Budapest University of Technology and Economics
Department of Telecommunications and Telematics
Laboratory of Speech Technology

Hungarian text–to–speech conversion: linguistic models, algorithms and implementation

Ph.D. theses

Author:
Péter Olaszi

Scientific advisors:

Dr. Géza Németh
Budapest University of Technology and Economics
Department of Telecommunications and Telematics
Laboratory of Speech Technology

Dr. Gábor Olaszy
Hungarian Academy of Sciences
Linguistic Institute
Kempelen Farkas Speech Research Laboratory

2002.

1 Introduction

Electronic speech synthesis has a tradition of some twenty years in Hungary. The first electronic speech production device was created by the concept of Gábor Olaszy at the Phonetics Laboratory of the Linguistic Institute, Hungarian Academy of Sciences (LI, HAS) between 1980–82. The system named Hungarovox was realized for linguistic and phonetic research on a DEC PDP–11/34 computer using the OVE III formant synthesizer. Formant synthesis is based on the study of the excitation and representative frequencies of the speech signal.

Parallel to the works at the Linguistic Institute, the Telecommunication Electronic Institute of the Budapest University of Technology (TEI, BUT) also hosted research related to speech production and telecommunication technology. The Scriptovox speech production system – based on the MEA 8000 formant synthesizer – was realized in 1983 in the teamwork of the two institutions.

Since then the collaboration between the LI and the TEI became persistent. The increase of the computing capacity of the personal computers allowed real-time, unconstrained vocabulary text–to–speech conversion using formant synthesis technology. Thus, in 1986, the development of the Multivox multilingual speech synthesis system began. The research program was lead by Gábor Olaszy and Géza Németh. The first versions of the program worked on the DOS operating system, and the speech signal was produced by the Philips PCF 8200 formant synthesizer chip connected to the computer. In later releases of the speech synthesizer the hardware emulation of the chip permitted solely software realization. The experiences of using formant synthesis technology for Hungarian are summarized in the comprehensive book of Gábor Olaszy.

Since the middle of the nineties both abroad and in Hungary research was focused on waveform concatenation technology. Due to the constant extension in the computing and storage capacity of personal computers, real-time speech synthesis based on the concatenation of waveform-elements of human speech became a realistic aim. In the basic form of this technology, connections of two speech sounds – so called diphone elements – are stored in a speech database. Appropriate positioning of the cutting points ensures natural character for the concatenated speech signal. The Department of Telecommunications and Telematics (DTT) – which can be considered as the successor of TEI – gave place for the development of the Profivox text–to–speech converter since 1998. This system builds the speech signal from a speech database of 1100 diphone elements.

The middle of the 90's also brought a significant progress in the speech synthesis related research of prosody. In the speech synthesis system prosody is considered as the change of pitch, accentuation, durations, pauses, rhythm and the intensity of the speech. Descriptive, theoretical surveys of the Hungarian prosody can be found in the works of Mihály Bródy, Katalin É. Kiss, Iván Fónagy, László Hunyadi, László Kálmán and László Varga. The relation of prosody and speech perception was studied primarily by Mária Gósy, Csaba Pléh and Klára Vicsi. Prosody in connection with artificial speech production was described in the works of Gábor Olaszy and Ilona Koutny. In the field of written language technology the work of Gábor Prószéky should be mentioned.

Other artificial speech production related research in Hungary took place in the Central Physics Research Institute, HAS from the beginning of 80's with the direction of András Arató. Their formant synthesizer was developed mostly with the needs of blind people in mind. József Király created a text–to–speech software for DOS without prosody in 1988. Recently Mindmaker, a Hungarian company produced a text–to–speech system primarily for the market, putting less emphasis on the research side.

I joined the ongoing research of the Laboratory of Speech Technology, DTT BUT in 1992 as an undergraduate student. Initially I aimed at getting acquainted with the formant synthesis technology; later I implemented applications. During my task of creating an algorithm to convert English text to phonemes, I came into contact with Péter Siptár (LI HAS), from whom I received valuable orientation in phonological questions, even later on. At that time my interest turned towards natural language parsing and generation. As a result of cooperation with Ilona Koutny (Department of General and Applied Linguistics, Eötvös Lóránd University of Sciences) and Gábor Prószéky (MorphoLogic, Ltd.), in my diploma thesis I presented an automatic prosody tagging algorithm based on the syntagmatic parsing of simple Hungarian sentences. Since 1997 I continued my research and development activity as a Ph.D. student. Following the international trends, our group focused on waveform-concatenation technology. Parallel to studying signal processing algorithms, my attention gradually turned to natural language processing.

I assign my future aims in the realization of multilingual text–to–speech synthesis, integration of different synthesis technologies (formant, linear predictive coding, waveform concatenation, and others), automatic prosody extraction from large amount of spontaneous human speech, researching very high fidelity parametric synthesis, and exhaustive syntactic parsing of Hungarian sentences.

2 The aim of the dissertation

The aim of the dissertation is to address the following problems: issues regarding Hungarian text–to–speech conversion, integration of existing and creation of new linguistic and signal processing models, the implementation of the theoretical results, examination of functioning systems, and to interpret the results. Therefore I set forth the following aims:

- Design of a Hungarian text–to–speech system based on waveform concatenation technology.
- Creation of algorithms to convert numerals to letters in Hungarian texts for the purpose of speech synthesis.
- Using the available models, realization of a prosody tagging algorithm that can be used both with the speech synthesizer and as a standalone application.
- Adaptation of the multi-layer model of classical generative phonology for speech synthesis.
- Construction of a universal speech sound representation system for multilingual text–to–speech synthesis on phonetic-phonological basis.
- Definition of the structure of the speech database containing diphones, triphones and larger elements and creation of an algorithm for optimal building of the speech sound sequence using the elements of the above database.
- Creation of a pictogram–to–speech conversion system used in the communication and rehabilitation of people with disabilities, and a Hungarian morphological and syntactic generating system for constrained vocabulary and grammar.

