7

Applications of Speech Technologies in Western Balkan Countries

Darko Pekar[1], Dragiša Mišković[1,2], Dragan Knežević[1,2], Nataša Vujnović Sedlar[1,2], Milan Sečujski[2] and Vlado Delić[2]
[1]AlfaNum – Speech Technologies, Novi Sad,
[2]Faculty of Technical Sciences, University of Novi Sad,
Serbia

1. Introduction

The chapter will present the first applications of speech technologies in the countries of Western Balkans, launched by the Serbian company AlfaNum. The speech technologies for Serbian and kindred South Slavic languages are developed in cooperation with the University of Novi Sad, Serbia. Most of these applications are rather innovative in Western Balkans and they will serve as a base for complex systems which will enable 20 millions of inhabitants of this part of Europe to talk to machines in their midst in their native languages, equally to their counterparts who live in more developed countries in the region.

Firstly, the importance of research and development of speech technologies will be stressed, particularly in view of their language dependence and, on the other hand, the possibility of their wide application. The central part of the chapter will focus on the results of the research and development of the first applications of automatic speech recognition (ASR) and text-to-speech synthesis (TTS) across Western Balkans – some of them are a novelty in a much wider region as well. The paper will be concluded by the directions of future research and development of new applications of speech technologies in the Western Balkan region and worldwide.

1.1 Relevance of the research and development of speech technologies

When communicating with others, people predominantly use the senses of sight and hearing – they speak, listen and watch. On the other hand, when communicating with machines (computers, telephones, robots, cars etc.), they mostly use the senses of sight and touch – they look at monitors and touch keyboards, mice or touch screen displays. It is worth noting that humans rarely address machines using speech and that machines rarely use speech to respond, although spoken communication is the most natural form of communication among humans. Apart from a number of fundamental problems related to ASR and TTS applications, addressed in more detail in (Delić et al., 2010), another possible reason for this is the fact that speech technologies are highly language dependent, and that a number of necessary resources and techniques have to be developed for each language separately. The most has been done for languages spoken by relatively large communities, but quality solutions for languages with smaller communities are beginning to emerge.

Source: Advances in Speech Recognition, Book edited by: Noam R. Shabtai,
ISBN 978-953-307-097-1, pp. 164, September 2010, Sciyo, Croatia, downloaded from SCIYO.COM

ASR is language dependent to a great extent, and TTS to an even greater. There are several aspects of this language dependence: (1) A database of at least several hours of recorded speech in a specific language must exist, in order to be able to produce high-quality synthesised speech, regardless of the method used. Speech databases for ASR training, which can be of much greater size, are also language dependent. (2) Morpho-syntactic analysis and syntactic-prosodic parsing of the input text have to be carried out, and both tasks are highly language dependent. (3) Based on the previous analysis of input text, appropriate prosody features (phone duration, f_0 contour and energy) have to be generated.

1.2 Integration of ASR and TTS engines into applications

AlfaNum TTS and ASR engines can be used through a number of interfaces, all of them built upon basic TTS and ASR libraries written in C++. The main reason for their design was to make engine integration into existing products as simple and as fast as possible.

- **C++ library (proprietary interface)** – TTS and ASR C++ libraries are at the base of all supported interfaces.
- **Microsoft SAPI** – The Speech Application Programming Interface or SAPI is an API developed by Microsoft to allow the use of speech recognition and speech synthesis within Windows applications. Both SAPI 4 and SAPI 5.x interfaces are implemented. In general, all versions of the API have been designed such that a software developer can write an application to perform speech recognition and synthesis by using a standard set of interfaces, accessible from a variety of programming languages. Speech engines are built as standard COM objects by implementing the required COM interfaces.
- **Media Resource Control Protocol** – The Media Resource Control Protocol (MRCP) is a protocol proposed by the Internet Engineering Task Force (IETF), which has the goal of standardising computer dialogues between the ASR and TTS with interactive voice response (IVR). Clients send MRCP messages to the server over a network usually by means of another protocol, such as Real Time Streaming Protocol (version 1) or Session Initiation Protocol (version 2). AlfaNum servers comply with version 2 of MRCP protocol.
- **AlfaNum IP server/client (proprietary interface)** – This interface is based on a proprietary protocol which includes additional functionality not found in any of the industry standard protocols. This protocol is designed to make the system more robust and provides faster content delivery. For this purpose speech engines (C++ libraries) are built into the AlfaNum IP servers. Along with the server, TTS and ASR client libraries are created to enable developers the use of AlfaNum IP server functionality from within different programming languages. Client libraries are developed for C++, C#, Visual Basic and PHP programming languages.

  The basis of AlfaNum IP server is a multi-threading protocol which accepts connections from client applications and is based on TCP/IP. The functioning of the server can be explained through two types of sockets that are created. The first one is the *listening socket*, which collects connection demands generated by clients. After the demands are received, a *service socket* is opened for each client, through which further communication is carried out, as shown in Fig. 1. Such a mechanism enables handling a large number of users and simple addition of new routines.

  Besides remote access, the client library that encompasses the communication between the applications and the server also enables the use of multiple ASR/TTS servers (located at different computers) in case of need for a large numbers of simultaneous requests for speech recognition.

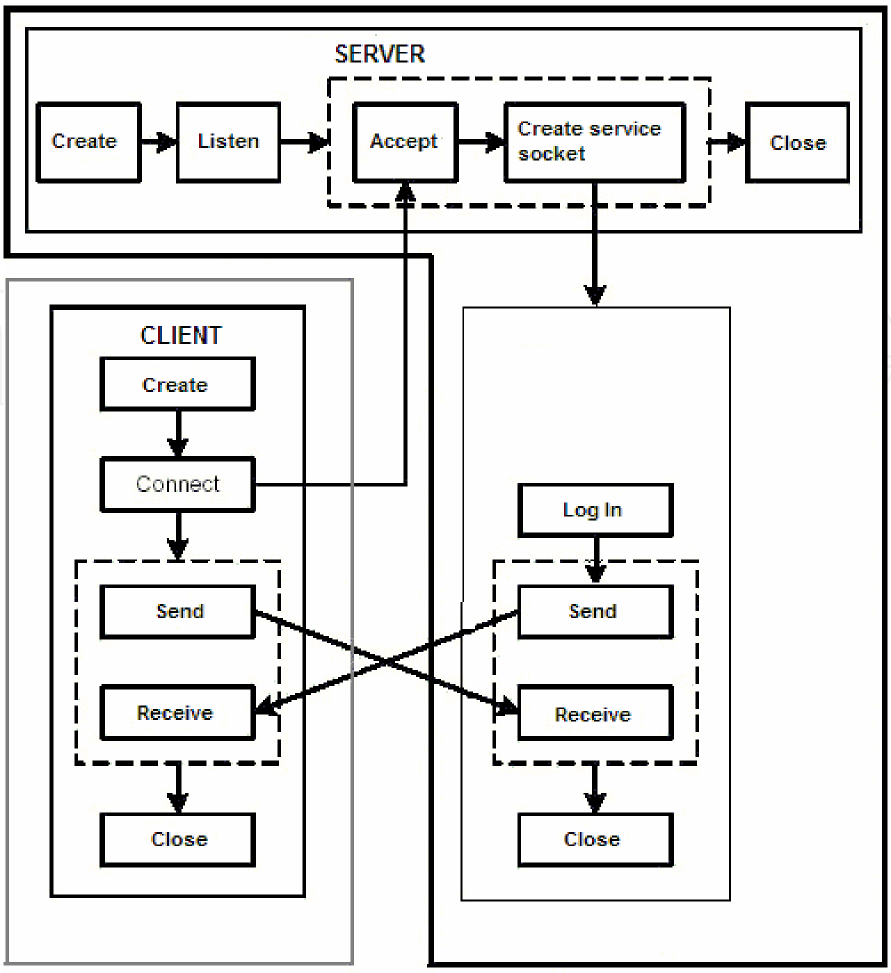


Fig. 1. Communication protocol of the AlfaNum IP server

Within speech-enabled applications ASR components are commonly coupled with TTS components, thus completing the cycle of human-machine speech communication (Delgado & Araki, 2005). Various areas of interaction have so far been covered, and today there are active systems offering e.g. information related to bus schedules and stock market, as well as any information that can commonly be found in electronic newspapers.

