Two Technologies Vie for Recognition in Speech Market

Neal Leavitt

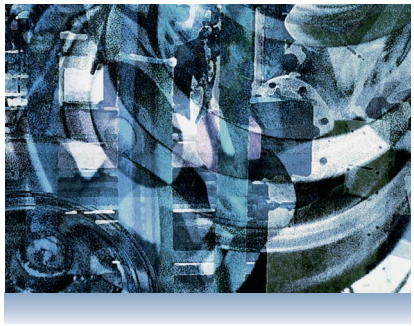
Improvements to voice-recognition algorithms and greater computing power have changed speech technology from an approach with limited uses to an increasingly important part of many applications.

As this process has unfolded, developers have searched for an open standard, rather than proprietary development and runtime environments, that would let them easily and quickly add speech input/output capabilities to applications that function across platforms, said Inderpal Singh Mumick, founder and CEO of Kirusa, a wireless-platform developer.

Two principal approaches have resulted from this search: VoiceXML and SALT (speech application language tags).

Each approach's goal is to let users with minimal training add speech interfaces to applications. Both utilize speech-recognition and speech-synthesis systems to convert voice input to a format that a computer system can understand, to retrieve or produce the requested information, and to convert it back into speech for the user.

VoiceXML is a completely self-contained language for writing speech-based user interfaces. SALT, a newer technology, is a collection of XML tags that developers can embed into an application written in other languages to create a speech or *multimodal* user interface. Multimodal interfaces can be controlled by voice and traditional input methods such as a keyboard or mouse.



Both technologies let developers design speech applications that are portable across multiple platforms.

"VoiceXML and SALT tags hide many of the low-level details often present in proprietary languages," said Jim Larson, principal of Larson Technology Services and chair of the World Wide Web Consortium's (W3C's) Voice Browser Working Group. "This decreases the time required to write and debug code. And because standard languages are more widely used, there will be more developers trained in using them."

The speech technologies would be particularly useful for mobile applications because many handheld devices have tiny keyboards and screens, as well as other limitations that make traditional I/O approaches difficult, noted Rob Kassel, a product manager for SpeechWorks International and a representative of the SALT Forum, an industry consortium that promotes SALT use for multimodal and telephony applications.

Key VoiceXML proponents include Genesys Telecommunications Laboratories, IBM, Motorola, Nuance Communications, Tellme Networks, VoiceGenie Technologies, and Voxeo. Primary SALT supporters include Cisco Systems, Intel, and Microsoft. Some companies support both, including Edify, HeyAnita, Intervoice, Kirusa, and SpeechWorks.

The stakes are huge. Allied Business Intelligence, an industry research firm, projects the global speech-technology market will increase from $677 million in 2002 to $897.8 million this year and to $5.3 billion by 2008, as Figure 1 shows.

DRIVING DEMAND

A number of forces are driving demand for speech technologies and, therefore, VoiceXML and SALT.

Text-to-speech applications let users convert text-based data to speech, noted Genesys product manager Srinivas Penumaka. For example, TTS technology, supported by VoiceXML and SALT, could read e-mail or information from text-based databases to users over the phone.

Individuals could also use phones to access the Internet and send e-mail. "Demand for providing Web data in audio form is increasing every day," noted Latifur Khan, an assistant professor at the University of Texas at Dallas.

The technology also permits voice-based data entry and enables companies to offer user-friendly, Web-based, automated, voice-activated transaction or customer-service applications.

Demand for open technologies such as VoiceXML and SALT has also increased as speech technology has become more accurate. Steve Chirokas, SpeechWorks' director of product marketing, said the technology now offers accuracy rates of more than 95 percent, much better than even a few years ago.

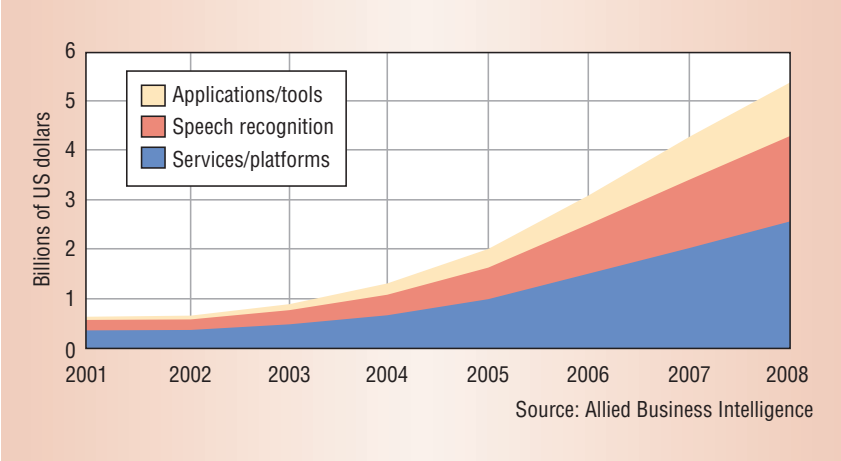


Figure 1. Allied Business Intelligence, an industry research firm, predicts the global speech-technology market will increase steadily during the next five years.

