Slovak TTS - From Rule Based To Unit Selection

Rusko Milan, Trnka Marian and Darjaa Sakhia

Institute of Informatics of the Slovak Academy of Sciences, Bratislava, Slovakia
{milan.rusko, trnka, utrrsach}@savba.sk

Abstract

Different types of synthesizers that have been developed at the Department of Speech Synthesis and Analysis of the Institute of Informatics of the Slovak Academy of Sciences from 1990 up till now are described.

The rule based synthesizer - Kempelen O.1 developed in 1990 was a memory-footprint optimized system with PC-speaker and parallel port outputs using a unique method of signal compression preserving transients and synthesizing stable parts of phonemes by repetition of the same microsegment (pitch period).

The Kempelen 1.x engine was based on concatenation of pre-recorded diphones with signal post-processing for intonation and rhythmical contours implementation. Some interesting features were added for commercial applications, such as multilinguality, singing voice synthesis and illustrative sounds (acousticons). These synthesizers have been used in several professional applications, such as voice operated information systems, interactive voice response systems and teleservices of the Slovak telephone operators as well as in special tools for visually handicapped people.

The Kempelen 2.x synthesizer is based on unit-selection. The speech synthesis database design is described in the paper as well as the experience resulting from the design and testing of Kempelen 2.0. A new approach in Kempelen 2.1 uses pre-selection of element-candidates based on a phonological analysis of the orthoepic transcription of the text. It is aimed at elimination of eventual concatenation in problematic areas of speech signal and on selection of candidate elements according to the phonetical context. Acoustical aspects are taken into account in the second run of the selection process.

Pretvarjanje besedila v govor za slovaščino – od sintetizatorjev na osnovi pravil do sintetizatorjev, ki temeljijo na izbiri govornih enot

Opisane so različne vrste sintetizatorjev, ki so jih od leta 1990 do danes razvili na Oddelku za sintezo in analizo govora Inštituta za informatiko Slovaške akademije znanosti. Sintetizator na osnovi pravil, Kempelen 0.1, ki so ga razvili 1990, je bil sistem za osebni računalnik, optimiziran na čim manjšo pomnilniško zasedbo, uporabljal je zvočnik osebnega računalnika in izhode na paralelnih izhodnih vratih, pri tem je uporabljal lastno metodo stiskanja signala, tako da je ohranjal prehodne in sintetiziral stabilne dele fonemov s ponavljanjem istega mikrosegmenta (osnovne periode). Kempelen 1.x je temeljil na združevanju vnaprej posnetih difonov z naknadno obdelavo signala za oblikovanje intonacije in izvedbo ritmičnih vzorcev. Za komercialne aplikacije so dodali nekatere zanimive lastnosti, kot npr. večjezičnost, sintezo pojočega glasu in ilustrativnih zvokov (akustičnih ikon). Te sintetizatorje so uporabili v več različnih (opomba: raje črtamo ali pa pustimo izraz profesionalnih) aplikacijah, kot so npr. govorno voden informacijski sistem, interaktivni govorni odzivniki in telekomunikacijske storitve za slovaške telefonske operaterje, kot tudi v posebnih orodjih za slepe in slabovidne. Sintetizator Kempelen 2.x temelji na izbiri osnovnih govornih enot. V prispevku sta predstavljeni zasnova podatkovne baze za sintezo govora in izkušnja načrtovanja in testiranja Kempelena 2.0. Nov pristop Kempelena 2.1 uporablja vnaprejšnje izbiranje kandidatov za osnovne govorne enote na podlagi fonološke analize pravorečne transkripcije besedila. Cilj tega je preprečevanje združevanja osnovnih govornih enot na problematičnih delih govornega signala in izbor kandidatov za osnovne govorne enote na podlagi fonetičnega konteksta. V drugem delu procesa izbire se upoštevajo še akustični vidiki.

1. Introduction

Early experiments with speech synthesis in Slovakia were made on RPP 16 mainframe computer developed in eighties at the Institute of Technical Cybernetics (which was later renamed to Institute of Informatics). The fist hardware formant synthesizer was built at the same institute in 1987. It was developed using a PC (IBM compatible PC PRAVEC, made in Bulgaria). The quality of the synthesized speech was not bad, but the hardware synthesizer board was expensive and the operation was not user-friendly. In that time a Department of speech analysis and synthesis was founded and led by a distinguished Slovak phonetician, Prof. Ábel Kráľ. His phonetic knowledge in combination with programming capabilities and signal processing skills of engineers from this department gave a birth to the first generation of software synthesizers in Slovakia.

