A Review of R&D of Speech Technologies in Serbian and their Applications in Western Balkan Countries

Vlado Delić
Faculty of Technical Sciences, University of Novi Sad, Serbia

Abstract

Automatic speech recognition (ASR) and text-to-speech conversion (TTS) are language dependent technologies. Considerable development requirements have restricted speech technologies to widely spoken languages, but now the solutions for other languages have appeared as well. For instance, ASR and TTS for Serbian have achieved a sufficient level of quality and their first applications have been launched. A brief review of R&D of speech technologies in Serbian and kindred South Slavic languages is given in this paper. The emphasis is on the key research contributions of a R&D group from the University of Novi Sad, which enabled greater progress in both development and applications of speech technologies in West Balkan countries (WBC). First applications are dedicated to the people with disabilities. They are described in more detail, as well as several other ASR&TTS applications.

1. Introduction

Unlike many other novel technologies, speech technologies cannot be so easily imported from abroad and applied in an area where another language is spoken. Just as humans are unable to correctly pronounce a text written in an unknown language or understand speech in an unknown language, speech enabled machines cannot be used in any language other than the one they have been originally developed for.

Development of speech technologies for a certain language requires knowledge in areas such as linguistics, phonetics, acoustics, mathematics, programming as well as digital signal processing [1]. It is necessary to bring together the knowledge from these areas and implement it into the available computer resources to enable a computer to understand human speech – using automatic speech recognition (ASR), as well as to respond via speech – using text-to-speech synthesis (TTS). Both are very complex multidisciplinary problems which require engagement of teams of experts from aforementioned areas, as well as ample time and financial resources. For that reason, solutions for widely spoken languages were the first to appear. However, due to the aforesaid language dependency, they could not have been used in areas where other languages were spoken.

Once humans are able to address a computer in their native language and a computer is able to respond in the same language, it will be possible to "talk" to other appliances, as well as industry machines, cars, toys and robots, or to a remote computer via telephone. We will have an increasing need to speak to appliances in our midst and thus it is extraordinarily important to develop speech technologies, each of us for their own language. There is a profusion of languages in the world and each nation strives to preserve and protect their language. Speech technologies have the potential to overcome an evident language barrier between collocutors who do not speak the same language. Figure 1.1. shows a block diagram of such an automatic speech interpreter. For instance, in a conversation between a Russian and a Serbian, speech in Russian language is converted into text using automatic speech recognition, then a translation of the text from Russian into Serbian is carried out, and the translation is finally converted into speech in Serbian language using text-to-speech synthesis. Evidently, such a system should be able to function in the opposite direction as well, using ASR in Serbian and TTS in Russian. For language pairs without appropriate machine translation systems a third language could be used as an intermediary.

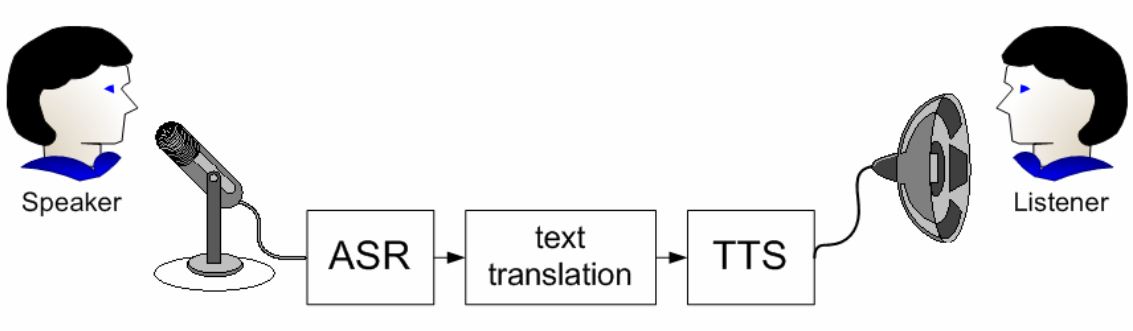

Fig. 1.1. Automatic speech translation

We in Serbia have understood the potential of speech technologies in preservation of the language as well. We have succeeded in overcoming certain research barriers related to the complex morphological structure of Serbian and kindred South Slavic languages. In this way further progress in development of speech technologies for the Serbian language has been made possible and we can now proudly state that we have introduced our language into a relatively narrow circle of world languages for which speech technologies have attained the necessary level of quality and that they find their first wide applications.

A brief review of experiences in development of speech technologies for the Serbian language as well as their adaptation to kindred South Slavic languages will follow [2]. The paper will also present their first applications in Western Balkan Countries (WBC): Serbia, Croatia, Bosnia and Herzegovina, Macedonia and Montenegro [3]. WBC include Albania as well, however the Albanian language does not belong to the South Slavic group.

The author of this paper has been the leader of an R&D team called AlfaNum, which has been dealing with technical aspects of research and practical implementation of speech technologies, as well as devising their applications. It is his intent to give a general presentation of speech technologies (ASR and TTS) in this paper, at a conceptual level, without going into technical details and mathematical foundations. Such a text should present speech technologies to linguists, phoneticians, acousticians, psycholinguists as well as other experts that deal with problems of speech communication from their own specific point of view.

2. The Rise of Speech Technologies and Their First Applications

The idea of speech communication between humans and machines has been motivating numerous researchers from all around the world for a number of years. Research and development of automatic speech recognition and text-to-speech synthesis has been invested into for a long time, but both technologies have turned to be very intricate multidisciplinary problems – one of the greatest problems humans have come across. The first applications, at the end of the last century, have come too early, probably under pressure of investors and their great expectations. Back then, speech technologies were not yet of sufficient quality and the designers of their applications did not fully take into account their actual capabilities, often making their application counterproductive. It was often the case that the user pronounced one thing and the machine understood another, which frustrated many users and lead them to the conclusion that the problem was simply too complex to be solved in a satisfactory way.

Speech technologies are still far from being perfect and will remain so for many years to come, unable to understand human speech as correctly as a human does or produce speech that would be indistinguishable from human speech. However, the author of this paper has already had the pleasure to hear a computer pronounce certain phrases faster and more precisely than any human can. The same computer is already capable of talking to tens of people simultaneously, understand

speech commands issued by humans and talk back using synthesised speech. This synthesised speech may not be as rich in prosody features as original human speech, but we already understand it much better than, e.g., speech produced by many persons with speech disabilities. This shows that at least persons unable to speak can make use of a TTS system to say what they want using synthesised speech. Quality of synthesised speech has been significantly improved in the last several years and it is being used by the visually impaired to read books, newspapers and letters unaided, using only a speech enabled computer. In this way the visually impaired are more equal in their education, access to information as well as privacy in written communication.

On the other hand, recognition of spontaneous speech with words from a large vocabulary (more than several thousand words) has still not been developed for a majority of world languages. However, small and medium-sized vocabulary speech recognition (several tens or hundreds of words) can be of quite satisfactory use in a guided dialogue between humans and machines. For instance, a cleverly designed and guided dialogue between a human and a computer (e.g. via phone) can use a different set of words in each phase, words that are easy to distinguish. In the next phase a new set of words – a new vocabulary – can be used, making the human-machine dialogue very rich and efficient. The most important thing for an application designer is to be aware of the actual capabilities of the technology, not to rely on theoretical data on error rates only, but put the IVR application to a proper test in real conditions.

