SpeechLab 2.0 – A High-Quality Text-to-Speech System for Bulgarian

Maria Andreeva and **Ivaylo Marinov** and **Stoyan Mihov**
Institute for Bulgarian Language,
Bulgarian Association for Computational Linguistics,
Institute for Parallel Processing
Bulgarian Academy of Sciences
stoyan@lml.bas.bg

Abstract

SpeechLab 2.0 is a high quality and very efficient text-to-speech engine for Bulgarian, which applies a sophisticated rule-based approach for tokenization, part-of-speech tagging, phonetization and prosody annotation. All stages of the text processing, including grammatical and accentual dictionary application, contextual rules for grammatical and prosodic annotations and phonetization, are implemented as a pipeline of finite state bimachines and subsequential transducers. The main advantage of the new method for text processing for speech synthesis is its very high efficiency for complex linguistic analysis. Some specific aspects of the Bulgarian phonology and prosody are presented as well. Using the FD-PSOLA algorithm for signal generation our system delivers Bulgarian speech with very high intelligibility and naturalness by a processing speed of 960 words/sec.[1]

Keywords: text to speech, text processing, finite state devices, natural language processing

1 Introduction

A modern text-to-speech (TTS) system has to provide a number of features including sophisticated text analysis and robustness, to deliver intelligibility and naturalness of the generated speech and to achieve high performance and robustness. Currently a number of high quality TTS systems for English, French and German have been developed (Santen *et al.* 97; Dutoit 94). The development of a high quality TTS system for a new language is still a challenge.

To the best of our knowledge, no comprehensive achievements have been made so far in the development of a high-quality speech synthesizer for Bulgarian. The first Bulgarian TTS system called "Betsy" has been developed by Borislav Zahariev in the framework of his Ph.D Thesis (Zahariev 93). While being quite an achievement for the time it was created, Betsy's voice sounds unnatural and applies only a few linguistic rules, which makes it hard to understand and does not suffice the need for a fast and reliable text-to-speech system. In (Totkov *et al.* 03) the authors report for another Bulgarian TTS system. This system also suffers from the lack of more advanced text analysis techniques for accent and prosody annotation.

In this paper we present a new approach for text processing, which we use in the SpeechLab 2.0 TTS system. We apply a rule-based approach implemented by a pipe finite state devices (Kaplan & Kay 94; Roche & Schabes 97). Although finite state devices have been used extensively for phonetic translation (Sproat 97; Laporte 97), our method differs significantly in respect to the following three features: First we use only subsequential transducers and bimachines, which have been constructed by determinization of regular relations. Second, all the finite-state devices (including the dictionaries) are text rewriting devices – the whole text is deterministically transduced by each of the transducers. Third, starting from the input text all finite-state devices are simply applied in a pipe, which provides a simple implementation, application of complex linguistic rules and dictionaries and results in very high efficiency.

The first phase of text processing proceeds with the GrammLab system (Doychinova & Mihov 04) for Part-of-Speech annotation. At the second phase we apply an Bulgarian accentual dictionary with 1 million wordforms, an English pronunciation dictionary with 60000 wordforms and a user defined custom pronunciation dictionary. The third phase proceeds with 98 contextual rules for phonetization, accent determinization, unknown words processing. At last 91 rules for prosody annotation are applied. For the signal synthesis we use a diphone concatenation system based on the well-known FD-PSOLA method (Moulines & Charpentier 90), which delivers a high quality Bulgarian speech.

In the next section we present the general text analysis provided by the system. Section 3 describes the details of the phonetization procedure

[1] This work was partly funded by a grant from Microsoft.

for Bulgarian. We present the Bulgarian prosody annotation in Section 4. The technology used for text processing in the SpeechLab 2.0 system is presented in Section 5 and the implementation details are given in Section 6. Finally the conclusion presents some general comments and directions for further work.

2 General text analysis

The general text analysis component in SpeechLab 2.0 consists of four stages – tokenization, grammar dictionary application, unknown words guessing and contextual part-of-speech (POS) disambiguation. At the end the input text is annotated with token tags and POS tags. We have used the tagger described in (Doychinova & Mihov 04) for the implementation of this subsystem.

