognition and VoiceXML.

An Example

A good software kit is required to develop navigable applications using the Web via voice (speech-enabled) and to test them on a desktop PC. An example of this is the Cambridge Voice Studio, which is a complete, integrated development environment for building VoiceXML applications that integrate TTS, digitised audio, voice capture and recognition, DTMF key input, telephony and mixed-initiative conversations.[2]

In addition to VoiceXML 1.0 compliance, a library of functions is also provided that puts many powerful messaging capabilities within easy reach of the VoiceXML developer. It is easy to learn and is based on European Computer Manufacturers Association (ECMA) script standards.

The Future of VoiceXML

In the future, many websites will use VoiceXML technology and the impact on the market of this technology (and its innovative navigation model) will be very strong. It will be simple to use, it will increase of the number of users, increase traffic and will develop increasingly sophisticated features.

The textual navigation model, such as WAP, for some applications is not necessarily natural. Vocal technologies use the simpler, direct and natural interface – the voice.

Probably the greatest drawback of the voice-based technologies is the excessive time required to explain all of the information, particularly if menus contain substantial amounts of information. Moreover, the processes of TTS and speech recognition are expensive and, with the increase in the number of connected customers, it will be necessary for them to be very good servers and devices.

Recent investments in technologies such as WAP (also UMTS™) and VoiceXML are great and a good future is predicted for them. In the next two years, more and more websites will become accessible using voice and not only using a mouse and keyboard. ■

[1] *http://www.tellme.com*

[2] *http://www.cambridgeworld.com*