3 Method of research

During my work, theoretical research aims were frequently motivated by practical applications. In other cases, on the contrary, research results induced new applications. I carried out my research activity partly independently, partly in cooperation with my scientific advisors and colleagues.

One of the methods used was to create new models and integrate them with existing ones. I follow this approach, for instance, when designing the text–to–

speech system architecture, constructing the prosody tagging algorithm or implementing the morphological and syntactic generator.

During the design of the number–to–letter conversion algorithm I used the method of corpus analysis to search large amount of electronic mails for structures containing numbers.

At the development of the prosodic rule system we used the method of analysis by synthesis: first we formed the prototype of the rule system, then we marked the points of further development through perceptual tests. We improved the model in an iterative process.

I deduced the problem of assembling the speech sound sequence from diphones and larger elements to an algorithm of graph theory, thus achieving the algorithm analytically.

The text–to–speech system was tested both in laboratory and real-life conditions. We submitted the system to stress tests for verifying its stability. In industrial conditions we studied the text–to–speech converter as part of an e-mail reading system, and registered customer feedback continually.

4 New scientific results and their technological realization

New scientific results can be summarized as follows:

1. I designed a Hungarian text–to–speech converter using waveform concatenation technology, and realized the system as a speech synthesis engine.
2. I studied the pronunciation of numbers in Hungarian texts, and created a method to convert numbers to letters in different contexts for speech synthesis.
3. I realized an automatic prosody tagging method. I implemented both the prototype, handling the surface phenomena, and the complex version, exploiting the results of morphological and syntagmatic analysis.
4. I studied the multi-layer model of the classical generative phonology. I formalized the model for speech synthesis applications.
5. I constructed a language independent, universal speech sound representation scheme for artificial speech production. Herewith I established the possibility of constructing flexible multilingual speech synthesis databases.

6. I defined the structure of speech databases containing diphones, triphones and larger elements. I deduced the problem of assembling the speech sound sequence to a well-known algorithm of graph theory.
7. I formed the structure of the Blissvox speaking communication aid, used in the rehabilitation of speech-impaired people. I implemented the Mondgen morphological and syntactic generator for constrained vocabulary and grammar.

In the following I describe the above theses. I also address questions that did not bring results to form a thesis, but I consider them important from the research point of view.

1. I designed a Hungarian text–to–speech converter using waveform concatenation technology, and realized the system as a speech synthesis engine.

I started my work from the Multivox 4 text–to–speech converter, which is based on formant synthesis technology. I studied the version running on the DOS and Windows 3.1 operating systems and implemented the program under Windows NT. I prepared the documentation of the modified system, and starting from this documentation I designed Multivox 5, the system using waveform concatenation technology [R2], [R3].

In the new model of the text–to–speech converter I strictly separated the phases of analysis and synthesis. The input of the analysis phase is a sequence of *graphemes*, that is, letters, numbers and other characters. The output of the synthesis phase is the speech signal. The two phases are connected by a parameter sequence holding both segmental speech sounds and suprasegmental prosodic information. Using the terminology of Olaszy and Németh, this sequence of parameters if referred to as the *prosody matrix*.

I realized that the conversion procedure of *written text* $\rightarrow$ *representation of speech sounds* $\rightarrow$ *prosody control parameters* $\rightarrow$ *speech signal* is strictly sequential, thus the best way of modeling it is to use the pipeline-method. From this standpoint I defined the five stages of the conversion process: (1) preprocessing at the level of graphemes, (2) grapheme–to–speech sound conversion, (3) operations on the speech sounds, (4) speech sound–to–waveform conversion and (5) operations on waveform level.

For tracing of the conversion process and giving points of intervention I described the states of the data flow, and defined the interfaces between the stages.

During the realization of the logical units I carried out the tasks set forth in the following.

I defined an advanced control command set that can be embedded in the input text. With the application of the commands, one can adjust, for instance, the pitch, speed, intensity of the speech, or can add prosodic information to the text. The command set is based on the Multivox 4 commands, uses commands of the DecTalk speech synthesizer and those that can be found in the Microsoft Speech Application Programming Interface (MS SAPI). The recommendations for the Extensible Markup Language (XML), Speech Synthesis Markup Language (SSML) and the specific features of Multivox 5 were also taken into account. The command set is described in the [R6] electronic report.

I created an algorithm to convert numbers to letters in Hungarian texts. I summarize my results in a separate thesis.

I implemented a prosody tagging algorithm. I realized both its prototype based upon the parsing of special words and the complex version exploiting the results of morphological and syntagmatic analysis. I summarize my results in a separate thesis.

At the level of graphemes I designed an efficient method to represent the abbreviation dictionary. I constructed a trie-based fast search algorithm for the dictionary. I described the handling of the grapheme-level rules in the [R8] electronic report.

I recognized that the realization of a multilingual text–to–speech converter demands flexible, language independent representation of the speech sounds. Therefore I constructed a universal speech sound representation. I summarize my results in a separate thesis.

I designed the structure of the pronunciation dictionary and grapheme–to–speech sound conversion rules [R11]. Trie-type organization of the dictionary guarantees efficient searching even at large dictionary sizes. From the practical point of view I underline that using the new dictionary structure, I eliminated the limit for the number of entries – being a critical point of the former system.