CTI applications of ASR and TTS should make most of the advantages of speech technologies over touch-tone dialling and reproduction of pre-recorded voice messages. In cases where human-machine dialogue includes selections among a small number of options and where it is possible to predict all possible answers or their combinations, there is no real need to include speech technologies. However, there are many more applications where the number of options is much greater and it is not practical or not possible to pre-record all the answers, and this is the area where speech technologies are of great use. For instance, if a speech technology application offers the user to listen to his/her e-mail messages, it is clear that the conversion of text into speech has to be carried out automatically, on-line. In the same way, an IVR application offering information related to a railway time-table would be much more practical if it allowed the user to ask for information using speech. Such applications are already quite common in countries where widely spoken languages are in use, but there are many countries where speech technologies have not yet been developed or are not yet in wide use.

3. ASR and TTS for South Slavic Languages Spoken in WBC

An interesting fact related to speech technologies is that unlike many other novel technologies they are extremely language-dependent and cannot simply be imported from abroad and applied in an area where another language is spoken. Speech machines “trained” to read texts (convert them into speech) in one language cannot read it in another, foreign language. It is impossible for a human as well to read aloud a text in a completely unknown language correctly. Likewise, just as a human is unable to understand speech in an unknown language, a speech machine cannot recognise speech in any language other than the one it has originally been “trained” for. However, it should be pointed out that ASR is somewhat less dependent on language than TTS. In particular, in case of phonetic speech recognisers (capable of recognition of phonemes in specific contexts) it is relatively easy to adapt the speech recogniser trained for one language for recognition of speech in a language with a similar phonetic inventory. An example of such closely related languages are South Slavic languages: Serbian, Croatian, Bosnian, Macedonian, Bulgarian and Slovenian, as shown in Fig. 3.1. The first three of them are actually virtually identical (until recently they have all been variants of a single Serbo-Croatian language) and collocutors speaking any of those three languages do not need an interpreter. Macedonian has some similarities to Serbian and Bulgarian, and Slovenian is the most specific of the six. At any rate, it is a less complex task to adapt a speech technology (speech recognition in particular) to a kindred language than to a completely unrelated language.

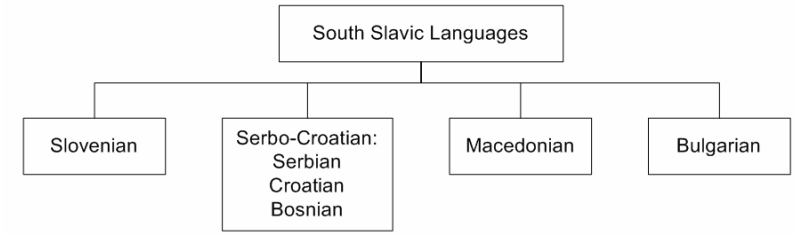

Fig. 3.1. Western Balkan countries and South Slavic languages

The difference between South Slavic and other Slavic languages is somewhat greater, but an exchange of experience between researchers who developed speech technologies for these languages is desirable. South Slavic area alone is relatively small (less than 30 million inhabitants) and there are no other research teams with practical results of satisfactory quality. Beside the R&D group at the University of Novi Sad (Serbia), there are two groups in Slovenia (Ljubljana and Maribor) focused on Slovenian language (less than 2 million speakers), as well as two smaller groups in Bulgaria (Sofia) and Croatia (Rijeka). The first applications in Western Balkan Countries are the result of the effort of the AlfaNum team at the University of Novi Sad (Faculty of technical sciences), as well as the "AlfaNum" spin-off company, dedicated to the development of applications based on technologies developed at the University.

3.1. AlfaNum TTS Conversion for Serbian and Kindred South Slavic Languages

The AlfaNum R&D group has solved certain research challenges that remained a barrier for other R&D teams from Western Balkan countries. During the development of the first high-quality TTS for Serbian language, this team has encountered many problems linked to bridging the gap between plain text and synthesised speech with all its typical features such as intelligibility and naturalness.

There is no explicit information in a plain text concerning prosodic features such as phone durations, pitch contours or energy variations [4]. These factors also depend on the meaning of the sentence, emotions and speaker characteristics, which further aggravates the task of attaining high naturalness of synthesised speech [5]. While Serbian and Croatian are tonal languages, having high-low pitch patterns permanently associated with words [6], Macedonian is a pitch-accented language with antepenultimate stress on most words, excluding clitics, words of foreign origin as well as some other word groups. However, a uniform dictionary-based strategy for lexical stress assignment has been successfully employed in all three cases.

This was the key step towards a high-quality TTS for Serbian and thereafter for other kindred South Slavic languages. Built using a specially designed software tool, the accentuation-morphological dictionary contains over 3 million inflected word forms. The dictionary is used in the first phase of text-to-speech conversion, while TTS engine tries to identify prosodic information hidden in the text and to convert text into a representation suitable for driving the low-level component of the system. Morphological and syntactic ambiguities are resolved based on the dictionary as well as rule-based syntax analysis.

However, a separate dictionary and syntax analysis techniques were required for each language. Lexical stress assignment, as one of the most important factors influencing the shape of the pitch contour, is based on the algorithm described in [7]. This approach has proved to produce reasonably correct stress pattern, with word error rate as low as 2.8% for Serbian language.

After a complete text analysis, TTS engine has a good idea how it should pronounce the given sentence, and thus the second phase – the actual low-level synthesis of the speech signal – can begin, as shown in Fig. 3.2.

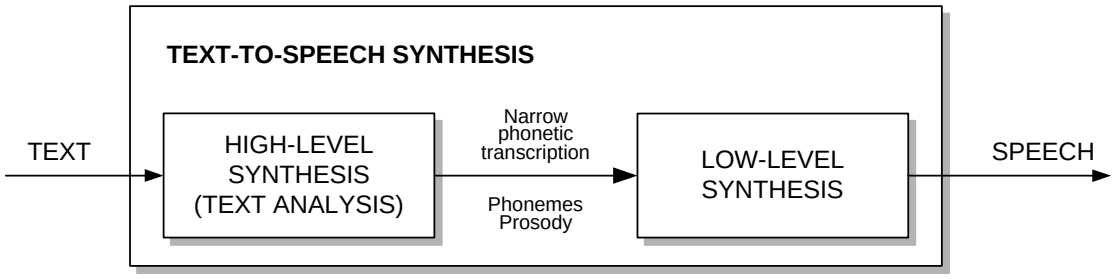

Fig. 3.2. The internal structure of the AlfaNum TTS system

Low-level speech synthesis is based on the concatenation of speech segments of variable size, selected at runtime from a large speech database containing more than two hours of labelled continuous speech. The boundaries of each phoneme are labelled semi-automatically (with the help of AlfaNum ASR) together with certain additional information regarding the manner of articulation, which will be of use in the process of speech segment selection. Although segment transitions are rendered nearly inaudible using the well known TD-PSOLA algorithm for speech segment processing and prosody manipulation, it is still desirable to use longer speech segments from a continuous speech database. Online speech segment selection is carried out according to the specifications resulting from the text analysis as well as the complete information on all speech segments available in the database. By keeping score of every phone in the TTS databases and its relevant characteristics, the use of phones in less than appropriate contexts was avoided, which further contributed to overall synthesised speech quality. This synthesiser is, thus, not diphone-based as a majority of other speech synthesisers 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. An alternative to TD-PSOLA has recently been introduced – hybrid harmonic/noise synthesis – with listening test still to be carried out.

