ISSN 1392 - 1215 ELEKTRONIKA IR ELEKTROTECHNIKA. 2004. Nr. 3(52)

T 191 AUKŠTŲJŲ DAŽNIŲ TECHNOLOGIJA, MIKROBANGOS

Voice Based Internet Services

A. Rudžionis, K. Ratkevičius
Speech Research Laboratory, Kaunas University of Technology
Studentų str. 65, LT-51369, Kaunas, Lithuania; phone: +370 37 35419; e-mail: alrud@mmlab.ktu.l, karat@mmlab.ktu.lt
V. Rudžionis
Dept. of Informatics, Kaunas Humanities Faculty of Vilnius Universitety
Muitinės str. 8, LT-44280 Kaunas, Lithuania; phone: +370 37 354191; e-mail: vyrud@mmlab.ktu.lt

Introduction

The beginning of the new millennium will be remembered for many things but in terms of the Internet, one of the most significant developments has been the rapid emergence of the voice based internet services, primarily of voice portals.

Voice portals began life as dedicated systems that provided access to a database of information via a voice channel. In a voice portal, input from the user is through spoken command, which the system can accept thanks to Advanced Speech Recognition (ASR) techniques. Output from the system back to the user is performed by text-to-speech (TTS).

Until recently, voice portals typically did not provide access to the whole Internet worldwide, but rather to the so called 'walled gardens' of content managed by individual service providers, which related to their customers' particular data and services. Now, the Web itself is becoming available through voice interfaces.

This has the potential to transform the nature of Web access, because at a stroke it makes Web content accessible via any telephone. And there are far more telephone users in the world than there are computer users: some 800m wireless phone users, and 1.2bn telephone lines, compared with a 'mere' 300m computer-based Internet users [1].

Voice portals are an important development because they bring the various benefits of voice-based access to the Internet. For one thing, rapid retrieval of information can be much easier via voice, because a user can simply state an item stored in a list or directory, without having to remember a number, scroll through a menu, or listen to each option.

Accessing the Internet from a telephone is nothing new, of course. Many wireless phones now come with microbrowsers that let you to check e-mail or to surf text-only sites. But up to this point, what they offer is a limited, often frustrating form of Web access, where information must be read on tiny screens and responses painstakingly entered on tiny keypads.

Voice portals take advantage of speech recognition and processing to eliminate these constraints. (The technology works with both wireless and desktop phones.) Instead of entering a URL, you dial a toll-free number.

Voice Sites are like Web sites for the telephone. Just like Web sites have interlinked Web pages, Voice Sites have interlinked voice pages. To use a Voice Site, you call from any telephone, listen to the options presented to you, and make your choice by speaking naturally.

Callers' spoken commands trigger events such as browsing audio information, being transferred to a phone number, and sending voice messages!

Voice Sites represent a breakthrough technology for businesses, organizations and the government to provide cost-efficient, high-quality customer service through advanced speech recognition applications. Voice Sites allow callers to get information, connect to people or leave messages from any telephone by simply stating their preferences.

Growth forecasts of the voice portal users and the number of voice sites in North America through the 2001-2005 are presented in the table 1[2].

Table 1. Growth forecasts of the voice portal users and the number of voice sites in North America in 2001-2005

Objects	2001 (fact)	End of 2005 (forecast)
Fixed voice portal users	4 millions	17 millions
Mobile voice portal users	1 million	56 millions
Voice sites	2000	250000

It is evident that the growth forecast is very optimistic, especially for the wireless applications. In the next three-to-five years, it is expected that the number of mobile handsets connected to the internet will exceed that of PCs. Speech recognition technology will follow the growth of users because internet connections and searches will be easier and faster using spoken words rather than tiny keypads [2].

Activities in speech processing area were carried out

in Lithuania for years. Some attempts to implement several test projects were done also in our country. Standardization efforts by famous international companies will effect further development and deployment of speech technologies in Lithuania. We think that it is important to evaluate and try to adapt our demonstrations and projects to current proposals of standards. We'll try to present first attempts to create and implement voice based internet services in Lithuania too.

In recent year two major standard proposals for voice based internet services has been prepared. One of them is called VoiceXML (Voice Extensible Markup Language) and is result of the initiative of IBM, Nuance and several other companies [2]. Another one is called SALT (Speech Application Language Tags) and is mainly result of Microsoft initiative [3]. Despite that both proposals are new and it is still unclear which of them will have better perspectives in the longer term we believe that it is important to begin develop Lithuanian internet services based on standard approaches and Lithuanian speech processing engines.