I designed a phonetic symbol set for the simple visualization and modification of the grapheme–to–speech sound conversion rules [R12]. In the earlier versions of the speech synthesis system, the sequence of speech sounds had to be given by the sound codes: the phonetic text thus was a string of numbers. This made the dictionary incomprehensible, and the editing of the rules inconvenient. The proposed phonetic symbol set offers a simple, feasible way for editing the pronunciation dictionary and tracing the stages of the conversion process in the speech synthesis system.

I made an attempt to design an English grapheme–to–phoneme conversion algorithm. The proposed letter–to–phoneme rules were based on phonological considerations. They took the number of syllables, prefixes and suffixes into account, and marked word stress using built-in stress patterns. In the design and implementation of the rule system I received a notable support from Péter Siptár. The tests were carried out using the NETalk pronunciation dictionary. I found that the algorithm resulted in the correct phonetic transcription for 58 percent of the words. The result was behind my expectations, and emphasized my later conclusion that practical English letter–to–phoneme conversion can only be efficiently realized by means of listing the words in a pronunciation dictionary. My conviction had been reinforced during studying the model of lexical phonology, where I found that the phonological rule system may remain hidden in the speech synthesis system, as it was enough to process only the surface forms of the words. I published my results in book [B1] and at presentation [P1].

I designed a rule system for handling and representing assimilation phenomena, such as voicing, unvoicing, merging, shortening and consonant drop. I prepared the experimental implementation of the rules.

I designed a method for the verification and correction of the structure and contents of diphone speech databases. The problem has been described in [R14].

I implemented Olaszy's algorithm to modify the specific duration and intensity of speech sounds [R13]. I solved the problem of superimposing suprasegmental clause melody and intonation curves on the segments of speech sounds. I implemented the diphone element concatenation algorithm and the algorithms to modify the pitch, duration and intensity of the speech sounds. I introduced the concept of *intonation profile* to represent the pitch changes at the level of pitch periods. We published our results in [B2].

I implemented the conversion algorithms between the 16-bit PCM and the 8-bit A-law and μ-law coding conventions. I designed a function library to play back the synthesized speech signal.

I implemented the text–to–speech converter system in the form of a speech synthesis engine (which cannot be run in itself; it should be built into applications) [R4], [R5]. The name of the speech synthesis engine is Multivox 5. The name of the text–to–speech server application is Profivox.

I carried out speed and stability tests on the system. The tests were performed under Windows NT running on a Pentium II 266 MHz computer with 64 megabytes of memory. In one group of the experiments I examined speed performance. I found that the content of an 800 000-word dictionary was proc-

essed at a speed of 240 words/second. For testing the stability, we added binary and text files to the input of the system, 1.4 GB of data in total. The program was able process all the incoming data without any difficulty.

The speech synthesis engine built into the Microsoft Speech API was checked according to the stress test requirements set forth in the MS SAPI documentation. Sentences, words and meaningless character sequences were given to the input. Some problems have been occurring during the test were eliminated gradually. We also carried out numerous perception tests. The subjective results were used in the iterative development process of the linguistic models.

The speech synthesis engine was built into an e-mail reading system. Our experimental system has been operating as the *Mailmondó* service of the Westel Mobile Communication Company since December 1999. The stability of the system is best characterized by the fact that, according to the service provider, due to software problem the system has not broken down since the installation.

Among the results introduced in the thesis I underline the importance of defining a modular and clear structure of the text–to–speech converter. By means of waveform concatenation technology and effective application of prosody I implemented a natural sounding Hungarian speech synthesis system.

Considering potential industrial applications, technical details and algorithms are mostly published in internal reports, being available only for the developers of the system. The complete text–to–speech converter is introduced in [J3]. The experimental version of the email reader system is described in [B3]; the version operating at the telephone service provider is published in [J4]. A limited vocabulary speech message composer based on the text–to–speech converter is explained in [C3] and [C4].

The architecture of the speech synthesis system described above forms part of the Speech Information System lecture at the Faculty of Electrical Engineering and Informatics of the Budapest University of Technology and Economics (BUTE).

2. I studied the pronunciation of numbers in Hungarian texts, and created a method to convert numbers to letters in different contexts for speech synthesis.

I studied the pronunciation of numerals in Hungarian texts. I surveyed the number elements constituting Hungarian cardinals, fractions and ordinals both written with digits and letters. I designed a method to convert numbers to letters.

I studied the role of punctuation marks in ordinals, decimal fractions and dates. I paid special attention to the dot as a thousand separator and sentence terminator, and the comma indicating enumeration and clause boundary.

I studied the writing of telephone numbers. I extracted telephone number formats from 20 000 e-mails by automated method, finding 70 different telephone number formats in the provided sample. I designed a technique to identify the most frequent formats.

I studied the writing of date and time. I designed a method to convert the most frequent date and time formats to letters.

I devoted special attention to the affixes modifying the pronunciation of numbers. Among others I examined the derivatives of ordinals and dates, the writing of vulgar fractions and the modifying effect of nominal case suffixes, such as stem alternations. I designed an algorithm to convert exponential numbers to letters.

Based on the works of Olaszy and Németh I studied the prosodic issues of reading numbers elements. I implemented their algorithm of assigning prosodic information to numbers.

Note, that Olaszy designed a set of number elements and speech database for reading Hungarian, German and English cardinals. This system was designed with linguistic considerations, such as the coarticulation effects in mind. In my case coarticulation was not an issue, as I aimed at developing a number–to–letter system. Postlexical effects are handled by subsequent modules in the speech synthesis system. My model, on the other hand, besides cardinals, aims at handling ordinals, fractions and any numbers embedded in any context.