2. ASR applications

The public telephone network is currently the most promising ground for application of speech technologies (Nöth, 2004). The first applications of ASR in Western Balkans have been launched at the public telephone network, with support by intelligent network functionalities. Some of the innovative applications of ASR in Serbia will be described in the following sections.

2.1 Interactive Voice Response systems

As has been mentioned before, ASR and TTS IP servers have found their first applications within AlfaNum Interactive Voice Response (IVR) systems. An IVR system is a computerised system allowing a user (a telephone caller) to choose among various options offered in voice menus. The first IVR systems played pre-recorded voice prompts to which the user would press a number on a telephone keypad to select the option. Integration of ASR and TTS components significantly improve this interaction and complete the cycle of human-machine communication. The foundation of all IP server based IVR systems developed is the simultaneous functionality of ASR and TTS servers and their communication with a required number of IVR processes (one per telephone line) via IP protocol.

Intel/Dialogic Telephony Cards provide a connection to the public telephone network. Through it, the calls are routed to any of the free channels managed by the IVR controller. At the same time, the controller provides a link to the database and ASR and TTS servers. The database represents an information source from which data is presented to the user by TTS in the form of synthesised speech, based on user requests that the system acquires via ASR. The ASR and TTS servers can reside on remote computers (dedicated if required) and can communicate with a number of different IVR applications.

Specific properties of such systems, from the point of view of ASR, manifest themselves in the need for activation of different recognisers according to options offered to the user in a given moment. Furthermore, the systems handle information that changes dynamically, and for that reason the grammars used for recognition often have to be generated dynamically according to the database contents. The basic organisation of an AlfaNum IVR system is shown in Fig. 2.

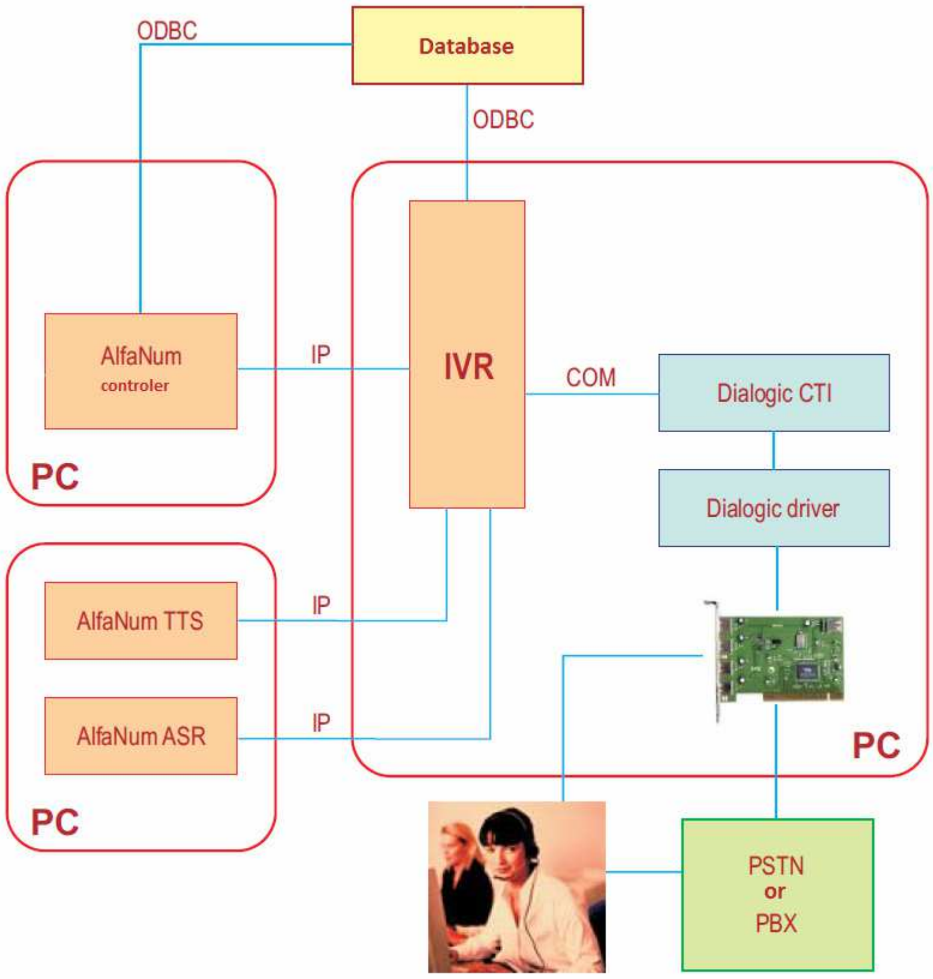


Fig. 2. The basic organisation of an AlfaNum IVR system

2.2 Advertising monitor

Besides application in the field of telephony, the AlfaNum ASR engine has also been implemented in systems that perform searches through a large amount of audio material. One of such systems, Advertising Monitor (Pekar et al., 2007), is an application that locates audio content such as jingles and commercials in audio archives. The system comprises a number of FM and TV tuners receiving signals from various radio and TV stations, and a search service that can be distributed to multiple computers. Specific properties of the material being searched allow the use of a simplified recognition process based on LPC coefficients. Unlike classical speech recognition, the input to this system is subject to different types of variations, which is reflected in the sound signal processing algorithm. However, there are also some alleviating circumstances for development of such a system:

- A complete absence of temporal variability between the reference recording and the test recording.
- A drastical reduction in acoustic variability in comparison with classical speech recognition. In this case, acoustic variability is the consequence of changes in channel properties (spectral changes and noise), which evolve slowly over time and the effect of which can be reduced to a sufficient degree using first time derivatives of acoustic features.

For that reason, processing of the sound signal amounts to calculation of its dynamic features, namely, first and second time derivatives of LPC coefficients. In this way there is no front-end processing in recognition and a significant portion of the processor time can be saved. Furthermore, because of the aforementioned absence of temporal variability, simple one-on-one comparison of the reference recording and the incoming signal can be applied instead of DTW or another, more complex time-alignment algorithm. Blocks containing the reference recording simply slide along the received signal and block-by-block comparison is carried out through calculation of the average distance between blocks of the reference recording and corresponding points in the received signal. When a very significant drop in the distance is observed, it can be concluded that the reference recording was located in the received signal, as shown in Fig. 3.

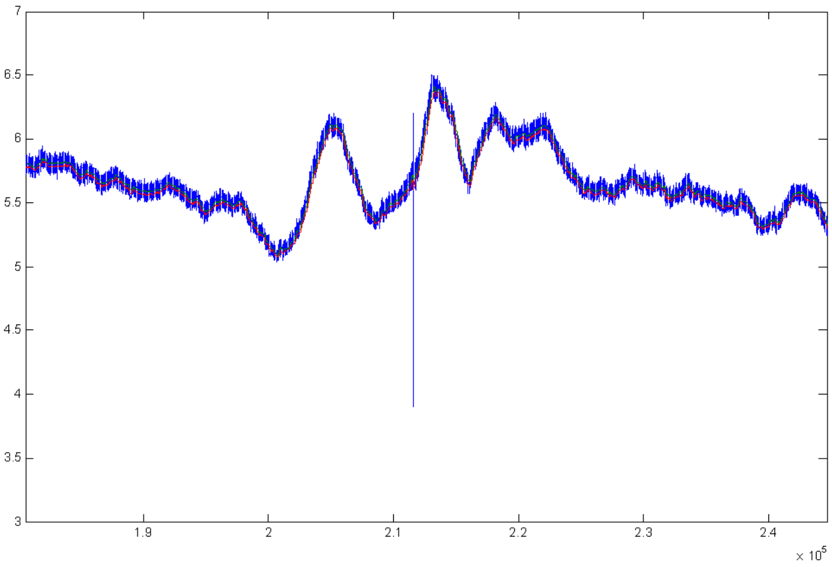


Fig. 3. Distribution of the average distance between the reference recording and the received signal

The system can fail only in cases where there is a significant mismatch between the original reference recording and the occurrence of the same audio material in the received signal (e.g. a truncated or overdubbed commercial). However, such problems can be efficiently handled by appropriate postprocessing techniques.