Two standards of speech-based applications

Implementation of speech technologies grows significantly worldwide in recent years and consequently grows necessity to implement suitable standards for speech technologies. Allied Business Intelligence, industry research firm, projects the global speech-technology market will increase from 677$ million in 2002 to 897.8$ million in 2003 to 5.3$ billion in 2008. This market growth is demand driving and is influenced by a number of forces: text to speech synthesis could read e-mail or information from text-based databases over the phone; permits voice-based data entry and enables companies to offer user-friendly, Web-based, voice-activated transactions, etc. From the technological point of view it is also important that mobile devices have fewer computing resources.

Despite the fact that there exists various standardization institutions, it is well known fact that most often standards are accepted de facto: most popular and best entrenched technical requirements and specifications are taken as a standard. In the speech technologies field there are at least two competing standard proposals at moment. It also could happen that these proposals will converge to new standard.

Several standard proposals have been prepared for speech based applications (voice portals, voice sites and so on). But only two of them are widespread enough and have best chances – VoiceXML and SALT. Both SALT and VoiceXML could have a future, but the bottom line is that it's not just in what language an application is developed; it's about what applications the customers want.

The advent of multi-model applications that are accessible from multiple devices will likely become the next step in computing evolution.

Comparison of VoiceXML and SALT

Here we want to present briefly characteristics of both proposals for standards, better suited application types for each proposal, their perspectives and importance for the development of speech technologies in our country.

Both proposed standards are markup languages. They aim to describe speech interface used in the application. But they operate using very different approaches, mainly due to two reasons: 1) different purposes; 2) different inheritance.

There are substantial differences between SALT and VoiceXML. The SALT specification defines a set of "lightweight" tags as extensions to commonly used Web-based programming languages, such as Java or ECMA Script, that are already well developed, as well as using the W3C standards in common with VoiceXML and some Internet standards from the Internet Engineering Task Force (IETF). VoiceXML is a programming language that does not require other programming languages. Another difference is that VoiceXML implements some functions at a higher level, particularly the "form" function for gathering specific information, avoiding the need to program that function specifically, while SALT operations are at a lower level, giving more control to the programmer, but, in its raw form, requiring more effort if the form function satisfies the application needs.

A fundamental difference is that VoiceXML does not currently deal with multimodal interactions - SALT was designed from the start to handle multimodal extensions.
VoiceXML was primarily developed mainly orienting to telephony application programs (applications which uses phone as one of the ways to access information). This standard has been developed to create specification for programs using interactive voice response (IVR) regime and to allow exploit speech technologies new capabilities that provides Internet. It is a simple high-level markup language that exploits system controlled or mixed initiative voice dialogs through mobile or usual wired phones.

Primarily SALT has been developed using orientation to the applications using speech technologies in the telecommunications and to the wide range of devices: mobile phones, PDA, tablet PC as well as desktop PC. Major factor in SALT technology is the emphasis on multimodality since most of the mobile devices have small displays and simple keyboards so it is possible to combine different modalities to get information but speech input and output is crucial due the limited capabilities of displays and keyboards.

In SALT, there are only four top-level commands: <prompt>, <listen>, <dtmf>, and <smex>. There are additional elements such as <record> and <grammar>, but there are only 10 XML elements total, versus over 30 elements in VoiceXML specification.

SALT proponents argue that VoiceXML is too inflexible and too much of a departure from current Web development tools to attract the "millions" of Web developers to speech.

SALT and VoiceXML have been considered competitors since SALT can be used for telephony applications that are not multimodal. However, there are strong indications that the standards will cooperate more than they will compete in the long run. SALT already uses some specifications proposed by the W3C Voice Browser working group, including Speech Synthesis Markup

Language (SSML), Speech Recognition Grammar Specification (SRGS), and a semantic interpretation language.

It is important to mention that Microsoft has a goal to teach about 6 million programmers worldwide to use SALT technology and encourage in this way to create more speech enabled Internet services. This could be significant factor in competition between VoiceXML and SALT.

Anyway the carriers, contact center managers, system integrators and outsourcers that deploy voice commerce technologies to better serve their clients will benefit from the momentum that VoiceXML and SALT bring to their business. Indeed, they bring the promise of standardized approaches to satisfying the demands of end users and a set of technological approaches that appeal to the broadest base of application developers. In this respect, SALT and VoiceXML are complementary.