I designed a three-layer model for converting numbers to letters. The lowest layer is occupied by the basic algorithm of converting a string of digits to letters. The second layer contains the procedures for converting formatted structures, such as telephone numbers and decimal fractions. The highest layer hosts the parser for the context of numbers. The number–to–letter conversion program module has been realized according to the above model.

The number–to–letter conversion module was tested as part of the speech synthesis system. For the experiments I used electronic letters from which I extracted 560 expressions containing numbers in different contexts: date, time, cardinals, ordinals, monetary units, decimal fractions, software version numbers, numbers with suffixes and others. I found that 96% of the numbers have been converted correctly. Later errors were detected and corrected by means of the iterative model development cycle.

The number–to–letter conversion algorithm was first described in internal electronic report [R18]. I published my scientific results in [J1].

3. I realized an automatic prosody tagging method. I implemented both the prototype, handling the surface phenomena, and the complex version, exploiting the results of morphological and syntagmatic analysis.

In the speech synthesis system prosody is considered as the change of pitch, accentuation, durations, pauses, rhythm and the intensity of the speech. Prosody, as a suprasegmental phenomenon is superimposed on the segmental sequence of speech sounds.

In the speech synthesis system the handling of prosody is interpreted by the model of Olaszy, Németh and Koutny. The model has three layers: (1) linguistic analysis at the level of graphemes and determination of the high-level prosodic control, (2) analysis at the level of speech sounds and production of low-level control and (3) realization of prosody on the speech signal as specified by the parameters.

I implemented Olaszy's model of prosodic parameters, according to the following: 1. Melody curve: 1a. Global melody curve (sentence, clauses, phrases). 1b. Word stress. 1c. Microintonation in speech sounds. 2. Time structure: 2a. Global time structure of the sentence. 2b. Word duration map. 2c. Context dependent specific duration of the speech sounds. 3. Intensity structure: 3a. Sentence level intensity parameters. 3b. Word intensity map. 3c. Intensity changes at speech sound level.

Starting from Olaszy's prototype of the prosody tagging algorithm I systematized the rules and developed the algorithm to produce the prosodic tags without deeply analyzing the sentence. The method is based on the identification of function words and typical phrases in the sentence.

Using Koutny's linguistic model I implemented the complex version of the prosody tagging algorithm, where prosody tagging rules exploit the results of morphological analysis and identification of syntagmatic structures.

For the morphological analysis I used the *Humor* Hungarian morphological analyzer. I created a morphological dictionary to complete the analysis results. I designed a method to correct analysis errors and created an algorithm for parallel handling of alternative analysis results. I constructed a system for the effective representation of morphological features.

The framework of the syntagmatic analysis was given by Koutny's model for syntactic analysis, based on dependency grammar. Starting from this point, my representation of phrases was extended with features of the unification gram-

mar. For the identification of grammatical structures I designed a quick phrase pattern matching method using the output of the morphological analysis.

I formalized Koutny's rule system developed for the prosodic tagging of simple sentences of constrained meaning. These rules can identify phrase boundaries, assign melody patterns to them, set the place and length of pauses and can determine the level of emphasis on the words.

I realized both the simple prototype and the complex version of the prosody tagging algorithm as part of the speech synthesis system. The algorithms were improved through lengthy perception tests.

The most important feature of the thesis is that using the new prosody tagging method, the naturalness of the synthesized speech improved significantly – allowing the system to read sentences of different kinds, including interrogative and optative sentences.

To differentiate myself from my co-authors, I consider the followings as my own results: The prosodic rules' systematization and test for consistence; the strict formalism of the rule system; the design of the model's operating algorithms and the implementation of the prosodic tagging program module.

I formalized Olaszy's prototype model in electronic report [R7]. The works regarding the complex system are described in my diploma thesis [T1], in conference papers ([A1], [A2] and [C2]), and at a presentation [P2]. Technical aspects are discussed in a final project report [R19]. The three-layer prosody model realized in the speech synthesis system was presented at conference [C5].

In the framework of research cooperation with the Porto University of Technology I prepared a report on the general questions of sentence prosody generation for speech synthesis by means of syntactic analysis [R1].

4. I studied the multi-layer model of the classical generative phonology. I formalized the model for speech synthesis applications.

The earlier version of the speech synthesis system tried to produce the sound sequence in one step, using only the written forms of the isolated words. Sound level modifications were implemented at an elementary level only. This approach did not allow the handling of those pronunciation rules, which are applied at word boundaries and operate at the level of speech sounds (and not at the level of letters). Through the works of Kiefer, Durand and Siptár I studied the lexical model of phonology, and realized that a text–to–speech converter of high standard has to include a multi-level rule system designed with the lexical phonology in mind.

In the model of lexical phonology the rules can be divided into three levels. The first group is containing the phonological and morphological rules, working in the lexicon, producing the lexical representation of the words. The second level are the rules governed by lexical insertion, form the words appearing in the sentences. At the third level postlexical phonological rules produce the surface word forms. The latter ones operate on the speech sound sequence and can reach over word and morpheme boundaries as well.

As a first step I studied the two upper levels of the model. I recognized that a significant part of the lexical phonological rules is unnecessary to realize, because either a) the pronunciation (phonetic forms) of the words could be listed in the dictionary or b) due to the phonemic nature of the Hungarian writing system, the written form represents the spoken form of the word very well. The rules of lexical and syntactic insertion are represented at the level of grapheme–to–speech sound conversion. The rule system incorporates the letter–to–sound conversion rules for single and double letters, spelling, duration changes of vowels, lexical *h*-deletion, lexical palatalization, *j*-assimilation and others. I also stored the pronunciation of irregular family names, foreign words and some abbreviations here.