2.3 Word Spotter

AlfaNum Word Spotter (Mišković et al., 2007) is an application that locates key words and phrases, given as text, in arbitrary audio material. The system relies on the ASR speech recognition engine, and the nature of the system indicates its areas of use and features expected by its users (various security agencies, media monitoring agencies etc.).

The functioning of the AlfaNum Word Spotter is based on the phoneme-based, speaker independent speech recognition system, AlfaNum ASR. Particular features of the application are related to the way trellises for given key words and phrases are built. If the standard approach to speech recognition is taken, with adaptation of syntax so as to allow for multiple pronunciations of a single word, word spotter produces the recognition result as a sequence of arbitrary number of silence models, noise ("garbage") models, key word models and wildcard models (universal models covering parts of an utterance that do not contain key words). This is the consequence of the way the trellis for each key word is generated, as shown in Fig. 4 (word 1,... word *n* represent transcriptions of basic and inflected forms of a word).

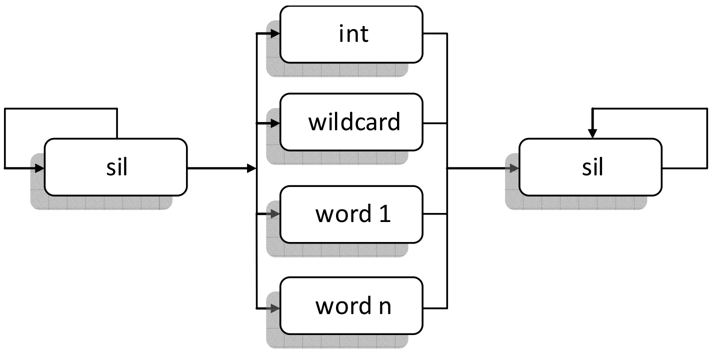


Fig. 4. Transition diagram of the AlfaNum Word Spotter

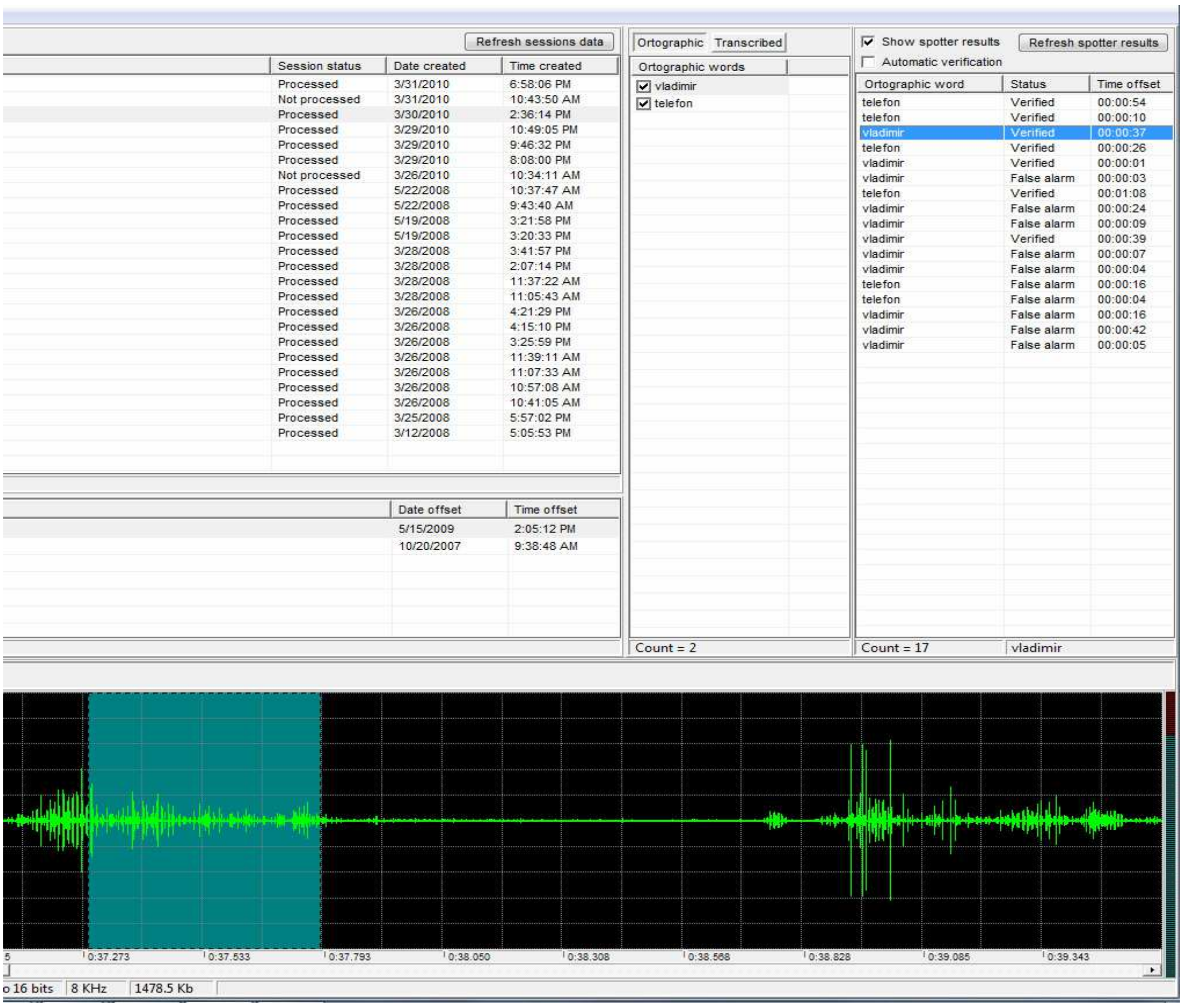
As can be seen in the figure, besides key words and phrases, trellis also contains non-speech states as well as states modelling various types of noise (INT). Having such a structure in mind, buildup of a trellis of the word spotter is based on the following rules: each sequence must begin with silence (non-speech state of an unlimited duration), from silence it is possible to traverse only into an initial state of a key word or phrase or into a noise or wildcard model, a word has linear structure and limited duration, from the final state of a key word or phrase it is possible to traverse only into the state of silence, which is at the end of every sequence.

Unlike application of ASR in interactive voice response systems or call centres, where a user can be asked to repeat the utterance more clearly in case of unreliable recognition, in case of a word spotter error rate has to be reduced to a minimum possible level.

Two types of errors can be identified. The first type of errors is related to key words that existed in the recording, but were not recognised by the system (false negatives), and such errors are critical from the point of view of system reliability. The second type of errors is related to the words that did not exist in the recording, but were nevertheless "recognised" by the system (false positives). Eliminating as much false positives as possible without creating significant false-negative results is a very demanding task, directly related to specific properties of ASR algorithms. Some of the false positives can be eliminated by subsequent comparison of the durations of particular phonetic segments of the recognised word or phrase to the expected ones (Mišković at al., 2007). The graphical user interface of the application has been designed so as to enable the user to eliminate a significant number of false positives, since the recognition results (recognition locations) are displayed in order of decreasing reliability. The user can thus decide to stop manually checking the results when a sufficiently high rate of false positives is reached. A portion of the graphical user interface related to recognition result verification is shown in Fig. 5. The figure shows the situation after some of the results have been checked, and the distribution of accurate recognitions vs. false positives can be observed.