In order to develop a multilingual text-to-speech conversion system, a separate speech database is generally required for each language. However, the Serbian database is actually used for Macedonian, leading only to a minor decrease in speech quality, due to the fundamental similarity between the phonetic inventories of the two languages. The Macedonian speech database is expected to be recorded soon. Curiously enough, in case of sufficient similarity between phonetic inventories across languages (such as Serbian, Croatian and Macedonian), the naturalness and overall quality of synthesised speech are influenced by accentuation and pitch contours much more than the original language of the database itself.

The speech synthesised by AlfaNum TTS is highly intelligible and reasonably natural-sounding, much more than any other attempts at speech synthesis in Serbian, Croatian and Macedonian so far. It is the first TTS engine that can read Serbian in both Cyrillic and Latin script, as well as in Latin without diacritics found on 6 out of 30 letters (š, č, ć, đ, ž, dž), which is the common way of writing e-mail and SMSes in most south Slavic languages. AlfaNum TTS correctly pronounces a number of foreign words commonly used in Western Balkan countries.

The most significant application of this system so far is *anReader*, a speech synthesiser for the visually impaired that, combined with software tools known as *screen-readers*, offers the visually impaired complete independence in computer access. The number of the visually impaired computer users in Serbia has increased significantly since *anReader* was presented for the first time, and its popularity in Croatia, Bosnia and Herzegovina, Montenegro and FYR Macedonia is also growing.

The main objection of many anReader users is related to a perceived lack of naturalness in sentence prosody. The accentuation pattern being correct, overall sentence f_0 contour is sometimes found to be too linear and not to relate to sentence syntax in a satisfactory way. For that reason, further research in this area will include overall improvement of sentence prosody using fully automatic corpus-based methods, and certain experiments regarding the synthesis of expressive content are also being planned.

3.2. ASR Issues and Solutions for the Serbian and Other South Slavic Languages

3.2.1. Human vs. Automatic Speech Recognition

Human listeners have the ability to focus their attention on the sound they are trying to listen to. A listener can recognise words in continuous speech and understand the meaning of the message without difficulty provided that he or she is familiar with the language used. Beside linguistic information, a speech signal also contains information about the speaker (his/her gender, emotional state, age). A human can extract all these additional information straightforwardly. Extraction of relevant acoustic features from short speech units (phones) is performed in the peripheral auditory system based on time-frequency analysis of a speech signal, and the information is used for phoneme recognition and ultimately for the recognition of the message itself at a higher cognitive level. For a sound to be perceptible by humans, it should have sufficient intensity and it should be located in the appropriate frequency range. This hierarchical structure enables humans to understand what was spoken even in case some features are missing from the received speech signal or are masked by other sounds. Humans are trained to recognise the essence of what was said, and usually not the exact word sequence. Current ASR systems do not exhibit such a level of intelligence, thus they focus on both informative and uninformative words equally.

An ASR system has a microphone and a processor instead of a peripheral auditory system and a brain. Speech sounds as well as all other sounds correspond to minor pressure differences in the air.

A microphone converts these varying pressure waves into varying electrical signals. These analogue electrical signals are then converted into digital signals i.e. arrays of numbers, which are forwarded to the processor. Continuous speech consists of a string of different speech sounds and the task of ASR is to track down relevant changes in the digitalised speech signal and to recognise the actual phonemes and words spoken. Unfortunately, these variations are not caused by different phone characteristics only. Speech signal is influenced by a number of different factors, making ASR quite a difficult task. For example, a recorded electrical signal depends on microphone characteristics, relative position of the microphone as well as background noise. Speech can also vary along with differences in the speech rate, pronunciation of words within and across speakers. Variations across a single speaker are caused by her/his health, emotional state and/or specific intentions of the speaker.

Both acoustical and time variability are present in the speech signal. Temporal variability is caused by differences in speech rate and dynamics. Acoustical variability is caused by phone characteristics as well as all other previously mentioned causes and their manifestations in the acoustical signal. Speech is a non-stationary process, consisting of a sequence of phones with various features and dynamical properties. One of the ASR aims is to extract features which are important in answering the question “What was said?”.

The ASR concept is based on the knowledge of how humans distinguish phones. A human distinguishes three physical properties: loudness (intensity), pitch (f_0) and timbre (spectrum envelope). Pitch is dependent on speaker and sentence intonation, loudness is related to sound energy, leaving the timbre as the principal feature for phone distinction. During the pronunciation of a single phone, vocal tract is appropriately shaped, modelling the air stream spectrum. Subjective frequency perception is nonlinear (it corresponds more closely to the Mel scale than to the linear frequency scale).

3.2.2. Speech Process Analysis and Modelling

Frequency analysis of speech signals is based on the Fourier transform of quasi-stationary segments referred to as frames. Frame duration is shorter than phone duration i.e. frame duration is usually from 10 up to 30 ms, while the average phone duration is from 50 up to 100 ms (average speech rate being about 10 phones per second). A shorter frame duration results in a more stationary segment, yet durations shorter than 20 ms cause the deterioration of the frequency resolution (spectrum component distinction). For example, a 20 ms long frame of a signal sampled at 8 kHz has a frequency resolution of about 100 Hz. Since frames are not chosen to be synchronous with any acoustic landmark, resulting features are smeared over transition regions. In order to overcome this shortcoming, a frame shift of about 10 ms is introduced causing successive frames to overlap.

In order to reduce system complexity, instead of several hundreds estimated spectrum samples only 10-18 cepstrum coefficients are actually used. These cepstrum coefficients represent the spectrum envelope which is the principal distinctive property of phones.

Temporal changes of the spectrum envelope contain important information for ASR. Features which describe these changes are referred to as dynamic features, consisting of the first and the second derivative of static spectrum envelope features. The use of these dynamic features results in an increase of feature vector dimension. This process of conversion of speech segments (frames) into feature vectors is known as feature extraction.

During feature extraction each frame corresponds to a point in a multi-dimensional feature space (MDFS). Successive speech frames form a trajectory in the MDFS. In order to reduce system complexity in the first ASR systems, these feature vectors were quantised and MDFS was a discrete space.

A single phone segment contains several (5-10) frames. If the spectrum envelope of all the frames which belong to a single phone were the same then all frames would correspond to a single point in

the MDFS. Unfortunately, spectral properties of a phone vary along with the speaker, background noise as well as different channel properties. For that reason, frames corresponding to the same phoneme cover an entire region. This dissipation of points is caused by coarticulation as well. The points corresponding to the frames near to the phone boundaries are closer to the regions that correspond to adjacent phones.

When a point describing acoustical properties of a current speech frame moves across the MDFS, it can remain within a single phone region for a shorter or a longer time, or pass into another, more or less remote region. The amount of time spent in a single phone region depends on the phone duration (e.g. a phone 100 ms long contains 10 frames with one analysis per 10 ms), whereas transition speed depends on the distance between successive phones in the MDFS.

3.2.3. ASR Phases: Preparation, Training (off-line) and Test (on-line)

The feature extraction described in the previous section (mapping a speech waveform into the MDFS) is a common speech signal pre-processing step, both for ASR training and test phases. The ASR test phase is a phase in which a trained ASR system is used for actual recognition.