As a second step I formalized the rules of the postlexical level. However I studied these rules in the representation of distinctive features of the classical generative phonology, I attempted to develop a representation more suitable for practical applications. The rules implemented are consonant gemination and degemination, assimilation of nasals, palatalization of absolute word-final *j*, postlexical palatalization, assimilation of voicing, sound drop, merge, syncopation and other rules appearing at the postlexical level.

The importance of the thesis is that the new formalism can handle all aspects of the theoretical phonological phenomena in the speech synthesis system. A further advantage is that the exact definition of the rules reduces the interaction among them, and all the rules can be implemented at their own place (that is, the lexical ones in the lexicon, the postlexical ones at the level of speech sounds). In the implemented system, converting the rules into binary format in compilation time ensures fast and effective operation in run time.

The rules of lexical and syntactic insertion are formalized in [R11]; the postlexical phonological rules are systematized in [R13].

5. I constructed a language independent, universal speech sound representation scheme for artificial speech production. Herewith I established the possibility of constructing flexible multilingual speech synthesis databases.

Multivox version 3 was implemented for eight languages. In this version speech sounds were identified by serial numbers. The number of speech sounds in a language varied from 30 to 50. The sound codes of one language often represented a different speech sound in the other one. The extension of one particular language also posed problems: due to the assignment method of the numbers, the developers may have run out of the unassigned codes, making it impossible to add a new speech sound to the system. These problems prevented us creating one speech synthesis system integrating multiple languages.

I must emphasize that earlier versions of the speech synthesis system (due to storage constraints) confined themselves with handling only the *phonemes* of a language – the description of specific *speech sounds*, including the allophones, arose later.

For the solution of the above problem I designed a speech sound representation scheme. I studied the phonetic alphabet of the International Phonetic Association (IPA) as well as other documents available in the field.

When designing the speech sound representation, I considered the following aspects: one code should unambiguously represent a speech sound; the representation should be as compact as possible; the code should be able to represent the relation of sounds considered close from phonetic-phonological point of view; the representation should be complete (should include all the allophones, for instance); and last, the representation of the sound features should be orthogonal – with other words, one property of a sound should be described by one feature in the code, and one feature in the code should be assigned to only one sound property.

The advantage of the IPA phonetic symbol set is that it attempts to describe all the speech sounds that appear in human languages. Its disadvantage is that it does not show cohesion: sounds similar in pronunciation are assigned symbols that are not similar to each other. Further problem is that the symbol system makes a compromise in order to make human comprehension easier: some of the symbols resemble to the letter to denote the sound, other symbols seem to be chosen by chance, giving no clue for the untrained reader. Sometimes diacritics supplement the symbols.

The SAMPA phonetic symbol set was designed to represent and transmit the IPA symbols on computers. Here each of the IPA symbols is assigned one or

more ASCII characters. In this system one sound is often represented by three or four characters. From the current problem's point of view, the SAMPA system has the same disadvantages as the IPA has.

The proposed speech sound representation scheme is based on phonetic-phonological features. It defines six groups of features for the vowels: roundness, the horizontal and vertical position of the tongue, length of the vowel, advanceness of the tongue and nasality. Consonants have the following features: voiced quality, place and mode of articulation, syllabic quality, and other features for describing aspirated, palatalized, velar and other articulation variations. There is one code for each affricate, but diphthongs are represented by two and tripthongs by three codes.

The symbol system does not attempt to be human-readable, whereas it can be mapped to either the IPA symbol set or any other human-readable speech sound representation easily.

The importance of the thesis is that the new representation gives a consistent handling of speech sounds for both multilingual text–to–speech synthesis and can represent allophones when synthesizing one particular language. I published the representation scheme in the [R9] and [R10] internal electronic reports. The complex version of the waveform speech database also utilizes this representation. Prior to broader publication I intend to prove the usability of the system in more applications.

6. I defined the structure of speech databases containing diphones, triphones and larger elements. I deduced the problem of assembling the speech sound sequence to a well-known algorithm of graph theory.

As an improvement to the diphone concatenation technology, for the initiation of Németh and Olaszy, we started to develop a system, which would be able to handle diphones, triphones and elements of more speech sounds. Speech databases containing such elements are called complex speech databases.

I started my work from the structure of the diphone database. I developed a method to verify the structure and contents of the database and created an algorithm to prevent database inconsistency. I created the structure of the new complex database.

In the database containing diphones and consonant-vowel-consonant triphones the number of database elements is close to 8000. When assembling the speech sound sequence, the optional database elements have to be chosen in run time. Covering a speech sound sequence is considered to be optimal if it uses the minimum number of database elements to assemble the utterance. I designed

a method for fast search in the complex database and optimal selection of the database elements. I deduced the problem to Dijkstra's well-known algorithm for finding the shortest path in a graph. Starting from the original algorithm, I created an algorithm of linear time to cover the sequence of speech sounds, which is particularly important at larger number of database elements. I showed that the algorithm is always optimal. I implemented the search algorithm along with the experimental version of the complex database. We performed basic perception tests on the segmental speech signal.

I introduced the concept of speech sound substitution. When assembling the sound sequence, the database may not contain one or more of the required speech sounds. In this case the system has to choose database elements that are closest to the required speech sounds. The specified speech sounds are mapped to the sounds of the database using the sound substitution table. If a desired sound cannot be found in the database, then a database sound is selected that best approximates the desired sound from phonetic-phonological point of view.

The structure of the complex speech database, the internal speech sound representation and the comparison of database versions are summarized in the [R15], [R16] and [R17] electronic reports, respectively.

The importance of the thesis is that (1) I organized the database elements into a trie structure, providing fast search even at large number of elements and (2) constructed an algorithm to give the optimal covering of the speech sound sequence.