The next step in the development of this tool would be in the direction of its integration with a system for recording telephone conversations.

Besides the applications described in this section, developed on the basis of the existing system for speech recognition, there is a number of areas in which the application of this system is yet to be expected. Ongoing development of a continuous speech recognition



Session status	Date created	Time created
Processed	3/31/2010	6:58:06 PM
Not processed	3/31/2010	10:43:50 AM
Processed	3/30/2010	2:36:14 PM
Processed	3/29/2010	10:49:05 PM
Processed	3/29/2010	9:46:32 PM
Processed	3/29/2010	8:08:00 PM
Not processed	3/26/2010	10:34:11 AM
Processed	5/22/2008	10:37:47 AM
Processed	5/22/2008	9:43:40 AM
Processed	5/19/2008	3:21:58 PM
Processed	5/19/2008	3:20:33 PM
Processed	3/28/2008	3:41:57 PM
Processed	3/28/2008	2:07:14 PM
Processed	3/28/2008	11:37:22 AM
Processed	3/28/2008	11:05:43 AM
Processed	3/26/2008	4:21:29 PM
Processed	3/26/2008	4:15:10 PM
Processed	3/26/2008	3:25:59 PM
Processed	3/26/2008	11:39:11 AM
Processed	3/26/2008	11:07:33 AM
Processed	3/26/2008	10:57:08 AM
Processed	3/26/2008	10:41:05 AM
Processed	3/25/2008	5:57:02 PM
Processed	3/12/2008	5:05:53 PM

Date offset	Time offset
5/15/2009	2:05:12 PM
10/20/2007	9:38:48 AM

Ortographic words
☑ vladimir
☑ telefon

Count = 2

Ortographic word	Status	Time offset
telefon	Verified	00:00:54
telefon	Verified	00:00:10
vladimir	Verified	00:00:37
telefon	Verified	00:00:26
vladimir	Verified	00:00:01
vladimir	False alarm	00:00:03
telefon	Verified	00:01:08
vladimir	False alarm	00:00:24
vladimir	False alarm	00:00:09
vladimir	Verified	00:00:39
vladimir	False alarm	00:00:07
vladimir	False alarm	00:00:04
telefon	False alarm	00:00:16
telefon	False alarm	00:00:04
vladimir	False alarm	00:00:16
vladimir	False alarm	00:00:42
vladimir	False alarm	00:00:05

Count = 17 vladimir

Fig. 5. A segment of the graphical user interface of the AlfaNum Word Spotter

system for large vocabularies expands the area of application of this technology. The first applications are expected to be dictation systems, spoken dialogue systems and applications for automatic subtitling of TV shows. Applications related to speaker identification and verification will also be developed.

3. TTS applications

This section will describe the first applications of text-to-speech synthesis in Serbian and kindred South Slavic languages. The AlfaNum TTS engine supports a number of interfaces in order to facilitate its integration into useful applications. Some of these interfaces are standard, such as the C++ library, Microsoft SAPI5, Microsoft SAPI4, and MRCP, however, communication with other components in sophisticated systems is also possible via a custom designed AlfaNum IP protocol (implemented through libraries in C++, Visual Basic, C# and PHP).

3.1 AnReader

The first widely applied TTS-based system in WBCs is *anReader* (Delić et al., 2005; Sečujski et al., 2007), used by almost one thousand visually impaired persons in Serbia, Bosnia and Herzegovina, Montenegro, Croatia and FYR Macedonia. Before the appearance of *anReader*, the most widely used system was *WinTalkerVoice*, originally built for Czech language. It produced synthesised speech of poor quality in Serbian and Croatian, and has therefore never been used for any other purpose than as aid for the visually impaired. *AnReader*, on the other hand, was initially developed for Serbian, and later for Croatian and Macedonian as well. The basic concepts are the same, but the morphological dictionary (especially the information related to accentuation) and the rules for morpho-syntactic analysis had to be modified. The Croatian *anReader* required that a new speech database in Croatian be recorded and processed, while the Macedonian version currently uses the Serbian speech database, with slight impairment of speech quality as a result. High-level speech synthesis of Serbian and Croatian is performed using expert POS taggers, while for Macedonian full POS-tagging is never performed since it is not necessary for reasonably natural pronunciation of Macedonian (owing to the simplicity of accentuation in Macedonian in comparison to the other two languages).

It should be kept in mind that, for a visually impaired user to be able to use a computer unaided, besides a synthesiser such as *anReader*, he/she also needs a screen-reader, an application attempting to identify and interpret what is being displayed on the screen, as well as to communicate information on menus, controls, and other visual constructs. Owing to a number of freely available screen-readers, a quality speech synthesiser remains the critical component needed by any visually impaired individual for unaided computer access.

Owing to its superiority, *anReader* has quickly gained popularity among the visually impaired computer users in all of the countries of Western Balkans, and its use has resulted in a tenfold multiplication in their number, earning it the status of an official aid for the visually impaired, available to the visually impaired in Serbia through the Institute for Health and Social Care of the Republic of Serbia.

The new, higher quality of synthesised speech in Serbian and the potentials of its TTS engine for South Slavic languages were recognised very soon and, consequently, *anReader* was awarded the first prize of the Serbian Society of Informatics as the best applied software product in 2004.

3.2 Audio library for the visually impaired

There are more than 10.000 persons in Serbia with a visual disability of some kind, and a much larger number throughout the region of Western Balkans. The greatest centre for education of the visually impaired in Serbia and the entire Western Balkans is the School for the Visually Impaired Children „Veljko Ramadanović“ in Zemun. Until the introduction of the Audio library for the visually impaired (ABSS) (Mišković et al., 2005), written information necessary for education of the pupils of this school had been available in the form of Braille books, which are well known to be very impractical and extremely expensive to prepare, store and maintain, as well as audio recordings of books read out by human speakers, which have basically the same drawbacks. Preparation of both Braille and audio-books is also a lengthy process, making them quite inconvenient as media for accessing constantly changing information.

The Audio library for the visually impaired was developed in answer to these problems. It is a web-accessible client-server system in which a large quantity of books and texts from other sources is stored at the server side, while the client application enables an individual user to access the desired text, download it and have it converted to speech using a TTS system, namely, *anReader*. Searches by author name, genre and content are supported, and navigation through texts is intuitive and efficient due to a number of useful options. The texts are stored in an encrypted format in order to ensure the legal rights of copyright owners, preventing the users from being able to copy or print them.

As mentioned before, the Audio library is organised as a client-server application (Fig. 6). The administrator application, the database of books and the server in charge of handling user requests and accessing the database are situated on the server side. The books are internally stored in HTML format, which facilitates the retrieval of particular paragraphs before actual synthesis of speech. The client side contains an application intended for direct interaction with the visually impaired users as well as network communication. The user interface comprises two modules – *anAdministrator* (on the server side) and *anKlijent* (on the client side). The *anAdministrator* module is in charge of library administration, enabling inclusion and management of new books as well as search for (and within) the existing. The latest version of the library (Mišković et al., 2006) is multilingual, taking full advantage from

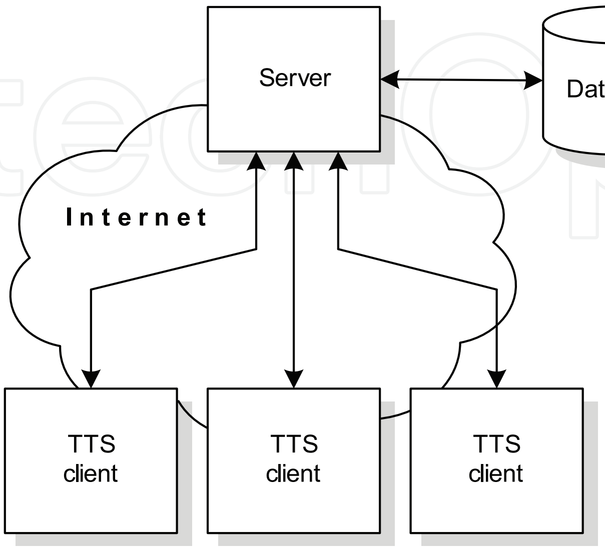


Fig. 6. Internal organisation of the Audio library

the fact that *anReader* has been developed for Croatian and Macedonian language as well, and that speech synthesis integrated into MS Windows can be used for reading books in English as well.