The goal of ASR training is to identify and model phone regions in the MDFS corresponding to a single phone (acoustical variability modelling) as well as time intervals that a point spends in particular phone regions (time variability modelling). A considerable effort has been made to find a solution which will yield the greatest accuracy in the test phase (e.g. maximum likelihood, minimum classification error, maximum mutual information, etc.). The goal of the recognition process is to identify the sequence of phoneme regions (i.e. phonemes) in the MDFS through which the trajectory corresponding to the analysed speech signal has passed. The algorithm which is most frequently used for these purposes is the Viterbi algorithm.

Significant variations of phone characteristics caused by different adjacent phones are the reason to treat phones in different contexts as separate modelling units. These modelling units are referred to as triphones. For that reason, each triphone corresponds to a different region in the MDFS. To model the trajectory of the speech signal within a single triphone, the triphone is modelled by several states (Hidden Markov Model). The number of states per modelling unit is usually 3. Each state corresponds to the sub-region of the triphone region with feature vector values closest to each other. Speech process passes through these states with certain dynamics – the time variability is modelled by state transition probabilities. In the pre-processing stage of the test phase feature vectors are extracted from the speech signal. The aim of the test phase is to find the state sequence of the Hidden Markov Model (HMM) which generates the input sequence of feature vectors with the greatest probability. In case there were no overlap between regions in the MDFS, a Markov model would not be hidden – a bijective mapping between feature vectors and states would exist i.e. once a feature vector were known, the corresponding phone would be known as well. Since this ideal situation is not realistic, the decision regarding the corresponding phone is made after a sequence of frames and phones is processed (the Viterbi algorithm).

The ASR system training is carried out off-line, using a large speech database. The database should contain speech samples with phones in as many different contexts as possible. In order to develop a speaker independent ASR system, the speech database should contain samples of speech from different speakers. This results in greater dissipation of points in a single triphone region and greater overlap between different triphone regions.

Variations in the spectrum envelope within the same phone across speakers can be reduced by forming separate models for female and male speakers and/or using special techniques of speaker dependent system adjustment.

Regions in the MDFS are modelled by probability distribution function of feature vectors as a weighted sum of Gaussian probability distribution functions, also known as Gaussian Mixture Models (GMM). Each Gaussian distribution is defined by its mean and covariance matrix. In case

feature vectors are mutually independent, the covariance matrix is diagonal i.e. its non-zero elements define the variance vector. The last presumption is usually true, leading to system complexity reduction. Parameters of Gaussian distributions are estimated in such a way that the minimum square distance between parameterised feature vectors distribution and distribution of the training set feature vectors be obtained.

The number of phones in a language is usually relatively small (several tens). As described above, basic model units are triphones (context dependent phones) instead of phones, thus the number of models can reach several tens of thousands. This results in an increase in ASR system accuracy, but also in its complexity.

3.2.4. AlfaNum ASR for Serbian and Kindred South Slavic Languages

The goal of ASR is to recognise spoken words in a speech signal, independently of the speaker, the input device, or the environment. A recognised 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 [8]. 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, Croatian and Macedonian ASR has been concentrated on four aspects that define the quality of a speech recognition technique [9]:

- *Accuracy* – WER is less than 5% for small and medium-sized vocabulary continuous ASR; it is achieved by developed acoustic modelling trained with 40 hours of speech databases; good results for large vocabulary continuous ASR in these languages 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* – extensive 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.

Due to the complexity of the problem, a system for isolated word recognition in Serbian language was developed initially. It was later upgraded into a system for connected word recognition in Serbian language. 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 compilation time [10]. The system takes

into account lexical stress (particularly vowel length), giving greater significance to stressed vowels at recognition time.

This system can be used for speech recognition in all three aforementioned languages because it is phoneme-based and because of the similarity of the phonetic inventories of these three languages. Its internal structure is shown in Fig. 3.3. No significant drop in performance for languages other than Serbian has been observed, but actual experiments will be carried out as soon as adequate ASR speech databases in Croatian and Macedonian are available.

Even state-of-the-art ASR systems cannot be successful enough if they are based on acoustic features only. In order to achieve natural dialogues in speech applications, AlfaNum 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 the design of front-end technology. The goal of SLU is to extract the meaning of recognised speech in order to identify a user's request. Dialogue Manager (DM) evaluates the SLU output in the context of the call flow specifications, which results in dynamic generation of the next dialogue turn. The DM may apply a range of strategies to control dialogue flow according to different application tasks. To provide a successful dialogue 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 dialogue applications without having to achieve perfect recognition accuracy and without dictating what a user should say.

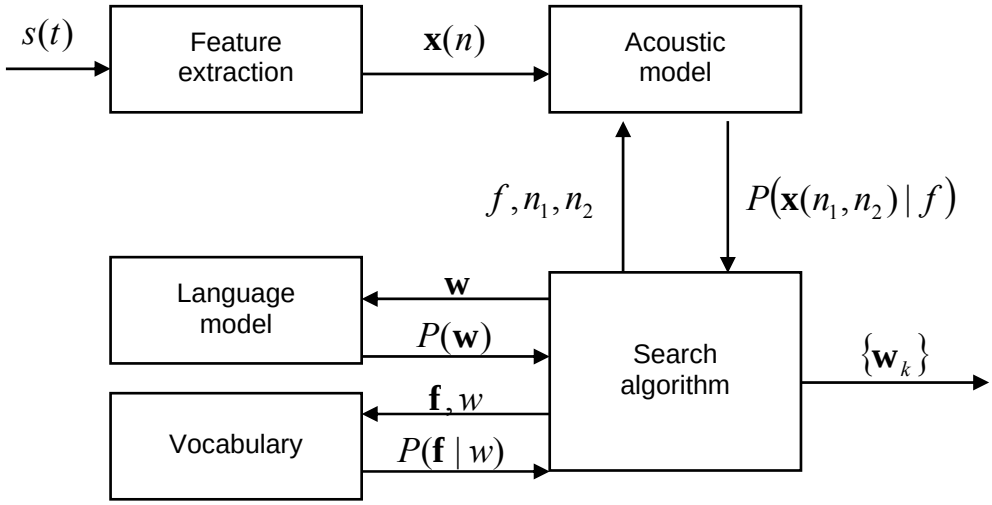

Fig 3.3. The internal structure of AlfaNum ASR

4. ASR & TTS Applications in Western Balkan countries

4.1 Aids for Persons with Disabilities Based on Speech Technologies

Many people have impaired sight or hearing, a speech disability or some other psycho-physical limitation. Although it is not clear which degree of disability should qualify a person as disabled, it is estimated that nearly 10% of the whole population are persons with disabilities (PWD). Developed countries take good care of the people with disabilities. The UN charter on equal possibilities anticipated this fact 15 years ago, making it the subject of debates and a series of meetings of an *ad-hoc* UN committee aimed at making a draft of the "All-inclusive and integrative convention about promotion and protection of rights and dignity of persons with disabilities" [11]. EU, in its turn, initiated a number of programmes aimed at the improvement of the quality of life

and the position in society of persons with disabilities. Although PWD are often limited by their incapability to use new technologies, these new technologies can be used to improve the aids for them and thus significantly improve their quality of life.