7. I formed the structure of the Blissvox speaking communication aid, used in the rehabilitation of speech-impaired people. I implemented the Mondgen morphological and syntactic generator for constrained vocabulary and grammar.

As a combination of the speech synthesis system and written natural language technology, in cooperation between the Laboratory of Speech Technology, DTT BUTE and the Helping Communication Methodological Center (HCMC), we constructed the Blissvox communication aid for the rehabilitation of disabled, handicapped children. The computer tool is based on the Bliss symbol system, used in the communication of disabled people. The users can assemble messages by concatenating simple symbols, pictograms. The messages are read in Hungarian by the formant synthesizer connected to the computer. I realized the first version of the system according to the concept of Doctor Sophia Kálmán (HCMC) and Gábor Olaszy (DTT BUTE) in 1994.

Originally the Bliss symbol system was developed for speakers of English. Sequential reading of the symbols' meanings resulted in a nearly correct English sentence. The Blissvox system used the Hungarian meanings of the symbols. Reading the meanings in Hungarian did not result in a correct sentence, as Hungarian has a highly agglutinative character, sometimes using a single word of multiple suffixes to express a structure where English uses a prepositional phrase. For solving this problem we designed the Mondgen system, intended to generate a sentence from a sequence of stems and affixes using a constrained grammar.

The sentence generator was formed for the constrained vocabulary of 2500 Bliss-symbols and for limited sentence structure. The sentence generator was designed to operate as a stand-alone module capable of being integrated to other applications as well. I prepared the Mondgen system according to the concept and under the supervision of Ilona Koutny. The synthesis of nominals was based on the work of Ferenc Papp, the synthesis of verbs was based on the work of László Elekfi. Both topics were systematized by Gábor Prószéky.

The process of grammatical generation can be divided into four steps: preprocessing, forming of nominal and verb phrases, constructing the sentence structure and morphological generation.

I designed a dictionary for the identification of affixes, and prepared an algorithm for the categorization of the words that do not appear in the dictionary. I formalized the rules building the phrase and sentence structure. I implemented an algorithm for the synthesis of nominals and verbs, using stem type, inflection and conjugation type, vowel harmony, and number and person. I implemented the Mondgen system as a part of Blissvox.

According to the experiments at the HCMC the system met the expectations. The basic version of the Blissvox system was introduced in conference paper [C1] and at presentation [S1]. The Hungarian translation of the conference paper was published in [B5]. The Mondgen system was first published in [S2]. The Blissvox system improved with Mondgen was presented in [J5].

5 Acknowledgements

I thank my scientific advisors, Gábor Olaszy and Géza Németh for introducing me several fields of language technology and helping me during my activity at the Laboratory of Speech Technology. They helped very much with advices, provided support for the research, and encouraged me to publish the results.

I thank Ilona Koutny for acquainting me with the linguistic theories during the development of the sentence generator and the prosodic tagging system, and for putting me on the way to study Hungarian grammar more deeply.

I thank Gábor Prószéky for providing us the *Humor* morphological analyzer for research purposes. Besides this, he helped the realization of the syntagmatic analyzer with numerous advices.

I received valuable help from Péter Siptár in forming the English letter–to–phoneme conversion rules. He spent tremendous time on testing and verifying my algorithm and drew my attention to several problems. I could also lean on his opinion and advices during the design of the language independent speech sound representation. I learned much from both personal conversations and reading his books.

I owe special thanks professor Géza Gordos, former head of the Department of Telecommunications and Telematics for his interest in my work and supporting my research at the department since my undergraduate years.

6 Publications related to the theses

Notation:

B Articles in edited books
J Journal articles published abroad in a foreign language
C Conference papers published abroad in a foreign language
A Conference papers published as an abstract only
P Presentations not published
R Research reports, electronic memoranda not available in libraries
S Papers at the Students' Scientific Conference
T Diploma thesis

6.1. Articled in edited books

[B1] **Olaszi P**. (1994). *Számítógépes algoritmus angol szöveg fonetikus átalakítására* (Algorithm for English Text–to–Phoneme Conversion). In: Mária Gósy (Ed.): *Beszédkutatás '94* (Speech Research '94), Hungarian Academy of Sciences, Linguistic Institute, Budapest. pp. 183–197.

[B2] Olaszy G., **Olaszi P**. (1998). *Hangidőtartamok mesterséges változtatása periódusok kivágásával, megismétlésével* (Artificial modification of speech sound durations by means of deletion and insertion of pitch periods). In: Mária Gósy (Ed.): *Beszédkutatás '98* (Speech Research '98), Hungarian Academy of Sciences, Linguistic Institute, Budapest. pp. 151–162.

[B3] Németh G., Zainkó Cs., Bogár B., Szendrényi Zs., **Olaszi P**., Ferenczi T. (1998). *Elektronikuslevél-felolvasó* (E-mail reader). In: Mária Gósy (Ed.): *Beszédkutatás '98* (Speech Research '98), Hungarian Academy of Sciences, Linguistic Institute, Budapest. pp. 189–203.

[B4] Olaszy G., Kiss G., Németh G., **Olaszi P.** (2000). *Profivox: a legkorszerűbb hazai beszédszintetizátor* (Profivox: The most advanced Hungarian speech synthesizer). In: Gósy Mária (Ed.): *Beszédkutatás 2000* (Speech Research 2000), Hungarian Academy of Sciences, Linguistic Institute, Budapest. pp. 167–179.

