ELECTRONICS AND ELECTRICAL ENGINEERING

ISSN 1392 – 1215 *2007. No. 6(78)*

ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY

T 120

SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Investigation of Voice Servers Application for Lithuanian Language

A. Rudžionis, R. Maskeliūnas, K. Ratkevičius

Speech Research Laboratory, Kaunas University of Technology
Studentų st. 65, LT-51369, Kaunas, Lithuania, phone: +370 37 354191; e-mail: alrud@mmlab.ktu.lt

V. Rudžionis

Dept.of Informatics, Kaunas Humanities Faculty of Vilnius University
Muitinės st. 8, LT-44280 Kaunas, Lithuania, phone: +370 37 354191; e-mail: vyrud@mmlab.ktu.lt

Introduction

Speech technologies are being developed for more than 50 years. We can ask: who needs it? The answer is simple – you. Remember how frequently you've heard this boring prompt, while accessing call center: „Currently all operators are unavailable to serve your call, if you want to access X, press 1, if you want to access Y, press 2, if you want to access Z, press 3 and so on". This button pressing information request form (DTMF) isn't very comfortable and accessible to every user. A lot of us forget what to press after a long list of commands, we get angry, unsatisfied with the service, etc. The use of speech technologies lets us to create more natural, intuitive, more reasonable information service for much lower operational cost. The use of speech interface means that information will be available to the user independent from live operator.

We can separate speech technologies into two groups: automatic speech recognition (ASR) and text to speech synthesis (TTS). The accuracy of ASR and naturalness of TTS (or prerecorded prompts) defines speech application and service quality.

Speech or voice servers, such as *Microsoft Speech Server (MSS)* or *IBM WebSphere Voice Server (WVS)*, offer ASR and TTS based speech interface. Special server-managed software controls human – machine dialog. Voice servers integrate together telephony, speech and internet.

Microsoft Speech Server was presented in [1], so *IBM WebSphere Voice Server* is more detailed analyzed in this paper. Examples of speech-enabled Web and telephony applications implemented on *Microsoft Speech Server* were described in [2]. The *VoiceXML* example for *WVS* was developed with *IBM's Websphere Voice Toolkit* in 2006 and illustrates human – machine dialog over telephone line. Application's purpose – to imitate IVR based information system, designed to inform user about selected item by voice. Scenario: user calls to IVR. System asks which shop he is interested in. After user's response, which item category he would like to learn about (milk, meat, fish and bread), system responds with requested information. Program allows both voice and DTMF commands.

Both servers don't support Lithuanian voice recognition and Lithuanian text to speech synthesis engines. Though such engines are prepared (first Lithuanian speech recognition engine "ARVRKRPK Lithuanian Recognizer" was created in 2006), only *Microsoft* or *IBM* can integrate them to voice servers. So far the using of English transcriptions of Lithuanian words is the only solution of voice servers application for Lithuanian language. The results of investigation of voice servers application for Lithuanian language are presented in this paper.

IBM WebSphere Voice Server

Based on open standards, *IBM WebSphere Voice Server (WVS)* is a software middleware product that provides breakthrough technology for quickly developing and deploying conversational solutions. Voice-enabled applications give your customers, employees and suppliers greater access to information and services. IBM support for open standards gives you freedom from proprietary technology, enabling you to manage costs and application deployment schedules for your business.

IBM WebSphere Voice Server provides the Automatic Speech Recognition and Text to Speech resources required to enable speech-based applications. It compiles grammars and uses them to perform Automatic Speech Recognition on a stream of audio data. A grammar is a set of syntax rules that specify what utterances comprise a valid word or phrase. The voice server synthesizes spoken voice from the supplied text and streams the audio back to the *VoiceXML* browser. *WebSphere Voice Server* uses *WebSphere Application Server* as its architectural

foundation. This allows *WebSphere Voice Server* to harness the advanced features of *WebSphere Application Server*, providing extensive administrative and performance benefits to the users (Fig.1).