The overall precision of the tagger is over 98.4% for full disambiguation and the processing speed is over 34K words/sec on a personal computer.

Below we give a short overview of the four stages.

2.1 Tokenizer

The system uses a sophisticated tokenizer, which marks and categorizes tokens as numeral expressions, dates, hours, abbreviations, URLs, items, punctuation, words, abbreviations etc. Words in Latin and Cyrillic are differently marked when capitalized or upper case. Punctuation is classified according to the type and the function of the sign. Sentence boundaries are recognized as well. The tokenizer is implemented as a composition of 53 contextual rewriting rules composed into 4 bimachines.

2.2 Grammatical dictionary

The dictionary assigns to each dictionary word in the text its ambiguity class and its initial tag. The initial tag is usually the most probable tag for the ambiguity class. For example, the class which consists of adverbs and neutral adjectives gets adverb as a most probable tag, because in an representative corpus these words are adverbs 2 times more often than adjectives. The dictionary contains about 1 million wordforms. Since each wordform can occur in lower case, upper case or capitalized, the dictionary contains 3 millions strings. It is implemented as one (big) rewriting rule represented by a subsequential transducer using the construction method presented in (Mihov & Schulz 04).

2.3 Unknown word guesser

The guesser handles the words that are not in the dictionary. The constructed rules are analyzing the suffix for guessing the word's morphology and for assigning the initial part-of-speech tag and the ambiguity class. The guesser is implemented as a composition of 73 rewrite rules all compiled into a single bimachine.

2.4 Contextual Disambiguation Rules

The part-of-speech ambiguity ratio for a representative Bulgarian corpus is 1.51 tags/word, which means that in average every second word is ambiguous. For solving the ambiguities we apply 148 contextual rules, which can utilize part-of-speech, lexical or dictionary information on the context. All 148 contextual rules are composed into 4 bimachines, which we apply in the pipeline.

3 Phonetization

We present a brief description of the phonetic inventory of Bulgarian, with a discussion of the approach used to select and segment phonetic units for the system. One of the main specifics of Bulgarian language is the combination of phonetic and morphological orthographic principle. This is the reason for us to choose the description for text-to-speech processing by means of phonetic and grammatical rules that rewrite the text by annotating it with phonetic description.

For the development of the formal rules a morphologically tagged corpus of 50000 words was transcribed. A set of rewrite rules were developed for the preprocessing of the orthographic format of written text into a phonetic alphabet, serving as input to the speech synthesizer. Our work was oriented towards capturing the specifics of the Literary Bulgarian speech and attaining intelligibility and expressiveness of the generated speech.

3.1 Phonetic Description of Bulgarian

The first stage of formal description included an large size Bulgarian accentual dictionary and the elaboration of letter-to-sound rules that capture the close correspondence existing between letters (symbolic representation) and pronunciation.

One of the specifics of the Bulgarian language is the free word stress. The change of stress position in the different forms of the word, homographs and the accent features and specifics of some clitics in different word order constructions gives rise to many problems in text-to-speech preprocessing.

Stress movement is accounted for by means of a dictionary consisting of over 1 million units with marking of primary and supplementary stresses. Most cases of homography are resolved by the tagger. For the description of the phonetic peculiarities of some specific word order models and grammatical constructions a set of rules were created as well as special dictionaries of some phrases. Those rules have been used to resolve cases like the one in “Добър вечер” (Good evening) where in the adjective “добър” (good) the vowel "о" is stressed, while usually "ъ" is the stressed vowel. Various combinations of prepositions, conjunctions and particles are analyzed for stress movements as well.

For the purposes of our text-to-speech system we adapted the traditional phonological system of Bulgarian to a phonetic system consisting of 45 phonemes, including 39 consonants and 6 vowels with additional accounting for certain specifics of speech such as stressed and unstressed vowels. etc. The precise number of Bulgarian phonemes was determined with regard to the necessary-and-sufficient condition in the representation of the groups of vowels and consonants needed for the diphone concatenation and its later implementation in the generation of natural sounding Bulgarian speech. Speech was modeled as a linear sequence of these phones. Table 3.1 shows the correspondence between the phonemes, their pronunciation being represented by the corresponding symbol established in the standard Phonetic Alphabet - SAMPA, and the characters of the Bulgarian alphabet, and an example of the occurrence of each phoneme in a Bulgarian word.