6.2. Journal articles published abroad in a foreign language

[J1] **Olaszi P**. (2000). Analysis of Written and Spoken Form of Hungarian Numbers for TTS Applications. In: Olaszy G., Gardner-Bonneau D. (Eds.): *International Journal of Speech Technology*. Kluwer, Boston. pp. 177–186.

[J2] Koutny I., Olaszy G., **Olaszi P**. (2000). Prosody Prediction from Text in Hungarian and its Realization in TTS conversion. In: Olaszy G., Gardner-Bonneau D. (Eds.): *International Journal of Speech Technology*. Kluwer, Boston. pp. 187–200.

[J3] Olaszy G., Németh G., **Olaszi P**., Kiss G., Zainkó Cs., Gordos G. (2000). Profivox—a Hungarian Text-to-Speech System for Telecommunications Applications. In: Olaszy G., Gardner-Bonneau D. (Eds.): *International Journal of Speech Technology*. Kluwer, Boston. pp. 201–215.

[J4] Németh G., Zainkó Cs., Fekete L., Olaszy G., Endrédi G., **Olaszi P**., Kiss G., Kis P. (2000). The Design, Implementation and Operation of a Hungarian E-mail Reader. In: Olaszy G., Gardner-Bonneau D. (Eds.): *International Journal of Speech Technology*. Kluwer, Boston. pp. 217–236.

[J5] **Olaszi P**., Koutny I., Kálmán S. L. (2002). From Bliss Symbols to Grammatically Correct Voice Output: A Communication Tool for People with Disabilities. In: Gardner-Bonneau, D. (Ed.): *International Journal of Speech Technology*. Vol. 5:1. Kluwer, Boston, pp. 49–56.

6.3. Conference papers published abroad in a foreign language

[C1] Olaszy G., Kálmán Zs., **Olaszi P**. (1994). BLISSVOX – Voice Output Communication System for Teaching, Rehabilitation and Communication. In: Zagler, W.L., Busby, G., Wagner, R.R. (Eds.): *Computers for Handicapped Persons, 4th International Conference*. Springer-Verlag, Wien. pp. 421–428.

[C2] **Olaszi P**., Koutny I., Olaszy G., Németh G. (1998). Syntactic Analysis of Hungarian Sentences to Predict Prosodic Information for Speech Synthesis. In.: *Proceedings of the 1998 Polish-Czech-Hungarian Workshop on Circuit Theory, Signal Processing and Applications*. Krakow, Poland. pp. 49–54.

[C3] Olaszy G., Németh G., **Olaszi P**., Gordos G. (1999). Interactive, TTS Supported Speech Message Composer for Large, Limited Vocabulary, but Open Information Systems. In.: *Eurospeech '99*. Vol. 2., Budapest, Hungary. pp. 943–946.

[C4] Olaszy G., Németh G., **Olaszi P**. (1999). Preparation of Limited Vocabularies with an Interactive Open TTS Based Development System. In: *Proceedings of the 1999 Polish-Hungarian-Czech Workshop on Circuit Theory, Signal Processing and Applications*. Herbertov, Czech Republic. pp. 73–76.

[C5] Olaszy G., Németh G., **Olaszi P**. (2001). Automatic Prosody Generation – A Model for Hungarian. In.: *Eurospeech 2001*. Vol. 1., Aalborg, Denmark. pp. 525–528.

7 Works not reviewed

7.1. Articles in edited books

[B5] Olaszy G., Kálmán Zs., Olaszi P. (1994). *A Blissvox – beszélő kommunikációs rendszer* (Blissvox – a speaking communication system). In: Mária Gósy (Ed.): *Beszédkutatás '94* (Speech Research '94), Hungarian Academy of Sciences, Linguistic Institute, Budapest. pp. 228–236. (Hungarian translation of [C1].)

7.2. Conference paper published as an abstract only

[A1] **Olaszi P**. (1998). Syntactic Analysis of Hungarian Sentences to Predict Prosodic Information for Speech Synthesis. In: *Conference of Ph.D. Students in Computer Science*. Advisors: Koutny I., Olaszy G., Németh G., "Best Talk of the Session" award. JATE, Szeged, Hungary. *http://www.inf.u-szeged.hu/~cscs/cscs.ps.gz* pp. 81.

[A2] **Olaszi P**. (1998). Syntactic Analysis of Simple Sentences to Predict Prosodic Information for Speech Synthesis. In: *Doximp 1998 – 3rd Doctoral Symposium*. Advisors: Koutny I., Olaszy G., Németh G. Eötvös Lóránd University of Sciences, Budapest.

7.3. Presentations not published

[P1] **Olaszi P**. (1995.07.14). The Multivox English TTS Presentation. *Private presentation*. BUTE DTT Laboratory of Speech Technology. `tts_pres.arj`

[P2] **Olaszi P**. (1997.06.12). Syntactic parsing of Hungarian sentences for presenting prosody information for speech synthesis. *Private presentation for the representation of the Ericsson company*. BUTE DTT Laboratory of Speech Technology. `ericsson.doc`

[P3] **Olaszi P**. (2001.02.19). Speech database size reduction by vector quantization. *Private presentation*. Nippon Telegraph and Telephone Corporation, Cyber Space Laboratories. Yokosuka, Japan. `ntt_final_presentation.doc`

7.4. Research reports, electronic memoranda not available in libraries

[R1] **Olaszi P**. (1997). Sentence prosody generation for speech synthesis by means of syntactic analysis – a Case Study for Hungarian. *Research report*. Faculdade de Engenharia da Universidade do Porto, Portugal.