The *anAdministrator* module is completely speech enabled, which in turn enables the visually impaired to administrate the library themselves. The functionality of the application (connecting to the database and performing queries) is realised through ODBC (Open Database Connectivity) drivers for MySQL. The *anKlijent* module is speech enabled as well, which eliminates the need for any additional screen reader (a solution that would be hardly possible to use anyway, since the complete interface is in the language of the user's choice).

The executive module of the entire application is *anKlijent*, which relies on the communication module and on Microsoft SAPI 5.0, which provides access to Windows virtual speakers. The SAPI interface offers a number of advantages related to automatic handling of audio-devices, multi-threading, speaker selection etc. For that reason, besides AlfaNum TTS, which is implemented as two virtual speakers, *anKlijent* can also use the original Windows speech synthesis, which is suitable for handling texts written in English.

The initial version of the Audio library used RPC (*Remote Communication Protocol*) for communication. However, introduction of web access in the version 2.0 required implementation of new routines and a higher degree of control. For that reason, it was necessary to implement a custom protocol, based on ASR and TTS IP servers, which better answered the needs of dial-up users in particular, and the AlfaNum IP server, described in detail in section 1.2, was used for this purpose. Thus, the library has become a system independent from the actual location of the server and the database of books, since it allows the use of Internet for client-server communication.

The Audio library, as such, represents a significant step towards equality in education and access to information for the visually impaired. It is also a very convenient tool for all those who prefer textual content to be read out to them aloud while they are busy performing other tasks at their computers.

3.3 Voice enabled web sites

One of the recently developed applications of speech synthesis is enabling arbitrary web sites with speech synthesis through an IP TTS server particularly designed for this purpose. Owing to this system, visitors of web sites are able to listen to textual content instead of reading it, leaving their eyes free for some other task.

The interface to the server is remarkably simple, based on a PHP library and an accompanying javascript, facilitating integration of TTS functionality into existing web sites with minimal human intervention. The PHP library is universal, and the javascript is easily adaptable to each particular website.

The TTS server is optimised for enabling websites with speech synthesis through streaming of mp3 compressed sound, which results in virtually instantaneous server response. The server contains a minimum HTTP server within, which supports GET requests for file delivery by responding to them by direct sending of mp3 streams from the buffer in case the entire text has not yet been synthesised or by sending the recorded file in case the synthesis has been accomplished. Mp3 content is delivered to a Flash mp3 player embedded into the client application.

The process of requesting and obtaining synthesised speech can be summarised as follows (Fig. 7):

- WEB browser issues a request for a speech enabled web page;
- WEB server issues a TTS request;
- TTS server initiates synthesis and responds by sending the synthesised file name back to the WEB server;
- WEB server responds to the WEB browser by sending HTML content (with the embedded player's ″file″ parameter set to the actual name of the synthesised file);
- WEB browser displays the page and requests mp3 encoded speech from the TTS server;
- TTS server's embedded HTTP server responds by sending the mp3 stream.

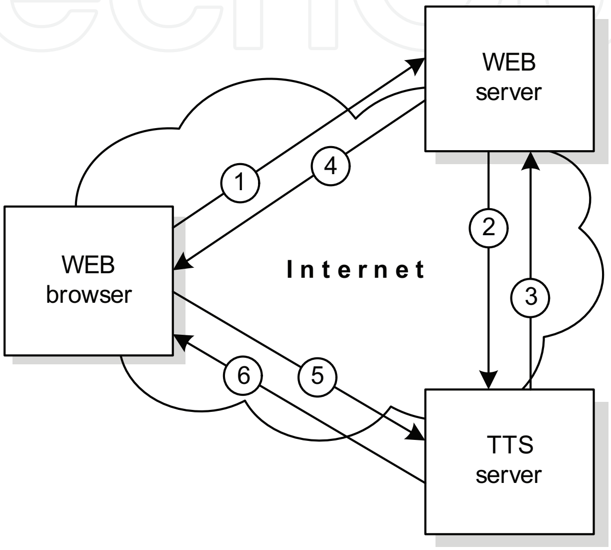


Fig. 7. Retrieval of synthesised speech from a speech enabled web site

The first speech enabled web site in the Western Balkan region is the site of Radio television Vojvodina (the northernmost province of Serbia) (http://www.rtv.rs), using the AlfaNum TTS engine. This web site was speech enabled in May 2009. After this pilot project has been successfully carried out, the interest for this web site feature across the Western Balkan countries has been on the increase and the AlfaNum team has recently obtained support from the Ministry of Science and Technological Development of the Republic of Serbia in the effort of enabling a significant number of Serbian web sites with speech.

4. ASR&TTS applications

4.1 Web portal Kontakt

Text-to-speech, as a technology with a wide range of application, becomes even more powerful when coupled with ASR. One of the examples of this is the web portal Kontakt, developed by the same team (Ronto & Pekar, 2005). This portal, intended primarily for visually impaired and elderly users, can be accessed by both computer users and those who do not own or use a computer since it is accessible by telephone as well, and it is essentially an Internet site whose contents are updated automatically from the websites of 4 well-known news sites in Serbia. Furthermore, authorised users can access it and submit information of particular interest to the visually impaired and/or the elderly. Each time the contents of the website are updated, the menu structure in the interactive voice response (IVR) interface is updated automatically as well. The users can, thus, navigate the site through voice commands and receive information via synthesised speech. Through the same interface, the users can change the speaker, speech rate and pitch, according to their own preferences. The portal is accessible via the intelligent network at a 0700 telephone number, which means that each user pays only the price of the local telephone call, regardless of the actual origin of the call.

The portal relies on a speech database as a source of information. In order to present the requested information to the user, it is necessary to send a query to the database, and receive the requested information in response. For the system to be efficient enough, it has to provide simultaneous access to information to a sufficient number of users. If the portal is accessed via telephone, the entire human-machine communication is carried out via speech.

In this case, as presented in Fig. 8, the communication in the system is based on interactive voice response (IVR) applications which handle one telephone line each through ASR and TTS IP servers and retrieve the requested information from the database mentioned above. The advantage of such a solution is in the fact that ASR and TTS servers can be remote and dedicated exclusively to speech recognition and synthesis. A possible disadvantage would be the delay in response in case of server overload.