3.2 Structural rule-based representation of Bulgarian speech

Rules are used to perform accurate phonological description of Bulgarian speech. The representation of Bulgarian speech is based on a set of phonological distinctive features – atomic units by means of which phonemes are described. The groups of consonants and vowels are further divided into subgroups on the basis of the correlative phonetic features regularly occurring in literary speech. The determination of the set of vowels differs from the traditional approach in that it includes all stressed vowels and their unstressed allophones, the reduction feature (in unstressed position) conforming to reduction rules established in literary tradition. The group of consonants (obstruents) is divided into subgroups based on the correlative characteristics +voice and -voice; +palatalness and -palatalness. The glide [j] groups together the sounds represented in spelling by the letters "й", "ь" and the glide formant of the sounds represented graphically by "я", "ю".

Consonants			
Non-palatalized	Palatalized	Vowels	Semivowel
б[b] баба	[p’] пял	ъ[@] пън	[j] ял
в[v] вада	[b’] бял	о[o] кон	
г[g] гарга	[t’] тях	у[u] тур	
д[d] домат	[d’] дял	е[e] фес	
ж[Z] жаби	[k’] кяр	и[i] пир	
дж[dZ] джам	[g’] гюл	а[a] чар	
з[z] змия	[ts’] цял		
дз[dz] дзвън	[dz’] дзян		
п[p] папка	[f’] фют		
Ф[f] фонтан	[v’] вял		
к[k] котка	[s’] сял		
т[t] телефон	[z’] зян		
Ш[S] шапка	[x’] хюм		
Ч[tS] чаша	[m’] мях		
с[s] сам	[n’] ням		
ц[ts] цаца	[l’] лях		
х[x] хора	[r’] ряз		
р[r] роза			
л[l] лекар			
м[m] мама			
н[n] нос			

Table 1: Bulgarian Phonemic System.

Other special rules describe the letters standing for two sounds like "щ", "ю", "я" or single phonemes represented by digraphs such as "дз", "дж". In these cases the letters may at the same time represent two distinct sounds occurring on a morpheme boundary. The problem of ambiguity in these cases is sufficiently resolved with special rules and a dictionary of words that include these digraphs, most of which are in fact rare or dialect words.

The organization of the phonemes into groups is the basis for the elaboration of formal rules for resolving problems such as stressed or unstressed vowels, voice assimilation and palatalization of consonants.

The possible sound alternations are described by means of context rules, based on the contradistinction of correlative features. The assimilation in Bulgarian may be described as an anticipatory (regressive) adoption ("copying") of a certain feature by the target sound from the sound in the immediate environment following it. The source of assimilation is the second sound in the sequence. One of the most important features of Bulgarian consonants is the alternation of voiced and voiceless consonants in certain positions in words.

The rules for the combination of two or more consonants replace voiced consonants with voiceless and vice versa under certain conditions (the influence of the first neighboring consonant regarded backwards). For example the word "отдавна" (a long time ago), in the combination of "тд", "д" voices "т" and the word is pronounced [odd‘avna]. In the word “безсилен” (feeble), in the combination "зс", "с" influences "з" and the word is pronounced [bess‘ilen].

A special place in the description is determined for the phonemes "в" and "в’", after which both voiced and voiceless consonants can be pronounced - for example “звяр” [zv’‘ar] (beast), “свят” [sv’‘at] (world), “двор” [dv‘or] (yard), “творчество” [tv‘ortSestvo] (creative work).

In connected speech words form a continuous speech chain. Another important phonetic characteristic of the Bulgarian language, which should be described with formal rules, is the word boundaries assimilation. This is the reason for the characteristic changes that occur word initially and word finally in neighboring words. Each word or a set of words is regarded as a string of phonemes, whose boundaries are determined by special rules. For example, some consonants are not influenced by the assimilation under certain conditions.