WebSphere Voice Server V5.1.x now runs as an *Enterprise Application* in *WebSphere Application Server V5.1.1*. This now extends the *WebSphere Application Server* benefits of reliability, scalability, and availability to *WebSphere Voice Server V5.1.x*. In addition, *WebSphere Voice Server V5.1.x* now supports the industry open standard called *Media Resource Control Protocol (MRCP) V1 Draft 4*. At the time of this writing, the latest release of *WebSphere Voice Server* is *V5.1.3* for *Linux* and *Microsoft Windows Server 2003*.

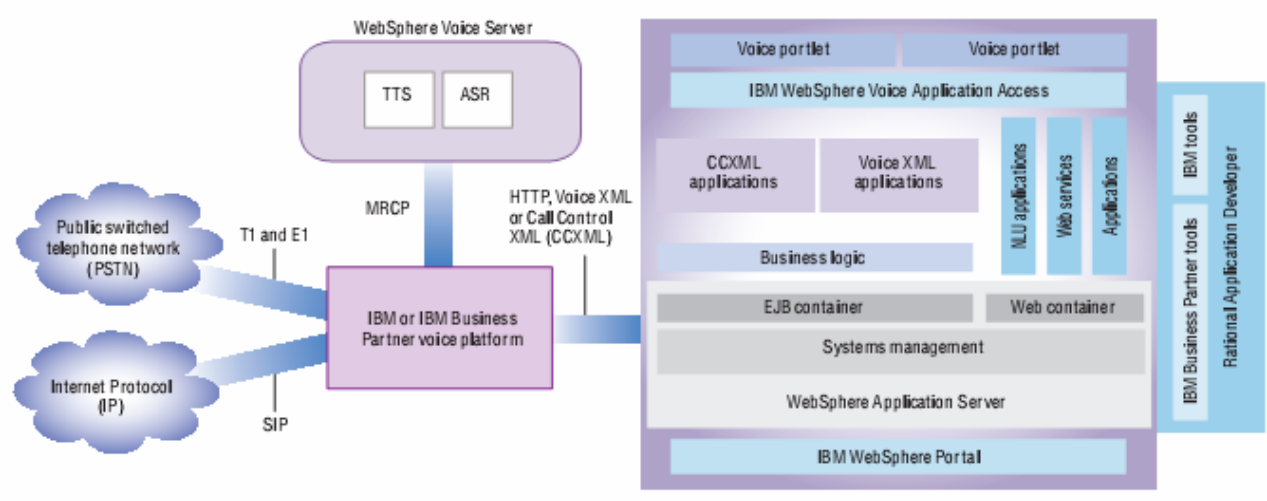

Fig. 1. The basic infrastructure of *IBM WebSphere Voice Server*

Typical "user – voice server" interaction in *WVS* can be described in 10 simple steps.

1. A customer places a call using a telephone to an *IBM WVS* based contact center. They may go directly to a telephony server or it may be routed from the *PSTN* to a *Private Branch Exchange (PBX)* and then to the telephony server.

2. The telephony server *(IVR)* answers the call and uses the *Dialed Number Identification Service (DNIS)* or *Automatic Number Identification (ANI)* information to determine which application the system is to fetch. For example, the application will be a company address book application, which was determined by the phone number that was dialed.

3. Next, the appropriate *VoiceXML* based document is retrieved from the web server.

4. The *IVR* then initiates a *MRCP* synthesizer and *MRCP* recognizer session with *WebSphere Voice Server*. The *IVR* makes a request to the *IBM HTTP Server* to send it a prerecorded audio file.

5. Next, the *IVR* plays the audio prompt to the caller requesting information. At this point, the prerecorded audio may continue playing so the caller can hear additional instructions.

6. The caller may interrupt (use of barge-in function) by saying the name of the person they would like to speak with.

7. *ASR* engine detects the caller interruption, tells the *IVR* to stop playing the message, and sends the speech audio to the recognizer.

8. Recognized data are then passed back to the *VoiceXML* browser.

9. Once the recognition is confirmed, it could lead to the retrieval of person's telephone information and an affirmative response that is played back as audio to the caller.

10. At this point, the *IVR* may have the capability to forward the call to requested person's phone or send data back to the user by other means (i.e. *SMS* message).