2. Rule based synthesizer – intelligible, but robotic

The development of the first generation TTS - Kempelen 0.1 speech synthesizer – started in 1989. The early PCs, equipped with two floppy disks and no hard disk, had 512 kB of operational memory, so the engine of our phoneme-based concatenative synthesizer was designed to require only 80 kB of operational memory for code and additional 80 kB was needed for the data. To keep the memory footprint as small as possible a unique method of signal compression was used. The stable parts of voiced phonemes were synthesized by repetition of the same microsegment (pitch period). Some unvoiced consonants and transients were kept uncompressed.

The synthesis process of the voiced phonemes merely consisted of concatenating the phoneme transients (the beginning and the ending segment) and the looped central “steady” part of the phoneme.

The full set of the Slovak transients was categorized into several classes and only one transient was chosen to represent the entire class in the database of elements. The transient also served as a joint with the neighboring phoneme. For better naturalness some of the problematic phonemes were stored as a whole.

With no soundcard available the PC-speaker and the parallel port equipped with simple resistor D/A converter were used as outputs.

In spite of the fact, that the repetition of central microsegment made the sound of the synthesizer considerably robotic, it was well understandable. According to the opinion of the users from the Slovak Union of Blind, who tested it, the generated speech quality was much better than that of the Czech speech synthesizer of the EUREKA computer that they had in use.

Kempelen O.1 was monotonous in its basic configuration. However the fact, that it had its samples stored in a form of pitch-periods (microsegments), made it relatively easy to manipulate the melodic and rhythmical contours. Simple deletion of the last samples of the period was used to shorten the period and zero padding was used to lengthen it. The first experiments with singing voice synthesis were accomplished. As the pitch shifts were realized mainly on vowels and voiced consonants with high degree of periodicity, the voice sounded a bit like sung by two people – one singing vowels and second one singing consonants.

3. Diphone synthesizer – versatile, but still a bit unnatural

The research on diphone synthesis and development of a concatenative speech synthesizer started in Slovakia approximately in the year 1994. It brought a synthetic speech of better comprehensibility and higher naturalness, together with elaborated interface that made this generation of synthesizers suitable even for professional telecommunication applications.

3.1. The diphone concatenative synthesizer

The second generation of Slovak TTS - Kempelen 1.x - was based on concatenation of small elements of a pre-recorded speech signal, mainly diphones. An algorithm similar to Time Domain Pitch Synchronous Overlap and Add (TD-PSOLA) (Hamon, Moulines, Charpentier, 1989) was used for concatenation. We have also developed a Linear prediction (LP) and Residual Excited Linear prediction (RELP) (Macchi et al., 1993) versions of the synthesizer.

Listening tests were carried out in order to evaluate three versions of our diphone text-to-speech system. The three synthesizers were based on linear predictive (LP) synthesis, residuum excited LP synthesis (RELP) and time-domain pitch synchronous overlap-and-add synthesis (PSOLA), respectively. All of them were in two versions – female and male voice. We tested the overall quality of voices and our aim was to reach MOS values for these synthetic speech signals. (Cernak 2005)

All the ten decades of Test words for Slovak audiometry (Bargár et. al. 1986) were synthesized by all the synthesizers and played from the PC to the test participant via Sennheiser HMD 25 closed-system headphones in laboratory conditions.

The subjects taking part in listening tests belonged to the normal PC using population, with the provisos that:

a) they have not been directly involved in the work connected with assessment of the performance of speech synthesizers, or in related work;

b) they have not participated in any subjective test whatever for at least the previous six months and not in any listening-opinion test for at least one year;

c) they have never heard the same word lists before.