A special case is presented by the negation particle "не" [ne] (not), which is unstressed in the construction "Той не ме удари" [t‘oj ne m‘e ud‘ari] (He has not hit me.), since it is followed by the short personal pronoun "ме" (me). In the fragment "не зелен, но не и червен" [n‘e zel‘en no n‘e i tServ‘en] (not green but not red as well) the particle is stressed, because it is followed by an adjective. The formal rules for the particle "не" (not) also subsume grammatical constructions with the auxiliary verb "съм" (to be), otherwise unstressed in all its forms, but assuming stress in all forms to the exception of the 3d person, singular, Present form “е” [e] (is), if it is preceded by the negative particle "не" (not), e. g. "Той е човек. Ти не си човек" [t‘oj e tSov‘ek t‘i ne s‘i tSov‘ek] (He is a man. You are not a man.). According to the their stress features function words were described and classified in three classes: always stressed, always unstressed and stressed under specific conditions.

Special rules for the determination of stress positions are needed for the definite forms of masculine nouns and definite masculine word forms of adjectives, numerals in some of which the definite article receives the word stress - "‘а/‘я" "‘ят" (pronounced [‘@/j‘@] and [j‘@t]) for example: "мъж" [m‘@Z] (man) - "мъжа" [m@Z‘@] (the man) - "мъжът" [m@Z‘@t] (the man); "ден" [d‘en] (day) - "деня" [den’‘@] (the day) - "денят" [den’‘@t] (the day). The same problem occurs also with first person, singular and third person plural present tense forms of Bulgarian verbs belonging to I and II conjugation type ending in "‘а/‘я" and "‘ат/‘ят" (pronounced [‘@/j‘@] and [‘@t/j‘@t]), for example "чета" [tSet‘@] (I read) - "чет‘ат" [tSet‘@t] (they read); "плат‘я" [plat’‘@] (I pay) - "плат‘ят" [plat’‘@t] (they pay).

Many problems are caused by so called homographs – words that are represented identically but differ in pronunciation. When the corpus was phonetized some interesting cases were described and classified in groups:

- Nouns in which the change in stress leads to a change in the category gender for example: “техника” (technician / technics), “физика” (phisician / phisics)

- Words in which the change in stress changes their part of speech, for example: “трупа” (the corpse / group), “душа” (the shower / soul)

- Verbs in which change in stress leads to change of the verb tense, for example “заделя”, “заделят” (put aside), “споделя”, “споделят” (share)

- Words that coincide in all their grammatical forms, but have different semantics for example “вълна” (wool / wave), “пара” (steam / penny), “блажен” (fat / blessed)

These cases cannot be resolved by the tagger and context rules, but can be partially reduced by placing secondary (weaker) stress to the vowels in question, or rather vowel length markers, to the wordform already recognized by the system as a homograph. In such a way an auditory impression of a correct pronunciation of these forms is created, while at the same time the possibility for a completely wrong pronunciation of the synthesized word in the concrete instance is eliminated.

4 Prosody annotation

We consider the intonation as a set of prosodic elements characterizing human speech and its realization in the act of the verbal communication. The grouping of the prosodic elements according to their significance and their classification in different groups shows that some of them are obligatory for the naturalness and intelligibility of speech, while others (not obligatory in this respect) are characteristic for individual speech and express the attitude of the speaker. The first group includes the different speech melody types characterizing the main communicative types of sentences in Bulgarian, the correct putting of stress of words and definition of pauses with respect to the punctuation as well as the logical accentuation of words. Non-obligatory prosodic elements include emotional speech indications such as sudden changes of voice, changes in voice power, lengthened vowels, slower or faster articulation of some words in the sentence.

Along with the description of the regularities of sound patterns of Bulgarian, phonology is concerned with the more abstract description of the intonation patterns viewed as a suprasegmental property of speech indispensable for the intelligibility of automatically generated speech.

We used a testing tool for interactive modeling of prosody changes to the synthesized speech. Using it we conducted a number of intonation modeling tests and experiments for the elaboration of formal rules describing the Bulgarian prosody.

