Speech-Enabled Computers as a Tool for Serbian-Speaking Blind Persons

Vlado D. Delić, Nataša M. Vujnović, and Milan S. Sečujski

***Abstract* — This paper is a review of speech applications and innovative systems in human-machine communication applied for visually impaired persons in Serbian-speaking areas. Three examples are described in more detail: a text-to-speech synthesizer providing independence in computer access to the visually impaired, an audio library for the visually impaired and a speech-enabled web site. These speech applications are predecessors to many commercial user services within areas where south-Slavic languages are spoken, such as voice portals and interactive voice response telephone services.**

***Keywords* — computer access, speech synthesis, speech recognition, visually impaired.**

I. Introduction

Automatic speech recognition (ASR) and text-to-speech synthesis (TTS) enable human-machine speech communication. In this way, humans can talk to devices in their midst such as household appliances, industry machines, cars, toys or remote computers via phone. However, speech technologies are language dependent and cannot be simply imported from abroad as most other technologies. They have to be developed for each language separately.

Owing to the accomplishments of the AlfaNum team at the Faculty of Engineering and the "AlfaNum" company in the field of speech technologies, the Serbian language has joined a rather small group of languages for which speech technologies have achieved a sufficient level of accuracy and robustness to be applied with success. In this way, people living in our country have been given the opportunity to take advantage of new speech technologies in their own language, side by side with people living in more developed countries, speaking languages for which speech technologies have already been developed.

Speech technologies have been of particular use to persons with physical disabilities, helping them to overcome their handicap. For example, a speech-enabled computer is a device which helps the visually impaired to read and write. They get access to written information and literature, have privacy in written communication and possibility to study and work, which in turn raises their self-esteem, confidence and independence. In this way, the visually impaired can get in touch with technology and make use of it more easily, and thus improve the quality of their lives.

Beside personal speech applications based on TTS and/or ASR engines, speech technologies open up a possibility to design and develop innovative systems that have the potential to redefine some aspects of education and social policy related to the visually impaired, such as audio libraries and voice portals.

The rest of the paper is organized as follows. In Section II it is explained how the visually impaired can use computers. Then, in Section III, a short review of speech technologies in Serbian language is given. Innovative systems and speech applications in Serbian language for visually impaired persons are presented in more detail in Section IV. Section V is a Conclusion with a vision of further applications of speech technologies in areas where south-Slavic languages are spoken.

II. Computers and the Visually Impaired

Persons with physical disabilities encounter many obstacles related to computer access. As for the visually impaired, these obstacles can be classified into three main functional groups: barriers related to absence of appropriate input devices, absence of appropriate output devices and access to black print information [1].

Computer input is not usually a problem for the visually impaired since they can get familiar with the classic keyboard layout quickly, and reach the same or greater typing speed than their sighted counterparts. The use of Braille keyboards is not so widespread due to their price and the fact that people who acquire blindness later in life are usually not used to Braille's alphabet. For a visually impaired person not familiar with computers, another speech technology – automatic speech recognition – can be of use. It enables them to control devices and initiate actions by voice. It can, thus, be concluded that

This work was supported in part by the Ministry of Science and Environment Protection of the Republic of Serbia within the Project “Development of speech technologies in Serbian and their application in ‘Telekom Srbija’” (TR 6144A).

Vlado D. Delić is with the Faculty of Engineering, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia & Montenegro; (phone: 381-21-4752999; fax: 381-21-450028; e-mail: vdelic@uns.ns.ac.yu).

Nataša M. Vujnović is with the Faculty of Engineering, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia & Montenegro; (phone: 381-21-4750080; fax: 381-21-450028; e-mail: natasav@uns.ns.ac.yu).

Milan S. Sečujski is with the Faculty of Engineering, University of Novi Sad, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia & Montenegro; (phone: 381-21-4752999; fax: 381-21-450028; e-mail: secujski@uns.ns.ac.yu).

speech technologies are a very efficient way to enhance computer accessibility to the visually impaired.