One of the recent EU initiatives is also *e-inclusion* [12], aimed at prevent a risk from being "digitally excluded", i.e. to make sure persons with disabilities are not neglected or overlooked because of their disability, lack of literature in an appropriate form or access to the Internet. At the same time, *e-inclusion* opens new possibilities for integration of marginalised groups and underdeveloped regions into the society. Information society provides all its citizens with more equal access to sources of information and opens new possibilities for employment. With the help of ICT, traditional barriers related to mobility and geographical dislocation are being broken. The "*i2010 Communication*" programme [133], aimed at an inclusive information society which should afford high quality public services and promote the quality of life, brings new challenges to the *e-inclusion* initiative. Finally, the year 2007 has been proclaimed as "the European year of equal possibilities for all", as an expression of the attempt to promote the rights and opportunities for persons with disabilities and to prevent their discrimination [144].

In accordance to European norms and standards in Serbia, the "Equal Rights for Persons with Disabilities" law has been adopted in 2006. The law proclaims equal rights for persons with disabilities, regarding education, information access as well as communication and employment. Nevertheless, many aspects of those rights will fail to be applied, unless modern technologies allow these persons to overcome their disabilities. Only then will they actualise their lawful rights to a more significant degree.

In the remaining part of this chapter it will be shown how speech technologies can help the visually impaired [15] as well as speech impaired, hearing impaired and physically impaired persons. Speech technologies can be applied as aids for persons with many types of disabilities. To the visually impaired, a machine can read books, newspaper articles from the Internet, *e-mail* or SMS messages. It can read aloud or talk instead of the speech impaired. It can receive voice commands from the physically impaired in order to control appliances such as telephone or any other household appliance connected to the system. It can convert input speech into text and thus serve as an aid to the hearing impaired.

4.1.1. The Speech Software for the Visually Impaired in WBC – anReader

Since the AlfaNum TTS was presented for the first time, its popularity has been growing within the population of the visually impaired. Computers previously had to be used with speech synthesisers designed for foreign languages.

The AlfaNum R&D team subsequently implemented appropriate interfaces (SAPI 4 and SAPI 5) which made possible the use of AlfaNum TTS with any SAPI compatible screen reader. This speech software was named anReader. Highly intelligible and reasonably natural-sounding speech motivated blind people to use computers as speech machines, helping them to communicate and offering them access to information and literature, as shown in Fig. 4.2. In three years, the number of visually impaired computer users grew up to several hundreds within the population of about 13.200 visually impaired in Serbia. Localisation for languages similar to Serbian (Croatian and Macedonian), doubled the population of anReader users in Western Balkan countries.

The popularity of AnReader gave rise to training projects for our visually impaired compatriots, especially within a chain project named "Vizija", initiated by the Faculty of Technical Sciences. Several training centres for the education of the visually impaired were created. In these centres, experienced visually impaired computer users work as instructors.

As the anReader was declared a "valuable resource for the visually impaired", the Serbian Ministry of work and social politics provided many of the visually impaired computer users with a personal

Fig. 4.2. The visually impaired can use computers as speech machines

copy. At this moment, several thousands of the visually impaired in Serbia and other WBC have the opportunity for significant improvement of the quality of their lives.

4.1.2. Audio Library for the Visually Impaired

As the population of the visually computer users remained smaller than it could have been, within the AlfaNum project a new resource for the visually impaired was developed. The audio library for the visually impaired (ABSS v1.0) was developed for the pupils of the School for the visually impaired "Veljko Ramadanović" in Zemun, the largest education centre for this part of the Serbian population [16]. The lack of literature for the blind was a significant problem for the education process in the school. Until the end of 2005, education was mostly based on books in Braille as well as low quality speech synthesisers or audio-books. All of these were costly and took a lot of space. The audio library was developed with the aim of overcoming all of the mentioned problems.

The audio library is a client-server system containing a large amount of textual data that can be accessed by a number of users over the local network or, since recently, over the Internet. The server enables centralised importing of text and additional data, as well as sorting and searches. The client application is adjusted to the needs of the visually impaired and does not depend on any screen reader. Speech synthesis, based on the anReader, is used for reading text as well as application commands. System offers the possibility of simultaneous access to the same book (all computers in the school being connected to the ABSS server). There is also a possibility of conversion of the book content to the audio format, as well as burning synthesised speech onto CDs or DVDs. In this way it can be used later by means of an ordinary CD/DVD player. A screenshot of the client application is shown in Fig. 4.3.

4.1.3. Voice Portal for the Visually Impaired

Newspapers in black print play a very significant role in our lives as a source of information and knowledge. However, black print is of little use for people with serious visual impairments without the help from somebody else. Having realised this problem, the experts from the Faculty of Technical Sciences and the AlfaNum company initiated a project named "Kontakt", and finished it by the first half of 2006. "Kontakt" is an Internet site which can be reached not only via web, but

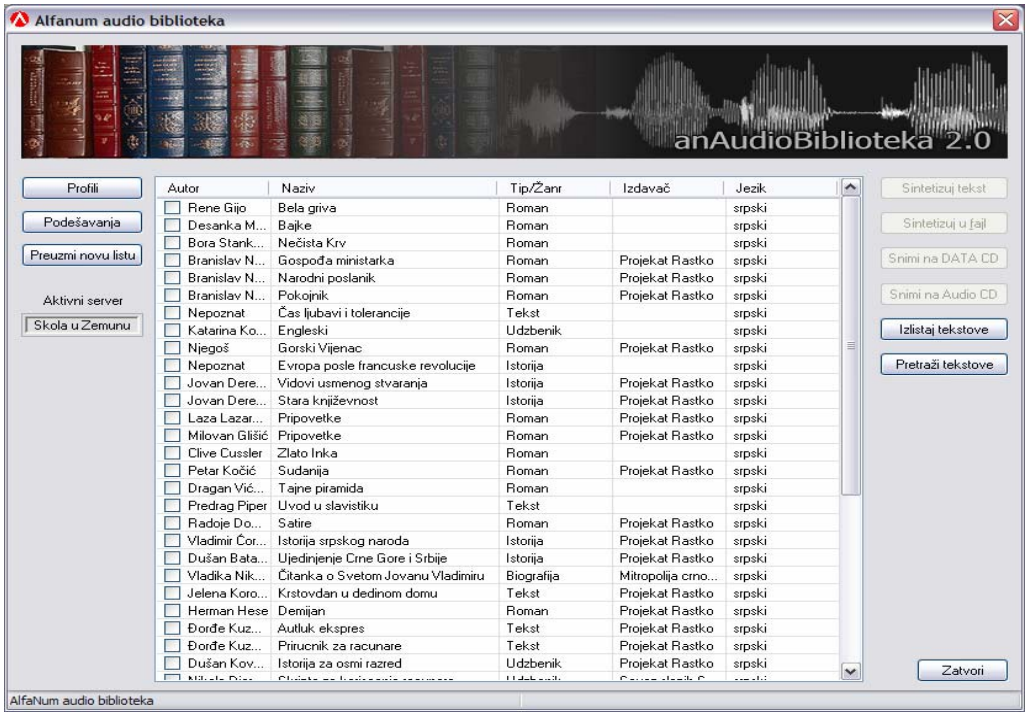

Figure 4.3. The audio library for the visually impaired – a client application screenshot

via a conventional telephone line as well, making it accessible to the visually impaired people, regardless of whether they possess a PC or not. Internet access is based on a classic approach, through a web-browser, while phone access is based on recognition of spoken commands by the system. In both cases, site contents are read out using a speech synthesiser. Currently available content includes articles from web pages of several newspapers. Additional content will include more news sources, topics related to rights of persons with disabilities, entertainment, etc. A screenshot of the application is given in Fig. 4.4.