[R2] **Olaszi P**. (1998.06.14). *Javaslat a Multivox 5-ös formánsszintetizátor alapú változatának architektúrájára* (Proposal for the formant synthesis-based architecture of Multivox 5). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `mv5arch.txt`

[R3] **Olaszi P**. (1999.11.04). *A Multivox beszédszintetizátor 5-ös változatának architektúrája* (The architecture of the Multivox 5 speech synthesizer). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `mv5arch2.txt`

[R4] **Olaszi P**. (1998.06.08). *Javaslat a Multivox 5-ös változat API függvényekre és a bemeneti szövegben elhelyezhető vezérlő szekvenciákra* (Proposal for the Multivox 5 API functions and command sequences in the input text). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `mv5api.txt`

[R5] **Olaszi P**. (1998.12.16). *Javaslat a Multivox 5-ös változat API függvényeire* (Proposal for the API functions of Multivox 5). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `mv5api2.txt`

[R6] **Olaszi P**. (2000.07.25). *Javaslat a Multivox beszédszintetizátor bemeneti szövegében megadható vezérlő szekvenciákra* (Proposal for the Multivox command sequences in the input text). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `control.txt`

[R7] **Olaszi P**. (1999.05.04). *Egyszerű prozódiai szabályok a Multivox rendszerhez Olaszy Gábor útmutatása alapján* (Simple prosodic rules for the Multivox system according to the guidelines of Gábor Olaszy). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `pr_szab.txt`

[R8] **Olaszi P**. (2000.05.16). *Graféma–graféma átalakítási szótár kezelése* (Handling the grapheme–to–grapheme conversion dictionary). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `gtg_dic.txt`

[R9] **Olaszi P**. (1998.06.17–2000.07.06). Language independent sound properties. *Multivox 5 source file*. BUTE DTT Laboratory of Speech Technology. `sound.h`

[R10] **Olaszi P**. (1998.06.14). *Javaslat a Multivox 5-ös változatában a beszédhangok ábrázolására* (Proposal for the representation of speech sounds in Multivox 5). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `mv5hang.txt`

[R11] **Olaszi P**. (2000.04.27). *Graféma–hang átalakítási szabályok magyar nyelvre* (Grapheme–to–sound conversion rules for Hungarian). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `gts_hu.txt`

[R12] **Olaszi P**. (2000.03.03). *Fonetikus ábécé a magyar beszédhangok ábrázolására* (Phonetic alphabet for representing Hungarian speech sounds). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `hangkod.txt`

[R13] **Olaszi P**. (1999.06.02). *Hang–hang átalakítási szabályok magyar nyelvre* (Sound–to–sound conversion rules for Hungarian). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `sts_hu.txt`

[R14] **Olaszi P**. (1999.03.10, 1999.09.27). *A Multivox beszédszintetizátor hullámforma alapú diádos hangadatbázisának formai és tartalmi vizsgálata* (Testing the structure and contents of the diphone waveform speech database of the Multivox speech synthesizer). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `dbase1.txt`

[R15] **Olaszi P**. (2000.03.29). *A Multivox beszédszintetizátor 2-es változatú (hullámforma alapú, triádokat vagy több hangot tartalmazó) hangadatbázisának formai és tartalmi vizsgálata* (Testing the structure and contents of the version 2 waveform speech database containing triphones and larger elements at the Multivox speech synthesizer). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `dbase2.txt`

[R16] **Olaszi P**. (2000.03.29). *A Multivox beszédszintetizátor 6-os változatától kezdve alkalmazott belső hangadatbázis ábrázolás* (Internal speech sound representation applied from version 6 of the Multivox speech synthesizer). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `dbase.txt`

[R17] **Olaszi P**. (1999.12.12). *A hangadatbázist kezelő függvények változásáról* (About the changes in the functions handling the speech database). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `db_diff.txt`

[R18] **Olaszi P**. (1999.04.28–2000.05.09). *A számok magyar szöveggé alakításáról* (About converting Hungarian numbers to text). *Internal electronic memorandum*. BUTE DTT Laboratory of Speech Technology. `szamok.txt`

[R19] **Olaszi P**. (1998.10.22). *Prozódiai előrejelző a Profivox beszéddel válaszoló rendszerhez* (Prosody predictor for the Profivox speaking system). *Final research report*. BUTE DTT Laboratory of Speech Technology. `zj981022.doc`

[R20] **Olaszi P**. (2001.02.21). Speech database size reduction by vector quantization. *Internal final project report*. Nippon Telegraph and Telephone Corporation, Cyber Space Laboratories. Yokosuka, Japan. `ntt_final_report.doc`

7.5. Papers at the Students' Scientific Conference

[S1] **Olaszi P**. (1994). *Blissvox – Beszélő kommunikációs rendszer rehabilitációhoz és oktatáshoz* (Blissvox – Speaking Communication System for Teaching and Rehabilitation). *Students' Scientific Conference*, BUTE Faculty of Electrical Engineering and Informatics. 3rd prize. Advisors: Olaszy G., Kálmán Zs.

[S2] **Olaszi P**. (1995). *Kötött szótáras mondatgenerátor magyar nyelvre; illesztése a Blissvox kommunikációs és rehabilitációs programhoz* (Hungarian Sentence Generator with Constrained Vocabulary – Connecting to the Blissvox Communication and Rehabilitation Program). *Students' Scientific Conference*, BUTE Faculty of Electrical Engineering and Informatics. 2nd prize. Advisors: Koutny I.

7.6. Diploma thesis

[T1] **Olaszi P**. (1997). *Algoritmusok az egyszerű magyar mondat szerkezetének meghatározásához prozódiai szerkezetek előrejelzése céljából* (Algorithms to determine the syntactic structure of single-clause Hungarian sentences to predict prosodic structures). *Diploma thesis*. BUTE Faculty of Electrical Engineering and Informatics.