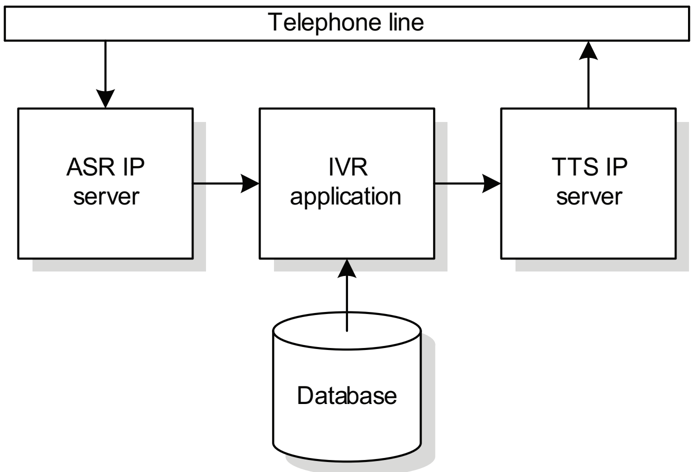


Fig. 8. Handling user requests in the system

The database in Fig. 8 is actually a MySQL database, which enables automated generation of web pages using PHP scripts, as well as a reliable connection with the IVR application realised through the C API of MySQL. The contents are refreshed using a puller application designed so as to gather new contents from the Internet on a periodical basis. These contents can be managed via an ordinary browser under administrative credentials. The system communicates with the telephone line using a Dialogic CTI card, and communication is controlled through Dialogic dx and Global Call API. The number of telephone lines that can be handled depends on the number and type of Dialogic cards integrated into the PC platform of the system.

Having in mind the technological limitations of ASR, users can form their queries in a standard format described by a number of *grammars*, which define the set of words, phrases and their combinations expected as input to the ASR process at any given moment. An example would be the initial grammar of the system, actual at the moment when the user initially addresses the system:

```
cmd = TEME | NASLOVI;[1]
gr = <gar>;
main = [$gr] [$cmd] [$gr];
```

where `<gar>` stands for any noise that is to be ignored during recognition. A successful navigation through a menu structure requires a new grammar to be defined at each point in the dialogue. The menu structure depends on the defined topics in the database, and on the other side, changes in the content of the database must have as little influence as possible on the design of the entire system. The only acceptable solution is to automatically generate grammars from the database. In the latest version of the system, grammars are defined at the initialisation of the IVR application. The ASR server is started thereafter, and thus it uses up-to-date grammars. The only deficiency of such an approach is that, if the database is refreshed while the application is active, the ASR server needs to be restarted.

For any newly generated grammar to be successfully used by the ASR server, it is necessary to communicate the location of each new grammar file to the server. This is done via the initialisation file of the ASR server, which contains all settings relevant to the functioning of the server, such as the parameters related to speech signal processing, recognition itself, IP port through which server communicates and other data related to the server. These data contain the vector of recognisers, defining the name, grammar file paths, postprocessor, pronunciation dictionary and phonetic transcriptor for each recogniser. In this context, the term "recogniser" denotes a set of rules to be used for recognition at a given moment. For the ASR server to be initialised with all newly generated grammar, it is necessary to establish a recogniser for each one of them, with all the necessary parameters, and include it into the vector of recognisers. From the point of view of the IVR application, defining all parameters of a recognition amounts to the selection of the appropriate recogniser.

The parameters of synthesis, on the other hand, can be configured by the users themselves. Each time a user logs out, the synthesis parameters of his/her choice are stored in the database, and the next time the user logs in, the same values are restored.

As such, the system was designed as a point of support to a number of the visually impaired and the elderly. As a project of great importance, the portal Kontakt has received support from Telekom Srbija, the Lottery of Serbia, as well as the community of the visually impaired in Serbia. The similar portal has been established in Croatia, with the only difference in that, at the moment, it updates its contents from a single news website.

4.2 iTEMA E-mail reader

iTEMA (Intelligent Telephone E-mail Access) is a multilingual CTI application for voice-enabled telephone access to user e-mails, developed within the joint EUREKA project E!3864 (Žganec Gros et al., 2006; Žganec Gros et al., 2008).

The architecture of the iTEMA system contains an interface towards a number of SAPI compatible TTS engines. The central element of the system is a dialogue manager connected to both telephone and Web interface (Fig. 9). Personal settings for each user, such as mobile phone number, PIN, e-mail access parameters, are stored in a database.

A user dials the number of the iTEMA user service and a human-machine dialog is initiated. Authentication is performed based on ANI and PIN, and followed by a personalised dialog enabling simple and intuitive navigation through a menu system. Through this dialog users

[1] which can be translated from Serbian as `TOPICS | TITLES`

can select messages they want to listen to, delete, or reply to using one of the pre-defined templates.

Beside drivers and business people, iTEMA also provides e-mail service to those who have difficulties when using a computer but use a telephone as a matter of routine (the visually impaired, many of the elderly etc.). The iTEMA project thus represents material support to the e-inclusion programme of the EU.

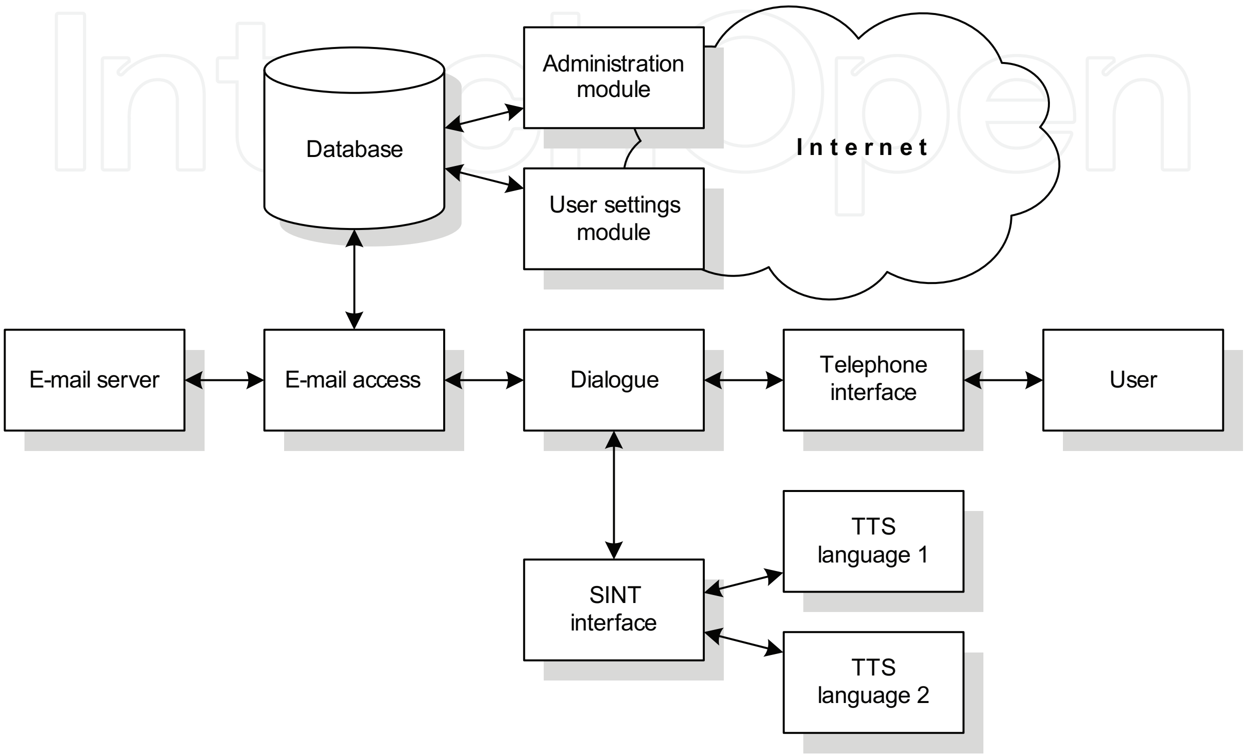


Fig. 9. Internal architecture of the iTEMA system

4.3 Computer games for the visually impaired

Besides the applications mentioned in the previous sections, the AlfaNum TTS engine, coupled with the AlfaNum ASR engine, was also used to create new computer games designed for entertainment and education of visually impaired children (Delić & Vujnović Sedlar, 2010; Lučić et al., 2009; Mester et al., in press).