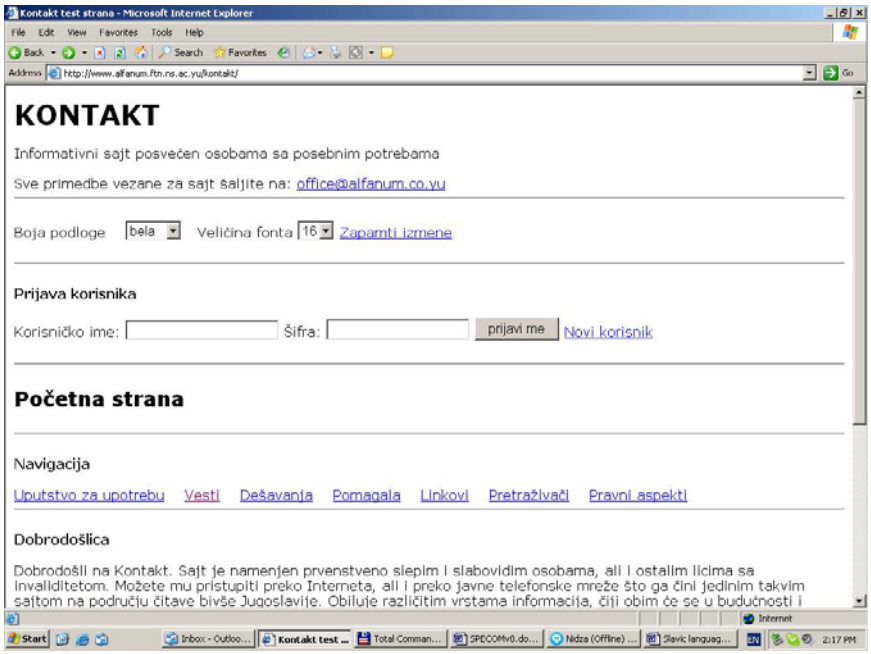

Fig. 4.4. The voice portal for the visually impaired "Kontakt" – a screenshot

The site can be found at http://www.alfanum.ftn.ns.ac.yu/kontakt. The web page is completely accessible by the visually impaired, since it is text-only, without pictures or banners, and the navigation is duly simplified. The primary design requirements were functionality and simplicity of use, rather than visual appeal, making it unconventional in comparison to a typical web site.

It can be concluded that now we have the opportunity to keep up with more developed countries and help the visually impaired in our community to improve their quality of life and be more actively involved in social life. The fact that the first steps were actually made in the biggest educational institution for the visually impaired people in Serbia, as well as in their unions, is of great importance, indicating at the same time a great interest for these innovative aids. Beside the audio library and the voice portal, the development of an SMS and e-mail reading system is under way. The development of these important resources for the visually impaired should set an example for the development of similar resources for people with other disabilities, in areas where Serbian as well as other Slavic languages are spoken.

4.2. The First IVR Services Based on ASR and TTS in WBC

ASR and TTS are applied in telephony services within IVR systems. The basis of all IVR systems developed within the AlfaNum project 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, as shown in Fig. 4.5.

A unique 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.

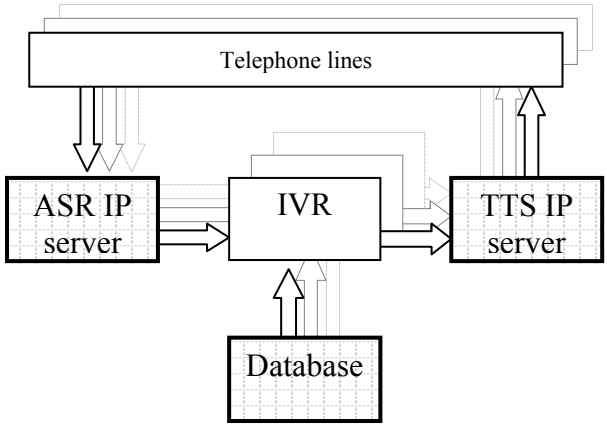

Fig. 4.5. IVR system based on ASR&TTS

4.2.1. IVR in Voice Portals

The first voice portals in South Slavic languages are the aforementioned portal for the visually impaired in Serbia (“Kontakt”) and a similar one in Croatia (“Televečernjak”). A fully automated human-machine dialogue is ASR and TTS based, as shown in Fig. 4.5. Daily information retrieval from various web sources is fully automated as well.

The experience acquired in the process of development and exploitation of these voice portals will contribute to the development of similar portals for other areas where South Slavic languages are spoken.

4.2.2. IVR in a Stock Market Application

This IVR system provides stockholders with information about their accounts at the Central Depositary Agency in Montenegro. Most of the application is standard, except for the recognition of isolated phonemes, since emission names are actually character strings such as URVF. Isolated phoneme recognition is unreliable as such; however, the number of combinations is drastically reduced due to the limited number of stock emissions that the clients actually possess. A dynamic design of grammars for a client's stock pool, without the need to restart the recogniser, was an additional requirement for the IVR system.

4.2.3. IVR in Entertainment Oriented Applications

“Sastanak” is an innovative and highly efficient phone dating IVR service. Human-machine communication enables users to execute criteria based searches through the database of registered users and to record voice messages intended for other users.

Selection of system functions (registration, search, playing and recording messages, ...) is carried out via ASR, while TTS is used for announcement of results of user actions (search results, data related to user profiles...), as well as outcomes of actual operations. Speech recognition is carried out by two independent ASR IP servers, which is sufficient for 30 telephone lines, while text-to-speech synthesis is carried out by two independent TTS IP servers.

4.3. New ASR&TTS Applications in WBC

4.3.1. Multilingual Intelligent Telephone E-mail Access

This system enables its users to access their e-mail messages by phone, and is being developed by the AlfaNum team as well as the Alpineon company from Slovenia [17].

Each day we spend more and more time reading e-mail messages. Most people also spend a lot of time travelling to work and from one business meeting to another, listening to music or radio news on their way. However, it is now possible to listen to e-mail messages via mobile phone, and thus save the time that would be spent at the beginning of the working hours. This has been made possible owing to recent improvements of the quality of text-to-speech conversion (TTS). An efficient phone e-mail access system should be able to handle a variety of e-mail messages in an intelligent way. An appropriate spam filter is also needed, as well as a simple navigation system (touch-tone and/or ASR based). The iTEMA system will offer personalization of the service regarding e-mail access as well as IVR communication, authentication and use of simple pre-defined replies. The internal structure of the system is shown in Fig. 4.6.

Multilingual nature of the iTEMA system is characterised by language recognition at the sentence level and activation of an appropriate TTS engine. This is necessary owing to the language dependency of TTS, requiring that a TTS system for the appropriate language be used for synthesis. The iTEMA system will support reading e-mail messages in several widely spoken languages such as English, German, Italian (for which off-the-shelf speech synthesisers are available), as well as most South Slavic languages such as Slovene, Serbian, Croatian and Macedonian. Reading e-mail messages in some of these languages is a special challenge due to the fact that for writing messages in the same language one can use different scripts (Cyrillic or Latin, as well as Latin without diacritics).

The architecture of the iTEMA system contains an interface towards a number of SAPI compatible TTS engines. In the middle there is a dialogue manager connected to both telephone and Web interface (see Fig. 4.6). 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 dialogue is initiated. Authentication is performed based on ANI and PIN, and followed by a personalised dialogue enabling simple and intuitive navigation through a menu system. Through this dialogue users can select messages they want to listen to, delete, or reply to using one of the pre-defined templates and recorded speech answer as an attachment.