Comparing of *Microsoft Speech Server* and *IBM WebShpere Voice Server*

The main difference between two servers is speech markup language. *IBM WebShpere Voice Server* is a *VoiceXML 2.0* (*Voice Extensible Markup Language*) based and *Microsoft Speech Server* is *SALT 1.0* (*Speech Application Language Tags*) based (latest 2007 beta version also supports *VoiceXML). SALT* is a small set of *XML* elements, adding speech-enabled interface to telephony or Web-based applications and bringing them into a multimodal mode. Multimodal environment allows user interaction in several ways: entering data using keyboard, pressing mouse buttons, speaking voice commands, playing back text information using text to speech engine or prerecorded prompts, showing graphical data on screen, etc. *VoiceXML* is aimed at developing telephony Web-based applications, although latest version also adds multimodal support. *VoiceXML* is a much more common standard (supported by more than 500 companies (leaders are ATT Labs, IBM, Lucent, Motorola) vs. 70 SALT supporting companies (leaders are Microsoft, Cisco, Intel, Philips, Speechworks) [3].

Different frameworks and programming tools are used for *SALT* and *VoiceXML* based application development. *Microsoft Speech Server* and *Speech*

Application SDK (SASDK) are based on the *MS .NET Framework*. During installation *SASDK* tools and documentation package is integrated into *Microsoft Visual Studio .NET 2003* environment, that allows developer to code in familiar environment. *IBM VoiceXML* development tools are *SUN Java Framework*-based. *VoiceXML* toolkit is integrated into *IBM Rational Application Development Tools* in similar to *SASDK* fashion. *IBM Voice Toolkit for WebSphere Studio* enables developers to create *VoiceXML* based speech applications using a familiar application development environment, with special tools, such as *VoiceXML* editor, grammar editor, and a pronunciation builder.

Both servers are based on different components. The main components of *Microsoft Speech Server* are: *Speech Engine Services (SES)* and *Telephony Application Services (TAS)*. *SES* uses *Speech Recognition Engine* for spoken input recognition and processing, *Prompt Engine* plays back prerecorded prompts, *Text to Speech Engine* synthesizes audio output from text data. The *IBM WebSphere Voice Server* includes *VoiceXML* voice browser, *Speech Recognition Engine*, *TTS Engine*, telephony and media components.

In order to combine telephony infrastructure and call center functionality, both *IBM* and *Microsoft* voice servers use special telephony hardware and software. *Microsoft Speech Server* relies on third party vendor telephony interface managers (*TIM*). Currently, the most popular two are *Intel Netmerge Call Manager* and *Intervoice TIMs*. *MSS* supports broad selection of telephony and *VoIP SIP* cards. Full list can be found on Microsoft websites. The *IBM WebSphere Voice Server* provides software, telephony, and media component, used to manage the telephony interface. It supports wide range of telephony boards, starting from basic analog telephony boards to complicated digital solutions with a T1/E1 interface.

Both servers have embedded call control tools, such as call transfer, call placement, call reply, etc. If developer wants to use more advanced call control functions, it is possible to use special *Computer Supported Telephony Applications* (*CSTA*). *CSTA* - is a set of *APIs*, that provides an international standard interface between network servers and telephone switches.

Both servers support different languages and operating systems. *Microsoft Speech Server 2004 R2* works only in *Windows 2003 OS* (standard or enterprise editions supported) and supports English (US), French (Canadian) and Spanish (American) dialects. *IBM WebSphere Voice Server* runs on AIX, Windows, and Linux operating systems. In AIX it supports most languages, including Portuguese (Brazilian), French (Canadian), Cantonese, Dutch, French, German, Italian, Japanese, Korean, Chinese (Simplified), Spanish, English (UK) and English (US). In Windows and Linux it supports much less languages.

Investigation of MSS application for Lithuanian language

Both servers don't support third party Lithuanian voice recognition or Lithuanian text to speech synthesis engines and our language support is obviously not included. In most cases synthesis can be fully replaced by prerecorded Lithuanian voice prompts. However recognition is different matter. If Lithuanian words are pronounced differently in English – recognition quality is not acceptable – word is either not recognized or recognized as another word. In order to improve Lithuanian speech recognition quality, it is possible to use English transcriptions of Lithuanian words. This way English speech recognition engine interprets spoken word as English one and sometime the improvement is quite noticeable.