In their study (IGDA, 2004) the International Game Developers Association discusses the availability of games to every person with a disability. The study presents speech synthesisers (TTS), screen readers and speech recognition (ASR) as assistive technologies which can contribute to a greater availability of games for the visually impaired. Unfortunately, the application of these new technologies which would allow the adaptation of the user interface to the ergonomics of the visually impaired is not the primary concern for the game industry. On the other hand, a connection between the visually impaired children and these technologies is of a crucial importance to their inclusion into the society (Perepatić, 2010).

User interfaces differ from game to game, and consequently, the ability to be adapted to the visually impaired, as well as the process of adaptation itself, also differs from game to game. Audio signals are the key factor of these games when adapted to the visually impaired. These audio signals have to differ from each other so that the player can easily identify them and respond at the right moment and in a right manner. At http://www.AudioGames.net, one of the best known web sites with audio games, there are more than 300 audio games, and their classification and examples are given in (Mester et al., in press).

In the process of creation of audio games particular attention is paid to presenting information in audio form, because sound presentation must carry all relevant information that allows the player to react timely and in the right manner. The GUI of a video game carries most of the information, which gives particular broadness and freedom while developing such games as opposed to audio games. Portraying all relevant information in audio form presents an interesting challenge because the presentation of audio information to the user is limited. In sound-based games the player gets a mental picture of all present objects and persons by listening to the sounds which characterise them. Stereo positioning is used to spatially distinguish the sounds of objects. It allows the sound to traverse from left to right and vice versa. These sounds are critical for the player and his/her understanding of the game. Yet stereo positioning only gives the player one dimension, which is a constraint compared to the two dimensions of a screen.

For example, Delić & Vujnović Sedlar (2010) have created the first audio game for the visually impaired with ASR and TTS in Serbian. It is a simple memory game with sixteen fields hiding eight pairs of objects. Having in mind the characteristics of binaural hearing the authors have decided to present the horizontal position of the field by simple stereo presentation (different interaural levels between ears), and to indicate the vertical position of the field by using different audio frequencies (pitch of synthesised speech – TTS) similarly to (Gärdenfors, 2003). The user has a sensation of sound coming from an exact position on a four-by-four grid facilitating memorisation of object locations. The user can select the square either using verbal commands (by pronouncing the coordinates of the square – ASR) or simply using the keyboard. The memory game has been developed as Microsoft application in C#, using Microsoft Visual Studio 2008, with sound supported by Microsoft DirectX SDK.

Another example of a computer game suitable to be adjusted to the visually impaired using audio and speech technologies is a set of very simple geometric puzzles named Lugram (Lučić & Vujnović Sedlar, 2009). Geometry as a branch of mathematics is one of the most difficult areas from the point of view of adaptation for the visually impaired, but on the other hand, it is very useful for orientation in space and executing everyday tasks. Following the example of the ancient Chinese Tangram, Lugram has been designed as a puzzle game aimed at composing given geometric figures. Elements to be used for assembling are square tiles containing geometric figures such as triangles, rectangles or squares, as shown in Fig. 10. One direction of the development of the game led to its successful adaptation for visually impaired users (Lučić et al., 2009), and opened the perspective for a special challenge of creating a new version of the game for the blind. Lugram has been developed using C++ and Macromedia's Director.

The audio interface of the game consists of speech, music and various sorts of audio effects. Speech is mostly used to introduce the user to the guidelines and rules of the game, and for this purpose TTS is most commonly used. Synthesised speech can be used in the game itself more or less, depending on the type of the game. Generally, if greater authenticity of the situation is to be achieved, or in order to ensure that the right reaction will be made, in most cases synthesised speech can be replaced by recorded – natural speech. Audio effects are commonly used to illustrate situations or various objects in the game (Ratanasit & Moore, 2005). For the visually impaired it is of particular importance to have certain audio effects which would tell them whether their reaction was suitable or not. Music can also be a good element for depicting states and situations a player can find himself/herself in, or it can be used just as a background.

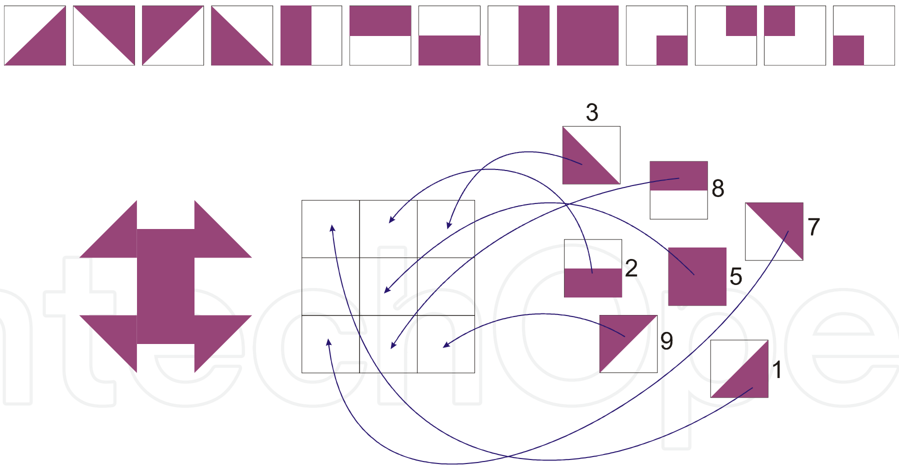


Fig. 10. The squares containing geometric figures and an example of a task in the computer game Lugram intended for the visually impaired

Due to the lack of sight, the blind rely on other senses heavily, especially on the senses of touch and hearing, making them more advanced (Doucet, 2005), and allowing them to easily learn to use keyboards very skillfully. The alternative to using keyboards is ASR. Because of intra- and interpersonal differences in the voices of speakers, different setups and qualities of microphones and the communication channels, as well as different levels of ambient noise, ASR is a very demanding task for the computer games and is not well developed for all languages. Studies usually mention using ASR for issuing certain voice commands, but unrestricted human-to-computer speech communication is not so common yet.

Audio interfaces enable the visually impaired to play games more equally to other players. As speech is an extremely important element of such an audio interface, speech technologies are essential for playability of games with audio interfaces. Development of speech technologies is thus a contribution to inclusion of persons with disabilities into the society.

5. Conclusion

The applications presented in this chapter clearly show the importance of development of speech technologies. Having in mind the extreme language dependence of these technologies, and the fact that, unlike most other technologies, they cannot simply be „imported from abroad", it is very important that scientific teams from the region should be actively engaged in their research and development. Only thus we can expect that the 20 million inhabitants of this part of Europe will be able to communicate with machines by speech in their native languages in a near future.

5.1 Directions of further research and development

One of the directions of furher research and development of speech technologies is multilingual and multimodal human-computer interaction involving not only ASR and TTS but speaker and emotion recognition as well. Besides further research aimed at increasing the quality of ASR and TTS components, research related to implementation of speech technologies on embedded platforms is also under way, aimed at their application in small portable devices. ASR and TTS have an extremely wide area of application, and some projects are initiated to apply developed speech technologies in South Slavic languages in smart homes, cars, industry, robots and toys. This would enable a number of other applications such as dictation, automated transcription of radio and TV programmes, meetings and sessions, telephone conversations etc.