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.

Further improvement of the iTEMA service requires continuous improvement of the quality of synthesised speech, introduction of speech synthesis in a larger number of languages as well as successful language identification between them. It is also necessary to profile models of provision of the iTEMA service through cooperation with telecom operators, mobile telephony providers and ISPs.

Since speech communication is mostly one-way (from the TTS engine to the caller), potential delays in transmission of digitalised speech are not critical. For that reason, possibilities such as VoIP communication as well as GPRS transmission of synthesised speech can also be taken into consideration.

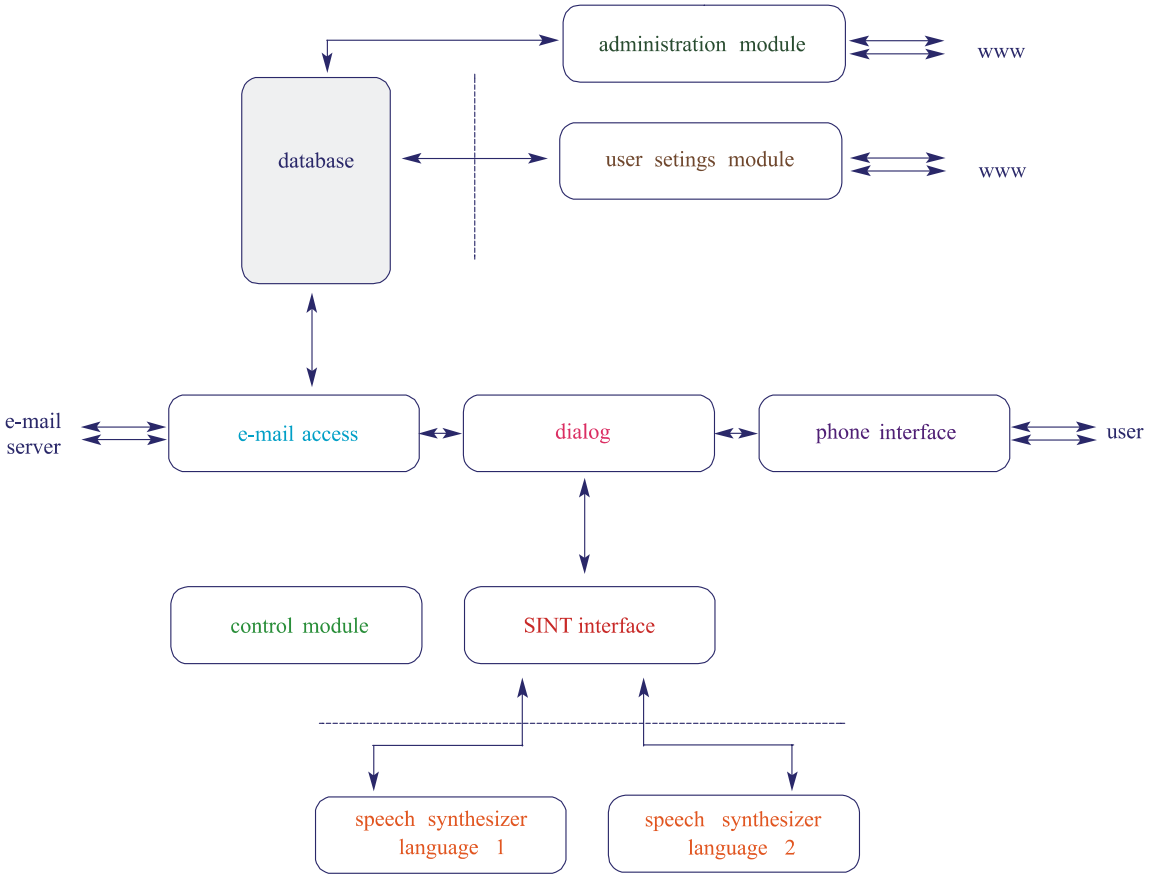

Fig. 4.6. System architecture of the iTEMA e-mail reader

4.3.2. Network Based SMS-to-Speech Conversion

SMS messages converted to speech and read out aloud are more practical for users in case their hands and/or eyes are busy (e.g. while driving), if they do not see (the visually disabled) or check their SMS messages irregularly (the elderly). This application enables that an SMS be sent to a classic (fixed) telephone with no display unit, by means of a speech machine which receives the SMS, dials the number indicated at the beginning of the message and pronounces the message by means of synthesised speech. Such a service already exists in several countries. It has been made possible for Serbian in spite of the fact that letters with diacritics in the Latin version of Serbian alphabet (š, č, ć, đ, ž and dž) are usually coded in unconventional ways in SMS messages.

4.3.3. Advertising Monitor

Reliable automation of the process of monitoring radio and TV programmes can be achieved by means of ASR. Automatic recognition of audio recordings can be carried out very accurately because there is virtually no time variability, and the acoustic variability is minimal.

The Advertising Monitor [18] is a system for audio surveillance of commercial TV and radio programmes. It tracks several channels of TV and radio stations, records them in real-time and enables automatic tracking of particular commercials, jingles, music acts etc. It can be used by media monitoring enterprises or as a service offered to individual advertisers.

The Advertising Monitor consists of several FM and TV tuners as sources of audio signals (video excluded) for a sound card which can record 10 channels simultaneously. After on-line MSGSM compression, the recordings are archived on local hard drives. Depending on the number of channels and disk sizes, up to several months of audio material can be archived. Hence the Advertising Monitor is able to carry out retroactive tracking, since complete recordings of a number of radio/TV stations are stored for a period of several months. The Advertising Monitor records selected TV/radio signals according to the never-ending tape concept. Once the storage device is nearly full, the oldest recordings are automatically erased.

A search for a set of commercials defined by a sound file and channel settings can be executed through the archived recordings. The search is carried out by independent ASR processes activated on any of an arbitrary number of machines in the system. The results of the search are written into a shared database. This database is later used for creation of daily or monthly client reports, including an optional CD with a complete recording.

It is possible to use the Advertising Monitor in combination with the WordSpotter, enabling search based on given keywords used during the programme.

4.3.4. Word Spotting System

Manual search for keywords in lengthy audio recordings can be a very troublesome and time-consuming task. For that reason, automation of such process is very important.

Automatic recognition of predefined keywords and phrases in an arbitrary audio context (word spotting) cannot be done flawlessly. Due to a variety of contexts in which a keyword can appear, errors such as a false recognition or omission can occur. By raising the recognition reliability threshold for ASR, the number of false alarms will be reduced, but the probability of a keyword being missed will increase as a consequence. For each application an acceptable compromise has to be made.

Owing to automatic tracking of audio information in electronic media, firstly in TV broadcasts, studies of media influence on the public opinion are also possible. Long-term recording of all important radio and TV channels enables a retroactive audio tracking, which can be required by law in some cases as well.

4.3.5. Automatic Telephone Inquiry

The system for automatic telephone inquiry (ATI) can be used for automatic market analyses and opinion polls, by telephone, through an automated human-machine dialogue. The entire process of the inquiry is automated: inquiry design, asking questions and noting answers, data processing as well as report design [19].