Computer output can be received via loudspeakers or a headset and in that way a visually impaired person can get the idea what is being displayed on a computer screen. Systems that support this function are called screen readers and they are developed for just a few world languages. A visually impaired person using such a system designed for a foreign language usually has to learn a set of unknown, most frequent keywords without really understanding them in order to get access to a computer. However, for reading any text in user's native language aloud, a speech synthesizer designed for that particular language is needed. With the aid of screen readers and speech synthesizers, a visually impaired person can use all computer functions and applications not directly related to graphic content. A Braille display can also be used for these purposes, with a line of Braille cells giving a tactile representation of the computer's text output.

The solution for the third group of obstacles are scanners enabled to "read" texts in black print and to convert it into electronic form using optical character recognition software (OCR). A text thus converted can be presented to the user using any one of the aforesaid means. Such a device gives the visually impaired user a possibility to access any book or magazine in black print unaided.

Speech technologies have so far been developed for not so many world languages, but Serbian language has recently joined this group, which is important not just for the visually impaired in Serbia and Montenegro, but for all visually impaired who live in former Yugoslav republics where languages similar to the Serbian are spoken, such as in Croatia, Macedonia and Bosnia-Herzegovina. It must be noted that there are only a few languages spoken by as few speakers as it is the case for Serbian, for which speech technologies have already been brought to the level where they can be applied. The synthesizer developed at the Faculty of Engineering in Novi Sad, within the "AlfaNum" project, produces highly intelligible and reasonably natural-sounding speech whose quality surpasses all similar systems developed for Serbian and other closely related languages [2]. The major part of this paper is dedicated to this synthesizer, commonly known as *anReader*, and to the resources built on its basis.

III. A Short Review of Speech Technologies in Serbian Language

During the development of the first high-quality **Text-to-Speech system (TTS)** for Serbian language, this team has encountered many problems linked to bridging the gap between plain text and synthesized speech with all its typical features such as intelligibility and naturalness [4]. There is no explicit information in a plain text regarding phone durations, pitch contours nor energy variations. These factors also depend on the meaning of the sentence, emotions and speaker characteristics [11], which further aggravates the task of attaining high naturalness of synthesized speech. The concatenative approach to speech synthesis has been selected as the most promising. The AlfaNum R&D team has recorded a large speech database and labeled it using visual software tools specially designed for that purpose. By keeping score of every phone in the database and its relevant characteristics, use of phones in less than appropriate contexts was avoided, which further contributed to overall synthesized speech quality. This synthesizer is not diphone-based as almost all other speech synthesizers developed for related languages are. The TTS engine can use larger speech segments from the database, according to both phonetic and prosodic requirements, and select them at runtime in order to produce the most intelligible and natural-sounding utterance for a given plain text [2].

The TTS engine has two main functions: text analysis and synthesis of the speech signal. Text analysis includes text processing such as expanding abbreviations, as well as resolution of morphological and syntactic ambiguities based on a comprehensive accentuation dictionary as well as rule-based syntax analysis. This approach yields highly intelligible and reasonably natural-sounding speech.

The goal of **Automatic Speech Recognition (ASR)** is to recognize spoken words in a speech signal, independently of the speaker, the input device, or the environment. A recognized sequence of words W_{ASR} for a given acoustic observation sequence X and all expected word sequences W are usually estimated using Bayes rule:

$$W_{\mathrm{ASR}} = \arg_W \max P(W|X) = \arg_W \max P(W)\cdot P(X|W)$$

where $P(W)$ is the **language model** estimated using *n*-gram statistics and $P(X|W)$ is the **acoustic model** represented by a Hidden Markov Model (HMM), trained using maximum likelihood estimation. HMM encodes the acoustic realisation of speech and its temporal behaviour, while prior probabilities for word sequences $P(W)$ lead to a choice of the word sequence hypothesis with the maximum posterior probability given the models and observed acoustic data. The best word sequence W_{ASR} is computed using a pattern recogniser based on a standard Viterbi decoder. A conventional approach to front-end signal processing of 30 *ms* frames, every 10 *ms*, results in a feature vector X that captures primarily spectral features of the speech signal estimated as cepstrum and energy, along with their first- and second-order time derivatives. A finite vocabulary defines the set of words (sequences of phone units) and phrases that can be recognised by the speech recogniser. The size of the recognition vocabulary plays a key role in determining the accuracy of a system, typically measured in Word Error Rate (WER), including insertion, deletion, and substitution errors.