"These accuracy improvements have been especially noticeable in wireless and hands-free environments," said Ken Waln, Edify's chief technology officer.

Therefore, researchers are working to improve speech technologies for mobile environments, in which there are fewer computing resources available and in which network performance can be inconsistent, said Les Wilson, IBM's chief multimodal architect.

VOICEXML

The founders of the VoiceXML Forum—AT&T, IBM, Lucent Technologies, and Motorola—developed the technology in 1999 and donated it to the W3C for formal standardization. The W3C has released VoiceXML 1.0 and 2.0.

VoiceXML was originally designed to support phone menus and other telephony functions, which comprised the primary speech applications at the time. VoiceXML extensions support some related features, such as call-control capabilities.

VoiceXML, designed as a standalone markup language, provides a full set of XML tags, as opposed to SALT, which provides just four primary tags that must be embedded into documents written in other languages such as XHTML (Extensible HTML) or the Wireless Markup Language.

Inside the technology

By using XML tags, which let heterogeneous systems understand the semantics behind a document's content, VoiceXML will let an application work across platforms.

VoiceXML also provides tags for manipulating user interaction through speech recognition, audio output, or telephone touch-tone inputs. Speech functionality is implemented along the execution paths specified by data-control-flow tags that enable different scenarios, such as taking callers to a set of prompts or letting them enter data in a form, said Kirusa's Mumick.

In addition, VoiceXML supports statistical language models, which, according to William Meisel, president of TMA Associates, a speech-technology consultancy, "are used in the most advanced and flexible speech applications."

Statistical language models accept speech input in almost any form and try to predict the most likely meaning. Grammar-based language models, on the other hand, can select only from a predetermined set of replies to user responses to prompts.

Both VoiceXML and SALT also handle speech via grammar tags based on the W3C's Speech Recognition Grammar Specification. SRGS is a language for specifying grammars used by a speech-recognition engine.

Because VoiceXML itself is dialog-based and has only elements needed for telephony applications, it doesn't support multimodal functionality. However, a proposal by IBM, Motorola, and Opera Software for W3C standardization, called XHTML plus Voice (X+V), uses XHTML and five W3C standards, including VoiceXML, for building multimodal applications.

Meanwhile, VoiceXML was designed primarily for use with telephones and supports only a verbal interface and thus must be used with other voice applications to handle graphics.

SALT and X+V, on the other hand, use existing Web tools for handling graphics and add the dimension of integrating a voice interface. The X+V specification uses XHTML to specify graphical content and VoiceXML to support voice capabilities.

Implementing VoiceXML

VoiceXML enables interactive Web access via a phone or voice-driven browser that accesses pages from a Web or VoiceXML server. Most telephones don't have enough processing, memory, or, in the case of cellular handsets, battery resources to support voice browsers, explained Distinguished Member James Ferrans of Motorola Labs' technical staff. Therefore, he said, browser functionality resides on a server that the phone can contact, as Figure 2 shows.

VoiceXML services use a VoiceXML gateway, which consists of an interface to a traditional phone or voice-over-IP system, a VoiceXML interpreter, and speech-recognition and TTS resources. The gateway fetches and interprets VoiceXML from a Web server so that users can work with it over a phone connection.

Meanwhile, users can employ servers and development software to generate VoiceXML Web pages, much as they do for HTML pages.

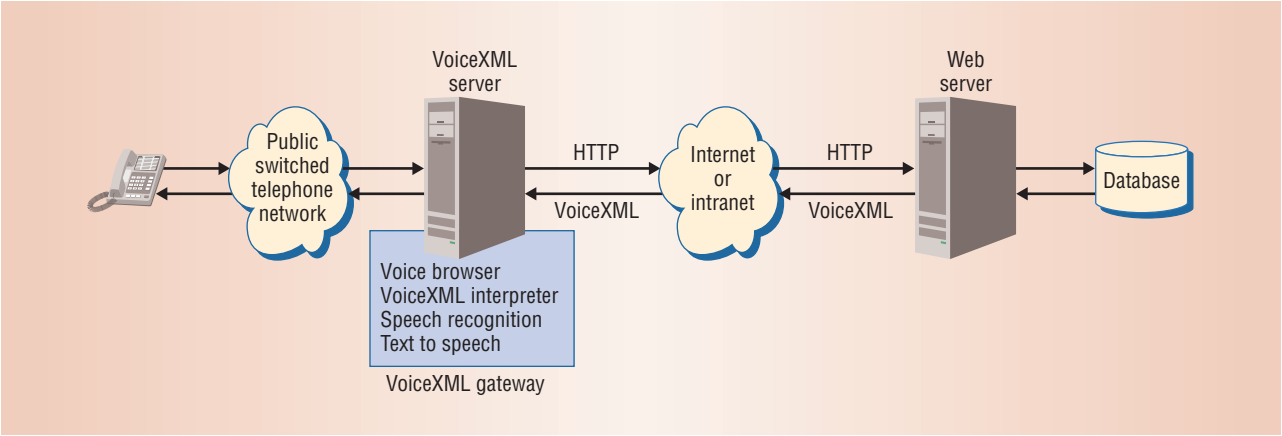


Figure 2. A telephone user can access information from a database via VoiceXML. The caller contacts a VoiceXML browser residing on a server/gateway, which also includes an interface to a phone system, a VoiceXML interpreter, and speech-recognition and text-to-speech resources. The server/gateway fetches and interprets VoiceXML data from a database via a Web server.