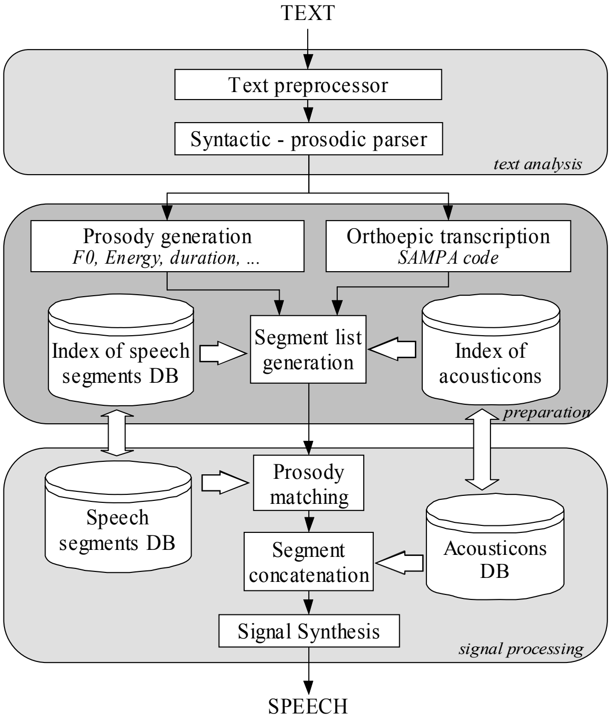

Figure 1. Schematic diagram of the Kempelen 1.x diphone concatenative synthesizer

8 subjects (2 males and 6 females) aged from 16 to 73 took part in the experiment.

Synthesis method	Mean opinion score (MOS)
LP-female	1.60
RELP-female	3.05
PSOLA-female	3.23
LP-male	1.53
RELP-male	2.84
PSOLA-male	3.34

Table 1: Subjective evaluation of the Kempelen 1.x synthesizers

The availability of synthesizers with more voices and different quality of speech made it possible to carry out experiments on voice quality measurement and to develop a method for objective synthetic speech measurement using PESQ measure (Cernak, Rusko, 2005).

The acceptable quality of the synthesized speech made it possible to use the synthesizer generated words as first draft templates for DTW word recognizer. These experiments were promising and the recognizer with male voice templates was able to recognize a majority of the words of its 1000 words vocabulary even when it was tested by female speaker. Anyway it of course could not compete with new technology - recognizers based on statistical models.

3.1.1. Rule based pronunciation

The pronunciation was controlled by the block of orthographical-to-orthoepical conversion (grapheme to phoneme) based on a sophisticated set of rules supplemented by a pronunciation vocabulary and a list of exceptions (Darjaa, Froněková, Rusko, 1994). This elaborated unit has proven to be more reliable than our similar data driven system based on CART trees (Cernak et al., 2003).

3.1.2. New voices

It generally takes several weeks to build a new professional quality diphone voice. To get an idea how the new voice will sound, we have designed a program that interactively records a set of nonsense words uttered by the tested speaker and immediately after a 10 minutes long recording session it automatically finds the needed diphones in the signal and creates a database for a draft new voice. The timbre of the new draft voice is the same as it will be in the definitive version of the new voice, only the appearance of concatenation discontinuities and rhythmical mistakes is much higher. So one can decide if the speaker is suitable for building a new voice.

Final recordings of the new voice were then realized in a studio under a permanent supervision of linguistic expert.

The diphone database building proceeded in two steps. A draft automatic phonetic alignment using a combined DTW/Rule-based recognizer. This had later to be checked and refined manually by a human expert to achieve a fluent and relatively natural voice.

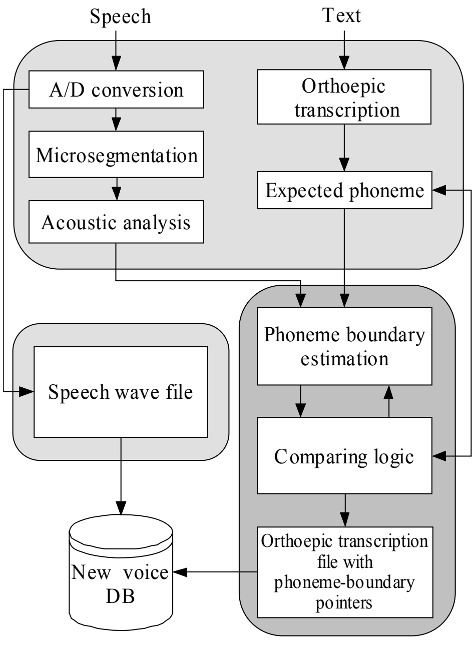

Figure 2. Diagram of automatic phonetic alignment used for generating a new voice

3.1.3. Multilinguality

One of the main trends in telecommunication services, information systems and computer speech interfaces is multilinguality. Developing English, German or French version of our synthesizer would probably not give any sense, as there are lots of high-quality synthesizers available for these languages, developed by reputable companies which have incomparably better financial and personal capacities for there development. We have however decided to make a Hungarian version of our synthesizer to broaden the rank of possible users by the Hungarian speaking fellow-citizens. We have used our synthesis engine and with a help of the students of Hungarian nationality and the employees of the Department of Hungarian language of the Comenius University in Bratislava we have defined rules and designed a database of synthesis elements as well as a block of pronunciation for Hungarian. As a result we have a synthesizer in two languages.