R&D for Serbian ASR has been concentrated on four aspects that define the quality of a speech recognition technique [13]:

- *Accuracy* – WER is less than 10% for small and medium-sized vocabulary continuous ASR; it is achieved by developed acoustic modeling trained with 40 hours of

speech databases; good results for large vocabulary continuous ASR in Serbian language are expected when a more complex language model and more comprehensive post-processing are implemented.

- *Robustness* – channel distortions are compensated by CMS (Cepstral Mean Subtraction), background noise spectrum is subtracted and speaker variations are treated by gender separation and speaker adaptation based on VTN (Vocal Tract Normalization).
- *Efficiency* – long work on software code optimization has resulted in fast decoder and small memory footprint. The ASR engine consumes 2% or more of CPU time on a Pentium IV PC, depending on vocabulary size.
- *Operational performance* – The ASR engine gives a useful confidence scoring and implements barge-in capability, improving operational performance. On the other hand, features such as rejection of out-of-vocabulary speech have not yet been enabled.

Owing to the complexity of the problem, a system for isolated word recognition was developed initially. It was later upgraded into a system for connected word recognition [6]. Eventually a system for continuous speech recognition (CASR) was developed, based on recognition of phonemes in particular contexts. Users of this system can define an arbitrary set of words (vocabulary) for each recognition at compiling time [7].

Even state-of-the-art ASR systems cannot be successful enough if they are based on acoustic features only. In order to achieve natural dialogs in speech applications, AlfaNum's ASR has to apply some post-processing such as Spoken Language Understanding (SLU), as well as a lot of experience in both machine learning and design of front-end technology. The goal of SLU is to extract the meaning of recognized speech in order to identify a user's request and fulfill their need. Dialog Manager (DM) evaluates the SLU output in context of the call flow specifications, which results in dynamic generation of the next dialog turn. The DM may apply a range of strategies to control dialog flow according to different application tasks. To provide a successful dialog progress, intelligent speech applications have to handle problematic situations caused by system failures or absence of concise or accurate information in a speech utterance. Post-processing makes it viable to adopt natural language dialog applications without having to achieve perfect recognition accuracy and without dictating what a user should say.

The two speech technologies developed for Serbian language have enabled two-way human-machine communication in Serbian language. This communication can be direct or remote (e.g. via telephone), which introduces the possibility of building various speech-enabled intelligent systems. This is a step of a human-to-machine interface in Serbian-speaking areas from touch-tone prompts toward multimedia and multimodal interface. While both these technologies are still under development [11, 12], they are already used by a number of persons with physical disabilities, especially those with a visual disability.

IV. Resources Based on Speech Technology in Serbian Language for the Visually Impaired

One byproduct of the evolution of communication user services based on advancements in speech and language technologies is enabling people with disabilities to engage in seamless, natural conversation with a new kind of intelligent user services. Speech is the most natural way of communication between humans and the most natural way of asking for information and obtaining them [8]. This is especially true for the visually impaired.

A. anReader – the speech synthesizer for Serbian

As has been explained, the concatenative approach has been used for speech synthesis in Serbian language, as well as online selection of speech segments from a large database, enabling production of highly intelligible speech. However, memory requirements of those two approaches combined are much higher than those of other synthesis methods [4]. Nevertheless, since processing power and memory capacity of a computer become less and less of a problem, this drawback was ignored. Since, on the other hand, most visually impaired computer users in our country possess rather outdated computers, some software optimizations have been carried out, retaining high synthesized speech quality. The final version of the software allows its user to adjust synthesis speed and quality with a single slide control (high synthesis quality corresponding to lower speed and vice versa).