Brief introduction to VoiceXML technology. VoiceXML originally has been developed to support phone menus and other telephony functions in applications using voice processing.

VoiceXML generally is w3c based markup language that enables developers to write telephony- oriented applications. Main property – such programs could be realized with such maximum level of simplicity. If you need to develop telephony based application (such as inter-office PBX system based on keyboard/dtmf input), then you will need to invest in the some expensive equipment, high qualification programmers to write and maintain software and place to install telephony equipment. Additionally you will need to upgrade hardware and software, what will allow to achieve new level of functionality. To realize this concept appeared VoiceXML language.

VoiceXML allows average level web designer to create telephony based applications with voice processing elements and simplicity to develop HTML based web pages of average complexity. As VXML is tag-based markup language, its structure in many aspects is similar to HTML, but instead of to be the visual medium VoiceXML is auditory medium, which allows user to browse through "telephony pages" using voice commands instead of pushing buttons on the web page.

Brief introduction to SALT technology. SALT specification was developed by joint efforts of Microsoft and several partner companies together with a groups of volunteers joined by the so-called SALT forum. The fact that SALT forum has been initiated by Microsoft is important: this company dominates on the desktop PC market worldwide and has growing market share in the PDA market.

SALT consists of three main top-level elements:

<listen …> - configures the speech recognizer , executes recognition and handles speech input events

<prompt …> - configures the speech synthesizer and plays out prompts

<dtmf …> configures and controls DTMF in telephony applications

The <listen> and <dtmf> elements may contain <grammar> and <bind> elements, and the <listen> element can also hold <record>. SALT also features ways to configure and manipulate telephony call control through both script and markup.

Each of three top-level SALT elements may be characterized briefly as follows.

The listen element is used for speech input: to specify grammars and a means of dealing with speech recognition results. It is also used for recording spoken input. To do this it contains elements <grammar>, <record> and <bind>. It also contains methods to activate and deactivate grammars, to start and stop recognition. Grammars could be either inline or referenced. Also multiple grammar elements may be used in single listen element. It shows that implemented principles allow flexible manipulation with grammars, particularly for dialog based applications.

The prompt element is used to specify system output. Its content may be simple text, speech output markup, variable values, links to audio files, or any mix of these. Prompt elements are executed declaratively on scriptless or SMIL browsers, or by object methods in script. To ensure inter-operatibility of SALT applications, it is intended that SALT browsers will support at minimum the W3C Speech Synthesis Markup Specification.

The dtmf element is used in telephony applications to specify DTMF grammars and to deal with keypress input and other events. Like <listen> command, its main elements are <grammar> and <bind>, and it holds resources for configuring the DTMF collection process and handling DTMF keystrokes and timeouts. Like <listen> it may be executed declaratively or programaticaly with start and stop commands.

Development of Lithuanian voice based internet services

The main tasks and problems for voice systems implementations. The task of bus schedule presentation through the telephone by voice was selected. The voice based timetable for long distance buses was created: the user collects the known phone number and listens to directions by voice from computer. In the initial stage the timetable of buses from Kaunas to Vilnius was realized. Later this system was expanded to the typical IVR (Interactive Voice Response) system: the user selects the departure town and the arrival town by DTMF means, the IVR systems presents some routes by phone reading prerecorded speech phrases.

The IVR system was reorganized to SLI (Spoken Language Interface) one. First of all speech recognition and text-to-speech programs were transferred to the telecom environment. The original projection based recognition algorithm was used [4,5] in the recognition of words spoken through the telephone. Detecting of word boundaries was carried out according the method presented in [6]. Lithuanian text-to-speech system AISTIS which transcribing and automatic stressing rules are presented in [7,8], was examined in the telecom environment. Some experience of voice operated informative telecom services was reported earlier [9].

Another task was the navigation of internet by voice (access to internet information by voice). Preliminary

attempts to combine three programs were carried out. These three programs were:

- recognition of voice commands;
- reading of text from internet;
- text-to-speech synthesis.

Demo version of program, which reads the text from internet and reports the weather forecast by voice after the appropriate voice command, was prepared. The main obstacle for this program development is the problem of useful text extraction from all text presented in the internet page. The removal of HTML tags is not complicated, because the special functions are prepared for such task. The one way to extract the useful text from all text is to use the HTML comments. Such comments point to the beginning and to the end of articles in the internet editions of newspapers "Lietuvos Rytas", "Lietuvos Žinios", but not in all internet pages are supporting comments. The view of the program, which reads the text from the selected internet page and synthesizes it, is shown in the Figure 1.