We think it would be interesting to have the source speech for synthesis recorded by one bilingual speaker in both languages, which would help to avoid timbre differences in the two languages.

3.1.4. Singing voice synthesis

Singing voice synthesizers have in general different purpose than speech synthesizers and they work on different principals. They are designed to provide enjoyable singing voice where intelligibility is not of highest importance. They may employ principals of music samplers, advanced methods of pitch processing and time stretching algorithms etc.

We decided to use the simplest and cheapest way – that is „to force the speech synthesizer to sing“. The basic formula for tempered tuning is:

$$f_{n+k} = kqf_n \qquad (1)$$

where q = 1,05946309, which is the twelfth root of two and k is the number of half-tones between f_n and f_{n+k}

It is obvious, that a direct mathematical representation of a note code does not give an acceptable pitch contour for the singing voice synthesis. Our analyses of the pitch contours of recorded songs had shown that at least several phenomena should be taken into account, such as rise and fall times of the tones, and vibrato, its depth, envelope and frequency. The introduction of these changes improved the synthesized singing significantly (Darjaa, Trnka, Rusko, 1999).

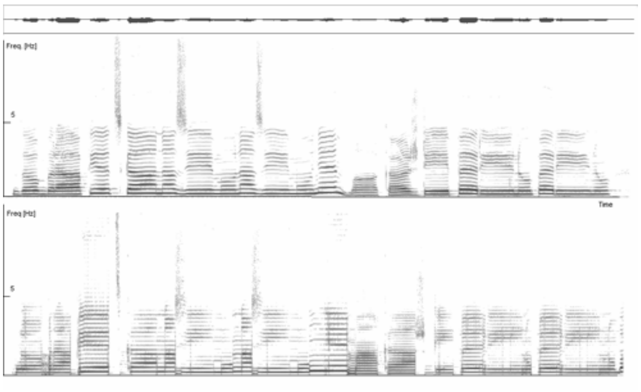

Figure 3. The spectrograms of a part of a song sung by a female singer (upper) and of a signal synthesized from the elements of speech of the same woman (lower). Only a simple rule for tone onset pitch changes and no vibrato is applied in this version.

In spite of imperfection of the solution we consider our singing voice feature to be fully functional and suitable to enrich the SMS to Voice service as a new entertaining feature.

4. The unit selection synthesizer – problems in unlimited domain

Advances in speech synthesis in the world led us to the decision that the third generation of the Kempelen synthesizers should be based on unit-selection from a speech database.

4.1. Speech database for synthesis

At the beginning of this project there was no annotated speech database available for unit-selection speech synthesis in Slovak. So it was an inevitable condition to design a professional-quality, one speaker, general purpose database for research, experiments and application building in unit-selection speech synthesis which would be extendible, but also down-scalable (e.g. for limited domain experiments).

4.1.1. Recording the database

The database consists of recordings of one male, non-professional speaker, experienced in speech processing. The recording took place in an unechoic room of a professional studio specialized to speech recording (radio commercials, dubbing etc.). The sessions lasted typically about two hours and were realized in irregular intervals from one week to one month. A Neumann U 87 cardioid condenser microphone with Focusrite Trackmaster pre-amplifier and a hard disk recording system equipped with AARK 20/20+ sound board was used in the sessions. The sampling frequency was 44.1 kHz and resolution was 16 bit.