Speech research laboratory conducted the following experiment to prove this. Digits from zero to one (in Lithuanian) were chosen: *nulis, vienas, du, trys (trees), keturi (kehtoori), penki, šeši (sheshi), septyni, aštuoni (ashtuoni), devyni (deveehni)*. Words in which Lithuanian letters are used were transcribed as English ones, i.e. *„šeši" (six)* as *„sheshi", „aštuoni" (eight)* as *„ashtuoni*" and so on. Each word was spoken 100 times and recognition accuracy was measured. Experiment results are displayed in Fig. 2.

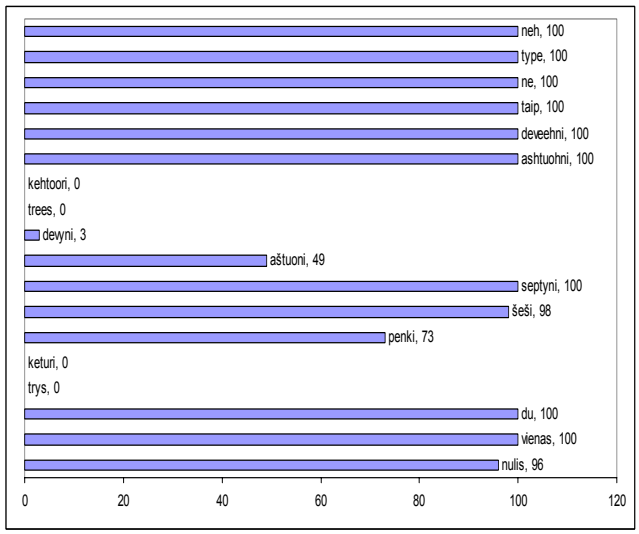

Fig. 2. Accuracy of Lithuanian words recognition by *Microsoft SR* engine through the microphone

Words „*trys*" (*three*) and „*keturi*" (*four*) weren't recognized at all. „*Trys*" in most cases was recognized as „*nulis*" (*zero*), „*keturi*" in most cases was recognized as „*du*" (*two*). Word „*aštuoni*" in most cases (recognition accuracy 49%) was recognized as „*devyni*" (*nine*). Word „*devyni*" was recognized only three times. In order to improve recognition accuracy we transcribed those words to English spelling: words „*trys, keturi, aštuoni, devyni*" were replaced with „*trees, kehtoori, ashtuohni, deveehni*". Recognition accuracy of words „*ashtuoni*" and „*deveehni*" was improved to 100%. But the results for the other two words were not improved: word „*trees*" mostly was recognized as „*nulis*", word „*kehtoori*" – as „*du*".

In order to demonstrate human – machine dialog over telephone, *Microsoft Speech Server* based voice application „Voting" was developed: user calls predefined

number and initiates information request dialog. System (voice application) gathers information, repeats it and asks to confirm data. The following speech commands were used in this example: *numeris vienas, numeris du, numeris trys, kitas, numeris pirmas, numeris antras, numeris trečias, taip.* Experiment was conducted by calling *Microsoft Speech Server* system by telephone and saying each command 100 times. In order to improve speech recognition quality, Lithuanian words were transcribed as English: *noomeris* (*numeris*), *trees* (*trys*), *keetaas* (*kitas*), *peermas* (*pirmas*), *traichas* (*trečias*). Experiment results are presented in Fig. 3.

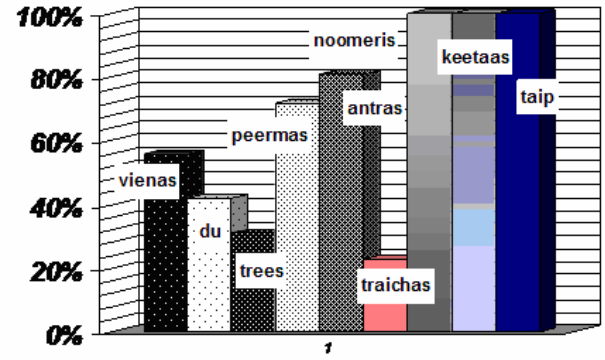

Fig. 3. Accuracy of Lithuanian words recognition by *Microsoft SR* engine through the telephone

Conclusions

Voice servers (*Microsoft Speech Server(MSS)* or *IBM WebSphere Voice Server(WVS))* combines Web technology with speech-processing services and telephony capabilities in a single system. The main difference between two servers is speech markup language: *WVS* is a *VoiceXML* -based and *MSS* is *SALT*- based.