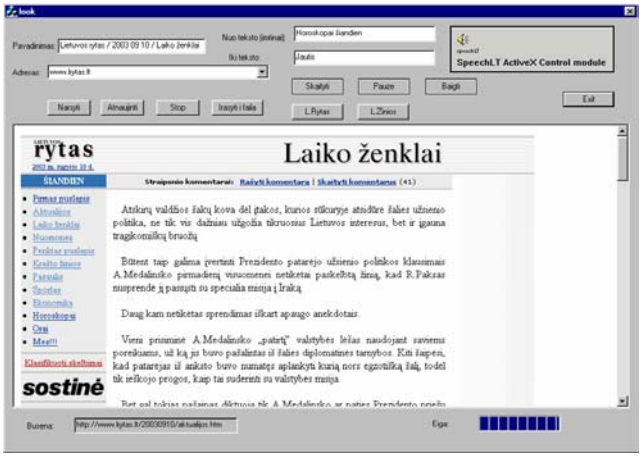

Fig.1. The view of program, which reads the text from internet and synthesizes it

This program enables to select the text that will be synthesized: the beginning and the end of selected text are indicated in the special areas for edit. It is adapted to read by voice two internet pages ("www.lrytas.lt" and "www.lzinios.lt"). After pressing, for example, the button "L.Rytas", Microsoft Internet Explorer opens the internet page www.lrytas.lt, the user selects the desirable article, the program analyzes HTML codes, finds the comments, which points to the beginning ant to the end of desirable article, extracts the text without HTML tags and passes this text to Lithuanian text-to-speech synthesizer AISTIS. This synthesizer is realized as ActiveX component.

Using of SALT or VoiceXML technologies enable to avoid the problem how to extract the useful text from all text presented in the internet page.

Today main manufacturers of programming and web design tools are trying to integrate support of SALT specification to the newest versions of their products. Still dominates two methods of SALT technology implementation:

"Voice Web Studio" programming tool which is used with "Macromedia Dreamweaver MX" programming tool [10];

"Microsoft .NET Speech (SDK) V1.0 BETA 3" which is used with "Microsoft Visual Studio .NET" programming tool [11].

"Voice Web Studio" imports into the "Macromedia Dreamweaver MX" package SALT components. Then with these components you could integrate to web pages output of audio files, text to speech synthesis, to carry out voice based dialogs between user and computer, to record speech to and to associate it with web page, etc. [12].

"Microsoft .NET Speech (SDK) V1.0 BETA 3" also is aims to integrate voice based technologies to made up web pages or pages under development. In this case are used control tools created with ASP.NET (ASP-Active Server Pages) technology that integrates SALT components to Web pages.

In 2003 work to create Lithuanian speaking Internet portal has started: "Microsoft .NET Speech (SDK) V1.0 BETA 3" package was mastered, "Macromedia Dreamweaver MX" tool was acquired, SALT technology was familiarized, demo version of Lithuanian speaking Internet server (http://www.kac.ktu.lt/kstl/test.html) was prepared. To test this web page you need to install Windows'2000 or Windows'XP system and freely distributed plug-ins for Internet browser "Internet Explorer" (so called Speech Add-in) (it is possible to download them from web site [10]). Additionally you need to install interface for Lithuanian text-to-speech synthesizer with SAPI (Speech Application Programming Interface) and to choose Lithuanian as the main language for synthesis.

Conclusions

The growth forecasts of the voice portal users and the number of voice sites are very optimistic, especially for the wireless applications. Speech recognition technology will follow the growth of users because internet connections and searches will be easier and faster using spoken words rather than tiny keypads.

Internet navigation by voice has one essential problem: how to find and extract the useful information from whole text presented on the internet page. This obstacle could be eliminated using SALT or VoiceXML technologies.

Two major standard proposals for voice based internet services has been developed: SALT and VoiceXML. They could be treated as competitors but also they could be regarded as complements. SALT proposal has more emphasis on multimodal applications. Some analysts predict that SALT will be the de-facto standard for integrating speech functionality into desktop, PDA, and Web applications. At the same time VoiceXML will likely remain the dominant standard for developing next generation IVR functionality that integrates with backend Web applications.

We think that it is necessary to begin implement Lithuanian voice based services using standard proposals. Demo version of Lithuanian speaking Internet server with SALT elements (http://www.kac.ktu.lt/kstl/test.html) was prepared.