6. References

Delgado, R. & Araki, M. (2005). Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment. Wiley, ISBN: 978-0-470-02155-2

Delić, V.; Sečujski, M. & Pekar, D. (2005). On anReader, its features and first applications (in Serbian), *Computers and the Blind*, Zagreb, Croatia

Delić, V.; Sečujski, M. & Tekić, Ž. (2006). A Contribution to Human-Machine Communication in Serbian, Croatian and Macedonian Language, *Proceedings of 17th DAAAM Symposium (Intelligent Manufacturing&Automation: Focus on Mechatronics&Robotics)*, DAAAM Publ., Vienna, TU Wien, pp. 101-102, ISSN 1726-9679, ISBN 3-901509-57-7

Delić, V. (2007). A Review of R&D of Speech Technologies in Serbian and their Applications in Western Balkan Countries, *Keynote lecture at 12th SPECOM (Speech and Computer)*, pp. 64-83, ISBN 6-7452-0110-x, Moscow, Russia, October 2007

Delić, V. & Vujnović Sedlar, N. (2010). Stereo Presentation and Binaural Localization in a Memory Game for the Visually Impaired. *Lecture Notes in Artificial Intelligence, Springer,* A. Esposito et al. (Eds.): COST 2102 Int. Training School 2009, LNAI 5967, pp. 354-363. Springer, Heidelberg, ISSN 0302-9743

Delić, V.; Sečujski, M.; Jakovljević, N.; Janev, M.; Obradović, R. & Pekar, D. (2010). Speech Technologies for Serbian and Kindred South Slavic Languages, Chapter in the book *Speech Recognition,* SCIYO, ISBN 978-953-7619-X-X (accepted for publication)

Doucet, M.-E.; Guillemot, J.-P.; Lassonde, M.; Gagné, J.-P.; Leclerc, C. & Lepore, F. (2005). Blind subjects process auditory spectral cues more efficiently than sighted individuals. *Exp. Brain Res.* 160: 194–202

Gärdenfors, D. (2003). Designing Sound-Based Games. *In Digital Creativity,* 14(2), 111-114

IGDA - International Game Developers Association. (2004). Accessibility in Games: Motivations and Approaches. Retrieved on February 15, 2005, from http://www.igda.org/accessibility/IGDA_Accessibility_WhitePaper.pdf

Janev, M.; Pekar, D.; Jakovljević, N. & Delić, V. (2008). Eigenvalues driven gaussian selection in continuous speech recognition using HMMs with full covariance matrices. *Applied Intelligence*, *Springer Netherlands,* DOI: 10.1007/s10489-008-0152-9, (Print, accepted) December 2008, ISSN 0924-669X (Print) 1573-7497 (Online), Available at: http://www.springerlink.com/content/964vx4055k424114/

Lučić, B. & Vujnović Sedlar, N. (2009). Geometric Puzzle Lugram - Development and Application (in Serbian). In *Proceedings of TELFOR: Vol. 17.* Belgrade

Lučić, B.; Vujnović Sedlar, N. & Delić, V. (2009). Computer game Lugram - version for visually empaired children (in Serbian). In *Proceedings of TELFOR: Vol. 17.* Belgrade

Mester, Gy.; Stanić-Molcer, P. & Delić, V. (in press). Educational Games, A chapter in the book *Business, Technological and Social Dimensions of Computer Games: multidisciplinary developments*, Publisher: IGI Global, PA, USA (accepted for publication)

Mišković, D.; Đurić, N.; Pekar, D. & Jakovljević, N. (2007). Alfanum word spotter as a form of ASR application. *Proceedings of 51th ETRAN*, Herceg Novi – Igalo, June 4 – 8, 2007

Mišković, D.; Zindović M. & Pekar D. (2007). Postprocessing methods for validation of Alfanum ASR recognition system. *Proceedings of TELFOR,* Beograd

Mišković, D.; Vujnović, N.; Sečujski, M. & Delić, V. (2005). Audio Library for the Visually Impaired as an Application of TTS Tehnology (in Serbian). *Proceedings of 49th ETRAN,* Vol II, pp. 400-402, ISBN 86-80509-54-X, Budva, Montenegro, Publisher: Society for ERAN

Mišković, D.; Vujnović, N.; Sečujski, M. & Delić, V. (2006). Audio Library for the visually impaired – ABSS 2.0. *Proceedings of DOGS (Digital Signal and Image Processing)*, pp. 67-70, Vršac, Serbia, Publisher: Faculty of Technical Sciences, Novi Sad.

Nöth, E.; Horndasch, A.; Gallwitz, F. & Haas, J. (2004). Experiences with Commercial Telephone-based Dialogue Systems. *it – Information Technology*, Vol. 46, No. 6, 306-314, ISSN: 1611-2776

Pekar, D.; Delić, V.; Molerov, S.; Kočiš, G. & Vuković, R. (2007). System for automatic recognition of audio clips in radio and TV programmes, Patent in Serbia P-2007/0505

Perepatić, J. (2010). Possible benefits of computer games to the visually impaired children - a survey of parents' opinions. Non-Government Organisation "Iskrica", Novi Sad, Serbia (unpublished).

Ratanasit, D. & Moore, M. M. (2005). Representing Graphical User Interfaces with Sound: A Review of Approaches. *Journal of Visual Impairment and Blindness,* 99(2), 69-84.

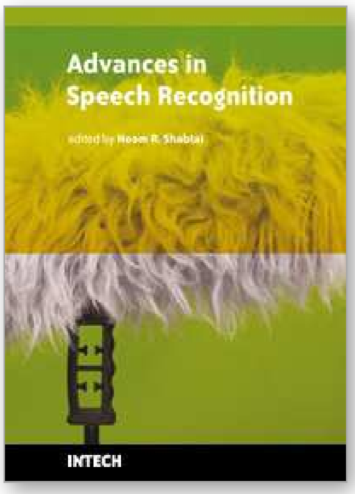
Ronto, R.; Pekar, D.; Đurić, N. (2005). Developing a Telephone Voice Portal with ASR and TTS Capability (in Serbian). *Proceedings of 49th ETRAN,* Tom II, pp. 392-395, ISBN 86-80509-54-X, Budva, Montenegro, June 2005, Publisher: Society for ERAN. Portal is available at: http://www.alfanum.ftn.uns.ac.rs/kontakt/

Sečujski, M.; Obradović, R.; Pekar, D.; Jovanov, Lj. & Delić, V. (2002). AlfaNum System for Speech Synthesis in Serbian Language. *Proceedings of TSD (Text, Speech and Dialogue)*, ISBN 3-540-44129-8, Brno, Czech Republic, September 2002. *Lecture Notes in Artificial Intelligence*, *Springer-Verlag Berlin Heidelberg,* LNAI 2448, pp. 237-244, ISSN 0302-9743

Sečujski, M.; Delić, V.; Pekar, D.; Obradović, R. & Knežević, D. (2007). An Overview of the AlfaNum Text-to-Speech Synthesis System, *Proceedings of 12th SPECOM (Speech and Computer)*, pp. Ad.Vol. 3-7, ISBN 6-7452-0110-x, Moscow, Russia, October 2007, Demo is available at: http://www.alfanum.co.rs/anreader.html

Žganec Gros, J.; Delić, V.; Pekar, D.; Sečujski, M. & Mihelič, A. (2006). The iTEMA E-mail Reader, *Proceedings of IS-LTC*, pp. 230-233, Ljubljana, Slovenia, ISSN 1581-9973, ISBN-13 978-961-6303-83-X.

Žganec Gros, J.; Delić, V. & Pekar, D. (2008). Listen to your e-mail through the telephone, *eStrategies|Projects EUROPE*, Vol. 2, No. 3, pp. 42-43, British Publishers, ISSN 1752-5152.



Advances in Speech Recognition
Edited by Noam Shabtai

ISBN 978-953-307-097-1
Hard cover, 164 pages
Publisher Sciyo
Published online 16, August, 2010
Published in print edition August, 2010

In the last decade, further applications of speech processing were developed, such as speaker recognition, human-machine interaction, non-English speech recognition, and non-native English speech recognition. This book addresses a few of these applications. Furthermore, major challenges that were typically ignored in previous speech recognition research, such as noise and reverberation, appear repeatedly in recent papers. I would like to sincerely thank the contributing authors, for their effort to bring their insights and perspectives on current open questions in speech recognition research.

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821