Users can choose between two virtual speakers, one female and one male, obtained artificially from the female one by particular signal processing techniques for speaker conversion. The synthesizer has many features particularly interesting for the visually impaired, such as speaking rate adjustment, average pitch adjustment, spelling out names and obscure words, and reading numbers in context correctly. It is the first synthesizer able to read Cyrillic script correctly, which is especially important in view of the fact that the Serbian is written with Cyrillic script, and much information in Serbian is available in Cyrillic script only [2]. It is compatible with all MS SAPI 5.0 (Microsoft Speech Application Programming Interface) compatible screen-reading software.

B. Audio library for the visually impaired

The audio library for the visually impaired [9] was developed for the pupils of the School for the visually impaired "Veljko Ramadanović" in Zemun, the largest education centre for the visually impaired in Serbia and Montenegro. The lack of literature for the visually impaired used to be a significant problem in their education. Preparation of books in Braille or audio-books is costly and their copies take a lot of space.

The audio library is a system containing a textual database, providing simultaneous access over a local network

and the Internet. The system does not rely on any screen reader and is adjusted to the needs of the visually impaired. Automatic conversion of texts into audio recordings is also supported. The library is a modular system, as shown in Fig. 1. The executive module relies on the communication module and MS SAPI 5.0 enabling access to a virtual speaker. SAPI interface makes controlling audio devices easier, allows multithreading, speaker selection etc. An RPC (Remote Procedure Call) mechanism enables client to server communication. A service on the server side manages incoming RPC calls and executes required functions.

C. A speech-enabled portal for the visually impaired

"Contact" is an interactive web site adjusted to the needs of the visually impaired regarding accessibility and contents. It is supposed to be a meeting point for the blind and a starting point for their future education and employment. The contents of the site are organised in a way easily presentable verbally, and it is accessible via phone as well, which is interesting for those visually impaired who do not have access to a computer [10].

This resource supports both speech technologies, speech recognition and synthesis (ASR&TTS), enabling two-way communication with its user. It is basically an IVR (Interactive Voice Response) system servicing telephone lines together with ASR and TTS servers relying on IP protocol and shared MySQL database, as shown in Fig. 2. In this way standard web page generation by PHP scripts as well as secure connection with IVR system via C API MySQL is achieved.

V. Conclusion

Speech technologies in Serbian are of great importance to 15,000 visually impaired in Serbia and Montenegro, as well as those living in former Yugoslav republics and abroad, speaking or understanding Serbian language.

Resources for persons with disabilities developed so far are mostly intended for the visually impaired. However, a survey is in progress, aimed at better understanding of the needs of the people with disabilities and the benefits that speech technologies in their own language may bring them. On the other side, these human speech applications and services based on ASR&TTS are predecessors to intelligent human-machine communication in broad perspective of commercial applications in the areas where south-Slavic languages are spoken.