References

1. **Boothroyd D.** Opening up the Internet through Voice Portals, http://www.hltcentral.org/page-883shtml.
2. **Talking** telecoms // www.telecommagazine.com, February 2001.
3. **SALT** forum, http://www.saltforum.org.
4. **Noreika S. and Rudzionis A.** Phoneme-like model of speech signal // Proc. of the XIIth International Congress of Phonetic Sciences, Aix-En-Provence, France, 1991. -Vol. 4 – P. 490-493.
5. **Rudzionis A., Rudzionis V.** Phonetical segmentation and averaging of the utterances in speech recognition // COST250 "Speaker Recognition in Telephony". Draft Minutes of 3rd Management Commitee Meeting, Lausanne, Switzerland, 1995. - P. 62-65.
6. **Rudzionis A., Rudzionis V.** Noisy speech detection and endpointing // Voice operated telecom services: Do they have a bright future?, Workshop Proceedings, May 11-12, 2000, Ghent, Belgium. – P. 79 – 82.
7. **Kasparaitis P.** Transcribing of the Lithuanian Text Using Formal Rules // Informatica, Vilnius, Lithuania, 1999. – 10(4). – P. 367-376.
8. **Kasparaitis P.** Automatic Stressing of the Lithuanian Text on the Basis of a Dictionary // Informatica, Vilnius, Lithuania, 2000. – 11(1). – P. 19-40.
9. **Rudžionis A., Ratkevičius K., Rudžionis V., Kasparaitis P.** Voice Operated Informative Telecom Services // Electronics and Electrical engineering. – Kaunas, Technologija, 2003. - No. 3(45). - P. 17-22.
10. **Voice** Web Community, http://www.voicewebsolutions.net.
11. Microsoft .NET Speech Technologies, http://www.microsoft.com/speech.
12. **Graham B..** Speak and Listen to the Web using SALT, April 2,2003,http://www.developer.com/voice/article.php/2174471.

Pateikta spaudai 2004 03 08

A. Rudžionis, K. Ratkevičius, V. Rudžionis. Balsinės interneto paslaugos // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2004. – Nr. 3(52). – P. 5-9.

Nagrinėjamos balsinės interneto paslaugos: jų evoliucija pasaulyje bei Lietuvoje. Apžvelgiama balso portalų ir interneto balso svetainių, pasiekiamų per telefoną plėtra. Pristatomi du balsinių interneto paslaugų kūrimo technologijų standartai: SALT ir VoiceXML, nagrinėjami jų privalumai ir trūkumai. Pateikiami Lietuvoje vykdyti darbai kuriant lietuviškas balsines interneto paslaugas. Trumpai apžvelgiamas programinis teksto iš interneto nuskaitymas ir perdavimas sintezės iš teksto programai, akcentuojami šio metodo trūkumai. Pristatoma demonstracinė lietuviškai kalbanti interneto svetainė, kurioje panaudoti SALT elementai. Il. 1, bibl.12 (anglų kalba; santraukos lietuvių, anglų ir rusų k.)

A. Rudžionis, K. Ratkevičius, V. Rudžionis. Voice Based Internet Services // Electronics and Electrical Engineering. – Kaunas: Technologija, 2004. – No. 3(52). – P. 5-9.

Paper deals with the voice based internet services: voice portals, voice sites, presentation of text from internet by voice. Two standards (SALT and VoiceXML) for creation of voice based internet services are analyzed. Advantages and shortcomings of both standards are introduced. Experience of creation of voice based internet services in Lithuania is presented. Program that reads the text from internet and synthesizes it to speech is described. The problem of extracting the useful text from all text is emphasized. Demo version of Lithuanian speaking internet page with SALT elements is presented. Ill. 1, bibl.12 (in English; summaries in Lithuanian, English and Russian).

А. Руджёнис, К. Раткявичюс, В. Руджёнис. Речевые интернетные сервисы // Электроника и электротехника. – Каунас: Технология, 2004. – № 3(52). – С. 5-9.

Статья представляет речевые интернетные сервисы: речевые порталы, речевые интернетные страницы, представление текста из интернета голосом. Анализируются два стандарта (SALT и VoiceXML) для создания речевых интернетных сервисов, представлены их преимущества и недостатки. Представлен наш опыт создания речевых интернетных сервисов в Литве. Описана програма, читающая голосом текст из интернета, акцентируются её недостатки. Представлена демонстрационная говорящая интернетная страница, в которой использованы элементы SALT. Ил. 1, библ. 12 (на английском языке; рефераты на литовском, английском и русском яз.).