A classic telephone inquiry is time-consuming and requires a lot of concentration from the inquirer, resulting in a variable quality in presenting the questions and reliability in noting the answers. ATI saves time and assets drastically, which means that small businesses as well could afford market analyses and public opinion polls, as well as those who wish to keep track of quick changes in the market or public opinion.

Beside the ATI application itself, applications for inquiry preparation, review and verification of answers and statistical analysis as well as graphical representation of results are developed. ATI users can carry out a statistical analysis at any time, making it easier to identify issues which require a modification of the inquiry, presented questions or multiple choice answers.

5. Conclusion

The Serbian language belongs to a relatively small circle of languages in the world for which speech technologies reached a level of quality necessary for their initial wider application.

We are being witnesses of a constant increase in ASR accuracy and robustness, and in intelligibility and naturalness of TTS. In areas where Serbian, Croatian, Bosnian and Macedonian are spoken, the first applications were aimed at people with disabilities, and subsequently expanded to the CTI market, mainly as IVR systems, as shown in Table 5.1. In the meantime, innovative applications such as the Advertising Monitor and Word Spotter were created. Horizons are expanding for different applications of ASR and TTS, like Serbian language learning for foreigners or even automatic speech translation from Serbian to other languages with ASR and TTS support, directly or (more likely) using English as an intermediary.

ASR and TTS are complex technologies which will most likely never function flawlessly, but they always function with constant quality and often more efficiently than human beings – at present, a single PC can talk to tens of clients simultaneously – to understand their voice commands and to provide them with information via speech.

Current research is expanding in the direction of speaker recognition, emotional state detection, and machine translation of speech. It is extraordinarily important that each of us strives for realisation of these applications in one's own language as well.

Table 5.1. ASR & TTS applications in WBC

	Croatian	**Bosnian**	**Serbian**	**Macedonian**
ASR	Televečernjak		Sastanak Kontakt Adv. monitor Telebanking	
TTS	anReader Televečernjak	anReader	anReader Call Centres Kontakt	anReader

Table 5.2. The list of abbreviations

ABSS	Audio Library for the Visually Impaired
ANI	Automatic Number Identification
ASR	Automatic Speech Recognition
ATI	Automatic Telephone Inquiry
CASR	Continuous Speech Recognition
CD	Compact Disc
CMS	Cepstral Mean Subtraction
CPU	Central Processing Unit
CTI	Computer Telephony Integration
DM	Dialogue Manager
DVD	Digital Versatile Disc
EU	European Union
FM	Frequency Modulation
FTS	Faculty of Technical Sciences
GPRS	General Packet Radio System
GUI	Graphical User Interface
HMM	Hidden Markov Model
IP	Internet Protocol
ISP	Internet Service Provider
IVR	Interactive Voice Response
iTEMA	Intelligent Telephone E-mail Access
MDFS	Multidimensional Feature Space
MSGSM	A Microsoft version of GSM 6.10
PC	Personal Computer
PIN	Personal Identification Number
PWD	Persons With Disability
R&D	Research and Development
SAPI	Speech Application Programming Interface
SLU	Spoken Language Understanding
SMS	Short Message Service
TD-PSOLA	Time-Domain Pitch-Synchronous OverLap-Add
TTS	Text-to-Speech
UN	United Nations
UNS	University of Novi Sad
VoIP	Voice over IP
VTN	Vocal Tract Normalization
WBC	Western Balkan Countries
WER	Word Error Rate

Acknowledgment

This work was supported in part by the Ministry of Science and Environment Protection of Serbia within the Project "Development of speech technologies in Serbian and their application in 'Telekom Srbija'" (TR-6144A).

References

1. *Vlado Delić, Darko Pekar, Radovan Obradović, Milan Sečujski*, Facta Universitatis Vol. 16, No. 3, Series: Electronics and Energetics: Speech Signal Processing in ASR&TTS Algorithms, Niš, Serbia, 2003
2. *Vlado Delić, Milan Sečujski, Darko Pekar, Nikša Jakovljević, Dragiša Mišković*, International Language Technologies Conference, IS-LTC: A Review of AlfaNum Speech Technologies for Serbian, Croatian and Macedonian, Ljubljana, Slovenia, 2006
3. *Vlado Delić, Milan Sečujski, Željko Tekić,* The 17th International DAAAM'06 Symposium, Intelligent Manufacturing & Automation: A Contribution to Human-Machine Communication in Serbian, Croatian and Macedonian Language, Vienna, Austria, 2006
4. *Sandra Sovilj-Nikić, Milan Sečujski, Vlado Delić, Ivan Sovilj-Nikić*, International Symposium on Social Communication: Analysis of Different Factors Influencing Vowel Duration in Serbian Language, Santiago de Cuba, Cuba, 2007
5. *Milan Sečujski, Radovan Obradović, Darko Pekar, Ljubomir Jovanov, Vlado Delić*: Text, Speech and Dialogue: AlfaNum System for Speech Synthesis in Serbian Language, Brno, Czech Republic, 2002
6. *Milan Sečujski*, MSc thesis: Prozodijski elementi u sintezi govora na osnovu teksta na srpskom jeziku, University of Novi Sad, Serbia, 2002
7. *Milan Sečujski*, IEEE EUROCON: Obtaining Prosodic Information from Text in Serbian Language, Belgrade, Serbia, 2005
8. *Radovan Obradović, Darko Pekar, Srđan Krčo, Vlado Delić, Vojin Šenk*, EUROSPEECH: A Robust Speaker-Independent CPU-Based ASR System, Budapest, Hungary, 1999
9. *Mazin Gilbert, Jay G. Wilpon, Benjamin Stern, Giuseppe Di Fabbrizio*, IEEE Signal Processing Magazine, Vol. Sept. 2005: Intelligent Virtual Agents for Contact Centre Automation, 2005
10. *Darko Pekar, Radovan Obradović, Vlado Delić*, DOGS: AlfaNumCASR – a System for Continuous Speech Recognition, Bečej, Serbia, 2002
11. http://www.un.org/esa/socdev/enable/rights/adhoccom.htm
12. http://europa.eu.int/information_society/soccul/eincl/index_en.htm
13. http://europa.eu.int/information_society/industry/comms/index_en.htm
14. http://ec.europa.eu/employment_social/equality2007/index_en.htm
15. *Vlado Delić, Nataša Vujnović, Milan Sečujski*, IEEE EUROCON: Speech-Enabled Computers as a Tool for Serbian Speaking Blind Persons, Belgrade, Serbia, 2005
16. *Dragiša Mišković, Nataša Vujnović, Milan Sečujski, Vlado Delić*, ETRAN: Audio biblioteka za slepe i slabovide osobe kao vid primene TTS tehnologije, Budva, Montenegro, 2005
17. *Jerneja Žganec Gros, Vlado Delić, Darko Pekar, Milan Sečujski, Aleš Mihelič*, International Language Technologies Conference, IS-LTC: The iTEMA E-Mail Reader, Ljubljana, Slovenia, 2006
18. *Darko Pekar, Goran Kočiš, Robert Vuković, Stevan Molerov,* TELFOR: AlfaNum Advertising Monitor, Belgrade, Serbia, 2006
19. *Vlado Delić, Nataša Vujnović, Branislava Kostić*, DOGS: Mogućnosti automatizacije telefonskih anketa pomoću govornih tehnologija, Vršac, Serbia, 2006
20. http://www.alfanum.ftn.ns.ac.yu
21. http://www.alfanum-global.com