4.1 Pauses

The rules determining the pauses in a string play a very important role in speech synthesis. The generated pauses should not interrupt speech or change the meaning, but should add to the melody contour of the segments. In different types of sentences different pauses operate. Their major role is in the process of the definition of the intonation segments' boundaries. All these factors were taken in consideration in the formulation of the formal pause definition rules to account for the different cases:

- the phoneme string is temporarily interrupted by a new line, new page or new paragraph mark;

- different length of pauses is associated with the different punctuation marks;

- in cases where two punctuation marks are combined the longest pause is considered;

- in sentences, which exceed a given number of wordforms determined by physiological factor;

- alternation on word boundaries where assimilation is forbidden a short pause interrupts the phoneme string. For example the cases when a word ends with "щ", "жд" и "ст";

4.2 Intonation

The rules describe the intonation contour of the four main communicative types of sentences in Bulgarian and their further subdivision into more detailed subtypes:

- Intonation of a text segment ending with a period "." - for example: “Той ходеше по улицата, по която беше вървял преди година.” (He was walking on the street, on which he walked a year before.)

- Intonation of a text segment ending with a question mark "?" which does not contain question word or particle - for example: “Вие мразите ония горе? Не знаете нищо? Имате домашно по математика?” (You hate those people above? You don't know anything? You have a homework on mathemathics?)

- Intonation of a text segment ending with a question mark "?", containing at least one question word - for example: “Колко струва това палто? Къде зимуват раците? Кой отговаря за тази работа?” (How much costs this coat? Where do the crabs spend the winter? Who is responsible for this job?)

- Intonation of a text segment ending with a question mark "?", containing a question particle - for example: “Сега ли трябва да ходя на работа? Той нали ще дойде на време?” (Should I go to work now? He will come on time, won't he?)

- Intonation of a text segment ending with a question mark "?", containing both question word and particle - for example: “Кой знае колко е страдала, колко е мислила за него и колко дълго го е очаквала?” (Who knows how much she suffered, how much she was thinking about him and how long she was wating him?)

The methodology adopted for the representation of these groups is based on the proper determination of a minimal intonation segment - the intonema. By intonema we mean any stressed content word, which attracts proclitics or enclitics and determines their tonal value.

For example in a sentence containing one content word and ending with a punctuation mark ".", as "Във вазата." (In the vase.), the system will track the main word stress and will add the unstressed preposition "във" (in) to construct an intoneme and will define its intonation contour as rise-fall marked by the punctuation mark ".".

If the sentence comprises two intonation segments, the first one ending with ",", and the second with ".", as in "Трябва да изберем човек, в който да сме сигурни." (You have to choose a man, of whom we can be sure.), the program will determine the intoneme of the first segment, finishing with "," as rise-fall and will apply it to the intoneme preceding the comma. The second segment in this sentence will be synthesized applying the rules for melody contour determined by a punctuation mark ".".

The intoneme generation involves the application of additional rules that take into account the characteristics of speech resulting from factors such as the number of intonemes in an intonation segment and the length of individual intonemes. On the operation of these rules, the system generates rules for segment initial, segment final and segment internal intonemes, as well as for special intonemes such as question words, question particles, etc.

5 Text processing technology

As already mentioned the whole text analysis module is implemented as a pipeline of bimachines and subsequential transducers. The text is transduced by each of the finite state devices deterministically. After each step the text is enriched with additional information, which is used by the following steps. At the end the text is annotated with token tags, POS tags, annotations for phonemes and prosody.

For the construction of the rewriting rules we used the methods presented in (Kaplan & Kay 94; Gerdemann & vanNoord 99; Ganchev *et al.* 03). All the rules are specified in the form $\alpha \rightarrow \beta / L\_R$, where α, β, L and R are regular expressions. This means that each occurrence of a substring presented as α is replaced by β if it occurs in the context of substrings of L and R ((Kaplan & Kay 94)). For solving the conflicts we used the “first left longest match” strategy with the implementation given in (Gerdemann & vanNoord 99). The actual realization of 2-tape automata and the operations on them are implemented using the “one-letter automata” methodology given in (Ganchev *et al.* 03). The rules are then composed and represented by bimachines by the techniques presented in (Roche & Schabes 97). For the construction of the sequential transducer for the rewriting dictionary rules we applied the new method we developed and presented in (Mihov & Schulz 04).