4.1.2. Choice of the source text material, database content

In spite of the fact, that we plan to extend the speech database in future, the initial elementary structure of the database had to be clearly defined first. Our ambition was to design a general-purpose database being at the same time suitable for experiments in limited domain synthesis. The other contradictive requirement for the database was not to be too big, but to be representative enough from the phonetical, phonological, and other points of view. Therefore we decided to design the database as a combination of several more or less independent parts:

4.1.3. Phonetically rich sentences

- Set of words covering all Slovak diphones
- Sentences covering intonation phenomena
- Spontaneous speech record (General topic story, Application oriented story)
- Set of prompted application-oriented phrases and embedded application commands
- Numerals

4.1.4. Database annotation

The annotation consists of several levels of information. In the case of need new levels of annotation can be added. Annotation techniques and choice of annotation levels belong to the subjects of research to be accomplished on this database, therefore the mentioned annotation levels serve only as a reference, as an initial annotation to start with.

4.1.5. Annotation levels

There are two text annotation levels:

- orthographic text
- orthoepic text (in SAMPA)
- Signal annotation levels are the following:
- microsegmental information – pointers to single pitch periods
- phoneme boundaries information
- diphone boundaries information
- syllable boundaries information
- whole words and phrases information

Suprasegmental annotation level consists of:

- melody contour information - smoothed f0 value, intonation phrase boundaries
- accent information

4.1.6. Automatic annotation

Automatic annotation consists of orthographical to orthoepical conversion, microsegmentation – pitch marking and segmentation to diphones

4.1.7. Orthographical to orthoepical conversion

The text in the orthographic form was transcribed to the orthoepic form by the block of pronunciation developed for earlier versions of our synthesizers [4]. The orthoepic text generated automatically was then manually checked and corrected by an expert with a degree in linguistics.

4.1.8. Microsegmentation – pitch marking

Microsegmentation – pitch period boundaries detection was accomplished by a rule based routine, which works well on a clean studio-quality full range speech signal (Darjaa, Rusko, 1997).

With a help of an orthoepically transcribed text and a rule-based phoneme recognizer based on pitch synchronous analysis (Darjaa, Kráľ, Rusko, 1993) correspondence of every microsegment to a particular phoneme was recognized and its boundaries were estimated.

4.1.9. Segmentation to diphones

One of the levels of annotation splits the speech signal into parts (elements - mainly diphones) which inventory matches to the set of the elements used in our diphone synthesizer Kempelen 1.4. The boundaries of the elements which the signal was generated from are known for the synthesized signal. Making use of the fact that we have a synthesizer with the voice of the same speaker, we applied a DTW algorithm in one of our phonetic alignment algorithms to automatically label element (diphone) boundaries in the recorded signal by mapping the labels from the synthetic speech to the recorded one.

5. Experimental synthesizer

We used Baum-Welch training to build complete ASR acoustic models from a part of the database. The HMM recognizer with these models was then used to label data. The whole labeling was realized in FestVox framework, where Carnegie Mellon University's SphinxTrain and Sphinx speech recognition system are used (Huang et. al., 1993).

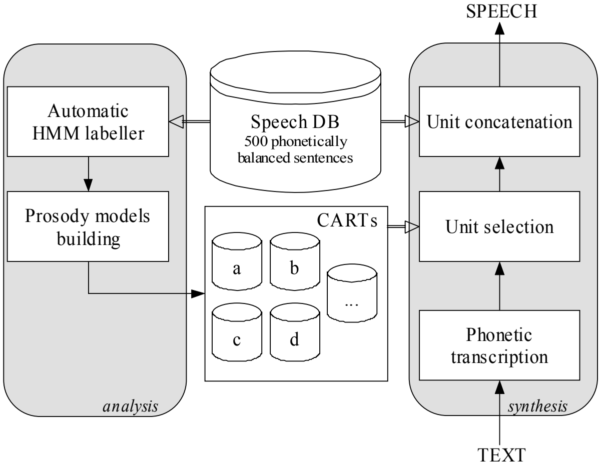

Figure 4. Schematic diagram of the experimental synthesizer

We used 500 phonetically balanced utterances for training and labeling. An experimental Slovak corpus-based speech synthesizer was built using the labeled data. The approach uses a technique of automatic clustering of similar units to build a CART for each phoneme with questions from NLP block in its nodes. We used duration model with average durations of phones. Then we applied simple multiplicative factors for the phones in phrase final and phrase initial positions. (Fig. 4)

5.1. Recent version - Kempelen 2.1 synthesizer

Recent version of the synthesizer, Kempelen 2.1. does not use any third-party components. It fully relies on our own annotation method, pre-selection of elements and unit selection algorithm. (Fig. 5)

5.1.1. Unit preselection

Generally speaking a syllable was taken for a basic element in our synthesizer. However the phoneme boundaries are annotated in the database and in the case of need smaller units than syllables are chosen for synthesis as well.