So far, the only acceptable solution to get good Lithuanian recognition accuracy in voice server environment is to transcribe Lithuanian words as English. This way, if transcriptions are properly chosen, we can achieve reasonable results when using English *Microsoft Speech Recognition* engine.

References

1. **Rudžionis A., Ratkevičius K., Rudžionis V.** Speech in Call and Web Centers // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2005. – No. 3(59). – P. 58–63.
2. **Rudžionis A., Ratkevičius K., Maskeliūnas R., Rudžionis V.** Review of Voice Dialogues in Telecommunications // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2006. – No. 5(69). – P. 77–82.
3. **Xiaole Song.** Comparing Microsoft Speech Server 2004 and IBM WebSphere Voice Server V4.2. Retrieved February 22, 2007, from http://www.developer.com/voice/article.php/b3381851.html.

Submitted for publication 2007 03 01

A. Rudžionis, R. Maskeliūnas, K. Ratkevičius, V. Rudžionis. Investigation of Voice Servers Application for Lithuanian Language // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 6(78). – P. 43–46.

Application of voice servers for Lithuanian language is analyzed. Voice servers (*Microsoft Speech Server* or *IBM WebSphere Voice Server*) integrate together telephony, speech and internet. *IBM WebSphere* voice server is analyzed more detailed. Both voice servers are compared according various parameters. Use of English transcriptions of Lithuanian words so far is the only solution of voice servers application for Lithuanian language. The results of investigation of Lithuanian words recognition by *Microsoft* speech recognition engine are presented. Very good accuracy of some Lithuanian words recognition by *Microsoft SR* engine could be achieved if English transcriptions of these words are properly chosen. Examples of prepared speech-based telephony applications could be heard at Speech Research Laboratory. Ill. 3, bibl. 3 (in English; summaries in English, Russian and Lithuanian).

А. Руджёнис, Р. Маскелюнас, К. Раткявичюс, В. Руджёнис. Исследование по применению речевых серверов для литовского языка // Электроника и электротехника. – Каунас: Технология, 2007. – № 6(78). – С. 43–46.

Анализируются речевые серверы (*Microsoft Speech Server* и *IBM WebSphere Voice Server)*, позволяющие объединить интернет, телефонию и речевые технологии. Речевой сервер *IBM WebSphere* описан более детально. Оба сервера сравниваются по разным параметрам. Представлены результаты исследования по применению речевых серверов для литовского языка. Получена очень высокая точность распознавания литовских слов при помощи распознавателя английского языка если удачно подобраны английские транскрипции литовских слов. Подготовленные интернетные-телефонные-голосовые диалоги могут быть прослушены в лаборатории исследования речи. Ил. 3, библ. 3 (на английском языке; рефераты на английском, русском и литовском яз.).

A. Rudžionis, R. Maskeliūnas, K. Ratkevičius, V. Rudžionis. Balso serverių taikymo lietuvių kalbai tyrimas // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2007. – Nr. 6(78). – P. 43–46.

Nagrinėjami *Microsoft* ir *IBM* balso serveriai, jungiantys telefoniją, kalbos technologijas ir internetą. Plačiau nagrinėjamas *IBM WebSphere* balso serveris, pateikiamas *Microsoft ir IBM* balso serverių palyginimas. Lietuviškų žodžių angliškų transkripcijų naudojimas kol kas yra vienintelis balso serverių taikymo lietuvių kalbai būdas. Pristatomi *Microsoft* balso serverio taikymo lietuvių kalbai tyrimo rezultatai. Tinkamai parinkus lietuviškų žodžių angliškas transkripcijas, gautas labai geras kai kurių lietuviškų žodžių atpažinimo tikslumas žodžiams atpažinti naudojant *Microsoft* anglų kalbos atpažiniklį. Paruošti internetiniai balso ir telefono sąsajomis praplėsti tinklalapiai demonstruojami galima Kalbos signalų tyrimo mokslo laboratorijoje. Il. 3, bibl. 3 (anglų kalba; santraukos anglų, rusų ir lietuvių k.).