We have developed a new rule compiler called “Siera” for practical implementation of the above functionality. This system allows the description and construction of all needed language resources including rule compilation from regular expressions to one-letter automaton; composition, union, concatenation of one-letter automata; conversion of functional one-letter automata to subsequential transducer or bimachine; compilation of rewrite dictionaries to susequential transducer and many others. In practice the entire text analysis is described as set of Siera scripts.

This methodology has provided a powerful, flexible and comfortable linguistic development environment while resulting to an exceptionally high performance.

6 Implementation details

The SpeechLab 2.0 system is implemented in ANSI C with a platform independent core. It was successfully tested under Linux and Windows.

The base pitch can be set in the range of one octave and the speed range is between two times slower and two times faster than normal. The usual options for reading the punctuation marks or for spelling given expressions are provided. For the fulfillment of specific requirements the system is supplied with a module for dynamic setting of specific configuration options. For example, the user can switch off the English dictionary, intonation, stresses and pauses between words, numbers can be read by digits, dates and abbreviation expansion can be controlled.

The size of text processing module is 120 MB and the size of the voice module is about 10 MB. Using memory map files the memory occupied in the computer is about 15 MB.

The speed of the synthesis including the saving of 16KHz 16 bit audio stream is 963 words per second on a 3 GHz Pentium 4 computer running Linux. Even on a 100 MHz Pentium II computer with only 32 MB SpeechLab 2.0 was able to provide real-time synthesis.

7 Conclusion

The presented SpeechLab 2.0 system provides a high quality Bulgarian speech. People with visual disabilities have extensively tested the system. Currently it is available free of charge for all Bulgarian visually disabled people from "Horizonti" foundation and the Union of the Blinds in Bulgaria. In result of the tests SpeechLab 2.0 was acknowledged to fulfill the requirements of the visually impaired people.

Although the tests revealed a high quality of the synthesized speech, there is still room for improvement in a couple of directions. In the near future we plan to create a more sophisticated syntax analysis, better homograph resolution rules and more complex intonation contour annotation.

References

(Doychinova & Mihov 04) Veselka Doychinova and Stoyan Mihov. High performance part-of-speech tagging of bulgarian. In *Proceedings of AIMSA 2004, LNAI #3192*, 2004.

(Dutoit 94) T. Dutoit. High quality text-to-speech synthesis: A comparison of four candidate algorithms. In *Proceedings of ICASSP 94 (1)*, 1994.

(Ganchev *et al.* 03) Hristo Ganchev, Stoyan Mihov, and Klaus U. Schulz. One-letter automata: How to reduce k tapes to one. Technical Report CIS-Bericht, Centrum für Informations- und Sprachverarbeitung, Universität Munchen, 2003.

(Gerdemann & vanNoord 99) Dale Gerdemann and Gertjan van Noord. Transducers from rewrite rules with backreferences. In *Proceedings of EACL 99*, 1999.

(Kaplan & Kay 94) Ronald Kaplan and Martin Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, 1994.

(Laporte 97) Eric Laporte. Rational transductions for phonetic conversion and phonology. In *E. Roche and Y.Schabes eds., Finite-State Language Processing.* MIT Press, 1997.

(Mihov & Schulz 04) Stoyan Mihov and Klaus U. Schulz. Efficient dictionary-based text rewriting using sequential transducers. *Natural Language Engineering*, Submitted, 2004.

(Moulines & Charpentier 90) E. Moulines and F. Charpentier. Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9(5-6), 1990.

(Roche & Schabes 97) Emmanuel Roche and Yves Schabes. *Finite-State language processing (Introduction).* MIT Press, 1997.

(Santen *et al.* 97) Santen, Sproat, Olive, and Hirschberg. *Progress in Speech Synthesis.* Springer-Verlag, 1997.

(Sproat 97) Richard Sproat. *Multilingual Text-to-Speech Synthesis, the Bell Labs Approach.* Kluwer, 1997.

(Totkov *et al.* 03) G. Totkov, D. Blagoev, and V. Angelova. Towards bulgarian text-to-speech system. In *Proc. of the Int. Conference "10 years Computer Systems Dept."*, 2003.

(Zahariev 93) Borislav Zahariev. *Microcomputer Systems for Text-to-Speech Synthesis.* Unpublished PhD thesis, Bulgarian Academy of Sciences, 1993.