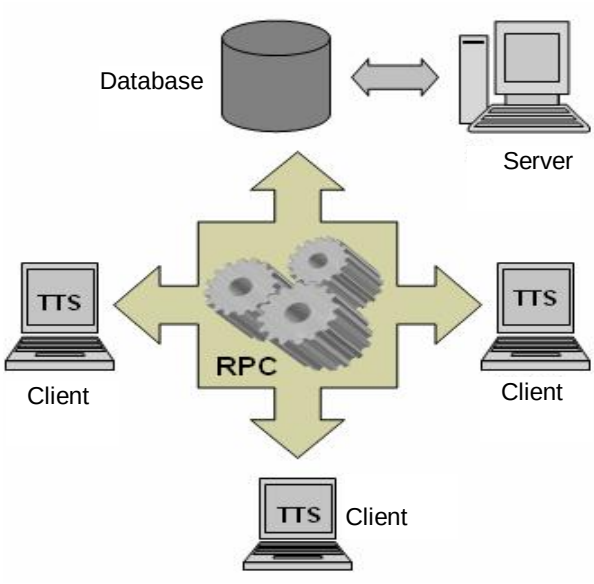

Fig 1. Internal organisation of the library

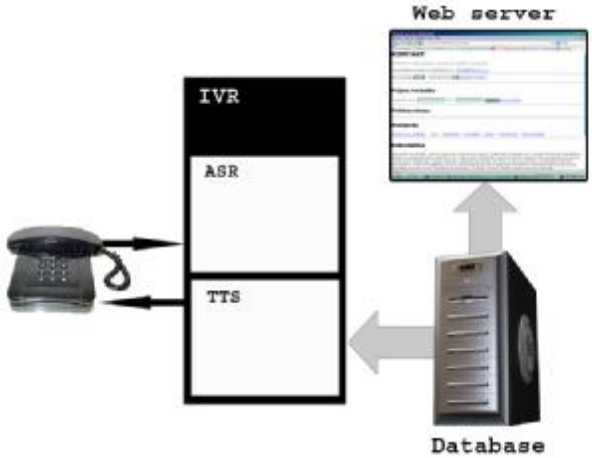

Fig. 2. Internal organisation of the Contact

References

[1] S. Bergstahler, *Working together: People with Disabilities and Computer technology*. DC: University of Washington, 2003.

[2] M. Sečujski, R. Obradović, D. Pekar, Lj. Jovanov, and V. Delić, "AlfaNum System for Speech Synthesis in Serbian Language," in *Proc. 5th Conf. Text, Speech and Dialogue*, Brno, Czech Republic, 2002, pp. 237-244.

[3] F. Jelinek, *Statistical Methods for Speech Recognition.* Cambridge, MA: The Massachusetts Institute of Technology, 1998.

[4] S. Lemmetty, "Review of Speech Synthesis Technology," M.S. thesis, Dept. Elect. & Comm. Eng., Helsinki, Finnland, 1999.

[5] J. C. Junqua, and J. P. Haton, *Robustness in Automatic Speech Recognition*. Dordrecht, the Netherlands: Kluwer Academic Publishers, 1995, pp. 125-147.

[6] D.Pekar, R.Obradović, V.Delić, "Connected Words Recognition," in *Proc. 9th Conf. TELFOR*, Belgrade, S&M, 2001, pp. 455-458.

[7] D. Pekar, R. Obradović, and V. Delić, "Programming package AlfaNumCASR – a system for continuous speech recognition," in *Proc. 3rd Conf. DOGS*, Bečej, Serbia, 2002, pp. 49-56.

[8] H. Levitt, "Processing of Speech Signal for Physical and Sensory Disabilities," in *Proc. N.A.Sci. USA*, 1995, pp. 9999-10006.

[9] D. Mišković, N. Vujnović, M. Sečujski, and V. Delić, "Audio library for the visually impaired as an application of text-to-speech synthesis," in *Proc. 49th ETRAN*, Budva, S&M, 2005, pp.400-402.

[10] R. Ronto, D. Pekar, and N. Đurić, "Realization of a speech-enabled telephone portal based on TTS and ASR," in *Proc. 49th ETRAN*, Budva, S&M, 2005, Vol. II, pp. 392-395.

[11] M. Sečujski, "Obtaining Prosodic Information from Text in Serbian Language," in *Proc. Conf. EUROCON*, Belgrade, Serbia, 2005, to be published.

[12] N. Jakovljević, and D. Pekar, "Description of Training Procedure for AlfaNum Continuous Speech Recognition System," in *Proc. Conf. EUROCON*, Belgrade, Serbia, 2005, to be published.

[13] "When Machines Talk – Speech Technology in Human-Machine Communication," *Special Section in IEEE Signal Processing Magazine*, Vol. 22, No. 5, September 2005, pp. 16-126.