The aim of our preselection is to avoid using improper joint points just by employing phonological knowledge. The phonetical context is checked carefully. If an element of the required context is not available, the database is searched for an element with a context belonging to the same phonetic category as the desired one. Different phonetic contexts are allowable only in the worst case, as they usually cause audible disfluences in timbre at the concatenation point. This approach is similar to that of Taylor and Black (1999).

In some of the triphones an extremely strong coarticulation at the central phoneme can be expected and it is very unlikely for the automatic annotation program to find the boundaries of such a phoneme correctly. Therefore we have defined a list of "forbidden joint point triphones" which can be split only if no other solution is possible. Typical representatives are VCV combinations with sonorants l, L, r, j, or fricative h (in SAMPA) as their central phoneme. The preselection takes into account also a syllable position – word initial, word center, word final and sentence final.

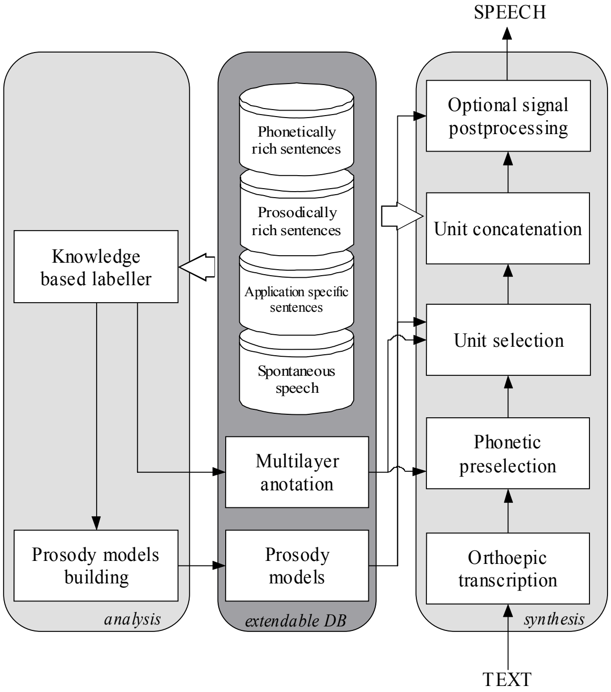

Figure 5. Schematic diagram of the Kempelen 2.1 synthesizer

6. Discussion

The rule "the more data the better" does not seem to hold in unit selection speech synthesis in all cases. Especially in the initial phases of the research it is necessary to have a smaller speech database recorded with relatively stable vocal effort and flat prosody. Only after the unit selection algorithm is well tuned, it is advisable to enrich the database with speech data covering intonation phenomena of expressive speech and rhythmically and phonetically problematic spontaneous speech. In this point it becomes even more important to have a reliable annotation method. We think that automatic HMM phoneme labelers should always be checked for typical errors and supplemented by knowledge based corrective algorithms. In our approach to phonetic alignment we strongly rely on secure identification of anchor points in the speech signal which are of three main categories:

- Vowels (high energy, periodicity, sharp formant structure)
- Fricatives (noisy spectrum with high frequency components)
- Plosives (pause plus burst structure)
- Phoneme boundary finding is always based on iteration.

In our recent approach to synthesis we apply a phonological unit preselection which reduces the universality and openness of the classical unit selection approach, but it excludes the most significant concatenative problems in advance, before the calculation of concatenative and unit costs has even started.

7. Conclusion

The paper presents a brief survey of research and development in speech synthesis in Slovakia.

The first generation of Kempelen speech synthesizers has proven a capability of software synthesizers to produce intelligible speech under very low computational expense. The know-how, from the first generation represented appropriate initial conditions for building a second generation with better performance, intelligibility, stability and versatility.

These reliable synthesizers have been integrated into voice services of all the three Slovak telephone operators. They are also in use by some members of the Slovak Union of Blind and Visually Impaired for screen reading and some of special tools for visually impaired are delivered with Kempelen 1.6. synthesizer too.

The Kempelen 2.1 synthesizer is the most recent of our products at the moment, which is still under development. We find it to be a promising successor of the popular Kempelen 1.x synthesizers and we hope, that the companies in Slovakia will discover the advantages of the unit-selection synthesis approach soon.