Many VoiceXML implementations already exist.

SALT

Companies such as Cisco, Comverse, Intel, Microsoft, and SpeechWorks founded the SALT Forum in October 2001. The consortium released SALT 1.0 in July 2002.

SALT, developed when speech applications were more advanced than when work on VoiceXML began, enables multimodal interactions within an application.

Inside the technology

SALT is a collection of four primary XML tags that developers can insert into a document written in a host language. According to SpeechWorks' Kassel, the tags identify actions to be taken by the browser and provide the data to control these actions.

In addition to basic XML functionality, SALT has a *prompt* tag for speech output, a *listen* tag for speech input, a *DTMF* (dual-tone multifrequency) tag for telephone touch-tone input, and a *smex* tag for call control and messaging with other systems.

Because tags can be inserted in documents written in other languages, SALT can add speech capabilities by leveraging existing Web standards and technologies, rather than creating new ones that users must learn and vendors must add to their products, as is the case with VoiceXML, said Genesys' Penumaka.

And because SALT tags can be inserted into existing Web applications and documents, the technology can provide multimodal capabilities by adding a speech interface to already-present I/O modules such as text, audio, video, and graphics.

SALT also uses the W3C's XML Synthesis Markup Language, which gives systems enhanced capabilities, such as the ability to emphasize certain words in a spoken passage, explained Sastry Isukapalli, Kirusa's director of engineering.

According to Penumaka, the differences between SALT and VoiceXML aren't significant when writing phone-based applications, as both provide telephony basics: speech input, speech output, and call control. However, he said, it is easier to write applications for PCs, PDAs, and other devices with SALT because developers can use SALT tags with other markup languages designed to create applications for different devices. Developers can also leverage existing development tools.

VoiceXML, on the other hand, requires special-purpose browsers and tools capable of handling the technology, said the SALT Forum's Kassel.

Implementing SALT

As is the case with VoiceXML, SALT users can access the Web via a phone that retrieves pages from Web servers. Also like VoiceXML, SALT can work via gateways and interpreters with browser functionality residing on a separate server. However, the relatively small SALT browsers could also reside on smart cellular phones.

Vendors are beginning to release a number of SALT implementations.

LOOKING AHEAD

SALT is newer and thus less mature than VoiceXML. However, SALT proponents say their technology benefits from lessons learned during VoiceXML's development and deployment.

Some industry observers say multimodal technology is too new to make SALT particularly useful in the near future. Thus, SALT's success depends to some extent on whether there will be significant demand for these applications. There already is proven demand for the telephony applications with which VoiceXML is used.

Benjamin Farmer, managing analyst in the Technology Division of market-research firm Datamonitor, predicted that VoiceXML will be more widely adopted than SALT because of its head start and the greater number of vendors who have already integrated

VoiceXML functionality and capability into their platforms.

"Also," Farmer said. "SALT is viewed by many as a Microsoft initiative. Therefore, there's some inherent wariness."

"Right now, VoiceXML is a more viable option but as soon as Microsoft releases its first SALT server, that may change," said Jonathan Eisenzopf, Intervoice's VoiceXML and SALT product manager.

According to Eisenzopf, small- and medium-sized businesses will adopt SALT faster than VoiceXML, whose developers are focusing on large businesses. He predicted SALT will begin capturing significant market share by 2005.

Many observers contend that VoiceXML and SALT will coexist and may even merge.

"At the end of the day, I think that VoiceXML will be the standard for pure voice applications," stated Motorola's Ferrans. "We may see a SALT-like language on Microsoft desktops. Ideally, we'll [eventually] converge on a single standard for multimodal, one that would leverage existing voice standards."

"Ultimately, the W3C will [bring] together the best features of VoiceXML and SALT into XHTML," predicted Igor Jablokov, chair of the VoiceXML Forum's Technology Council. "Speech concepts in SALT that appeal to the speech community will be adopted in a future revision of the VoiceXML standard."

In fact, the W3C has already started working on a dialog markup language, whose working title is VoiceXML 3.0, that could include VoiceXML and SALT features, said Larson. The W3C plans to release a first draft next year.

Regardless, said Kassel, "VoiceXML and SALT will completely displace proprietary systems for deploying speech applications. Both are flexible enough to support many generations of improvements in the underlying [speech] markets." ■

Neal Leavitt *is president of Leavitt Communications, an international marketing communications firm with affiliate offices in Paris, France; Hamburg, Germany; Hong Kong; Bangalore, India; and Sao Paulo, Brazil. He writes frequently on technology issues. Contact him at neal@leavcom.com.*

Editor: Lee Garber, *Computer*, 10662 Los Vaqueros Circle, PO Box 3014, Los Alamitos, CA 90720-1314; l.garber@computer.org