8. Acknowledgements

This work has been funded by the Ministry of Education of the Slovak Republic, task number 2003 SP 20 028 01 03 and Slovak Agency for Science, VEGA, grant No. 2/2087/22.

9. References

Hamon, C., Moulines, E., Charpentier, F., 1989. A diphone synthesis system based on time-domain prosodic manipulations of speech. *Proc. Int. Conf. Acoustics, Speech and Signal Processing*, 238.

Macchi, M., Altom, M.J., Kahn, D., Singhal, S., Spiegel, M.F. 1993. Intelligibility as a Function of Speech Coding Method for Template-Based Speech Synthesis. *Proceedings of Eurospeech 93*, Berlin, 893-896

Cernak, M., Rusko, M. 2005. An evaluation of a synthetic speech using the PESQ measure. *Proceedings of Forum Acusticum 2005*, Budapest

Darjaa, S., Franěková, L., Rusko, M. 1994. Conversion and Synthesis of the Slovak Speech. (in Slovak), *Jazykovedný časopis, 45,(Bratislava) 1994, No. 1*, 31-34.

Cernak, M., Rusko M., Trnka, M., Darjaa, S., 2003. Data-Driven Versus Knowledge-Based Approaches to Orthoepic Transcription in Slovak. *Proceedings of ICETA 2003, Kosice (Slovak Republic)*, 95-97.

Darjaa, S., Trnka, M., Rusko, M., 1999. The Application of Text-to-Speech System in Slovak to Singing Voice Synthesis. *Proceedings of SPECOM'99, Moscow,* 162-165.

Dutoit, T., 1997. An Introduction to Text-To-Speech Synthesis, *volume 3 of Text, Speech and Language Technology.* Kluwer Academic Publishers, The Netherlands.

Huang, X.D., et. al., 1993. The SPHINX-II Speech Recognition System: An Overview, *Computer Speech and Language (1993)*, 137-148.

Lee, K.F., 1989. *Automatic Speech Recognition: The Development of the SPHINX SYSTEM*, Kluwer Academic Publishers, Boston, 1989.

Rusko, M., 2000. Definition of corpus, scripts and standards for Fixed Networks – Slovak. *SpeechDat-E deliverable-ED1.2.3, http://www.fee.vutbr.cz/SPEECHDAT-E*

Kráľ, A. 1996. *Pravidlá slovenskej výslovnosti*, Slovenské pedagogické nakladateľstvo Bratislava, 163 – 200

Rusko, M., Darjaa S., Trnka M., Petriska M., 2000. SpeechDat-E, the First Slovak professional-quality telephone speech database, *In: Research Advances in Cybernetics.*, ELFA Publishing House, Košice, 187-211

Moulines, E., Charpentier, F., 1990. Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones, *Speech Communication 9*, 453-467.

Darjaa, S., Kráľ, Á., Rusko, M., 1993. Phoneme-oriented Approach to Speech Recognition in Slovak. *in D. Mehnert (Hrsg.): Elektronische Sprachsignalverarbeitung in der Rehabilitationstechnik, Berlin*, 83-89,

Darjaa, S., Rusko, M. 1997. Automatic Labelling of Speech Signal for Slovak Speech Database Building, *Proceedings of the 31st Int. Conf. ACOUSTICS-High Tatras 97,* 124-125.

Rusko, M., Darjaa, S., Trnka, M., 2001. Databases for speech recognition and synthesis in Slovak. *In: Proceedings of the conference SLOVKO - Slovenčina a čeština v počítačovom spracovaní*, Bratislava , VEDA, 88-97

Rusko, M., Darjaa, S., Trnka,M., 2002. Automatic design of the elements database for speech synthesizer in Slovak, (in Slovak), *in Proceedings of the conference Noise and vibrations in practice - Kočovce 2002*, Slovak Technical University Bratislava, 75-78

Black, A.W. and P. Taylor 1997. Automatically clustering similar units for unit selection in speech synthesis. *in Proc. of the European Conference on Speech Communication and Technology.* Rhodos, Greece.

Bargár Z., Kollár A., 1986. *Praktická audiometria.* Osveta, 159-160 [In Slovak].

Taylor P. and Black A. W., 1999. Speech Synthesis by Phonological Structure Matching, *in Proceedings of Eurospeech'99*, 623-626.