# RUSSIAN TEXT-TO-SPEECH SYNTHESIS SYSTEM FOR MOBILE TELEPHONES

*Liliya Tsirulnik and Dmitry Pokladok*

*United Institute of Informatics Problems of the National Academy of Sciences of Belarus*
*l.tsirulnik@newman.bas-net.by*

The paper is devoted to development of Russian TTS-synthesis system for mobile telephones, which are characterized by small memory capacity and low operating speed. The suggested architecture of the system includes processing of the incoming text on the server and processing of voice database and generation of speech signal on the telephone. The peculiarities of Russian TTS-synthesis for mobile telephones are shown. The created system generates speech signal in real time on devices with ARM-processors with CPU clock frequency 100 MHz.

## 1 Introduction

The TTS-synthesis systems are applied in many practical applications, such as call centers, compound object control, electronic book reading, etc. Using of TTS-synthesis system is the most cost efficient way for creating and playing audio-books. It became more popular recently and is applied to variety of devices including mobile telephones. The restricted memory capacity and low speed of mobile telephones requires changing the architecture of the TTS-synthesis system. The paper describes the traditional and modified architectures of the TTS-synthesis system.

## 2 General Structure of TTS-synthesis system

The TTS-synthesis system [1] consists of two main parts (fig. 1): the text processing module and speech signal processing module. On the first stage the incoming orthographic text is transformed into a sequence of prosodic phrases with intonation phrase type indicators, where each phrase is represented by a sequence of allophones (the phoneme tint in speech flow). On the second stage the corresponding allophones natural waves (ANW) are extracted from the ANW DB, the target prosodic parameters: fundamental frequency ($F_0$), amplitude ($A$) and duration ($T$) for each allophone are calculated, the ANWs are modified in accordance with target prosodic values and concatenated into continuous speech signal.

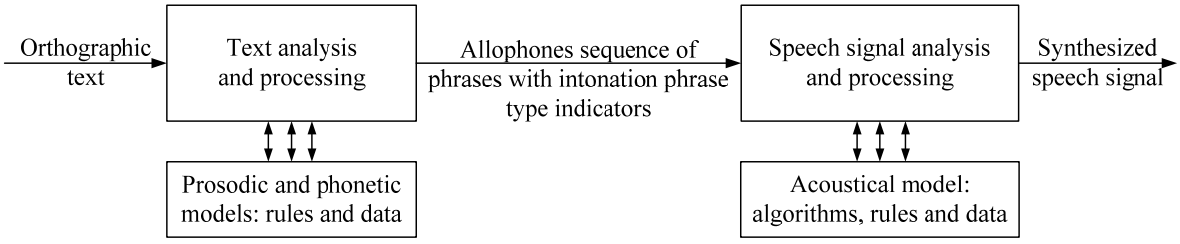


**Figure 1** – General structure of TTS-synthesis system

The text analysis and processing stage (fig. 2) contains linguistic, prosodic, and phonetic processing of text. The linguistic and prosodic processing includes dividing an orthographic text into utterances; transforming numbers, abbreviations, contracted forms; dividing an utterance into phrases; placing word stresses; dividing phrases into accentual units (AU[1]); marking the intonation type of the phrases. The main resources of linguistic and prosodic

[1] Accentual unit is a word or group of words with one strong stress.

processing module are grammatical dictionary and morphology and syntax rules. The dictionary is used to identify stress position and grammatical characteristics of each word of text. The morphology and syntax rules are used to divide the text into utterances, the utterances – into phrases, the phrases – into AU, and identify intonation types of phrases.

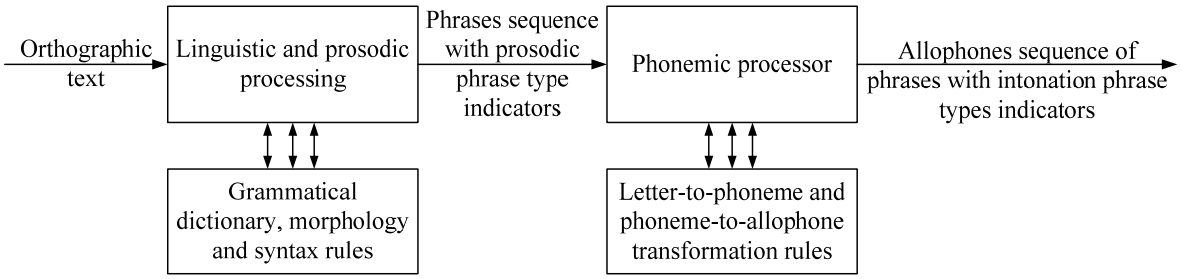


**Figure 2** – Structure of text processing module

Then every prosodically marked phrase is sent to the phonemic processor that performs the following tasks: phonemic transcription of the orthographic text; determination of positional and combinatory allophones from the incoming phonemic text; generation of the allophone and multi-allophone sequences that are necessary to synthesize.

The result of text processing module, which is a sequence of phrases with marks of intonation type of each phrase, where each phrase is represented by the sequence of allophones, is coming to speech signal processing module.

The speech signal processing module (fig.3) uses the Accent Unit Portraits (AUP) prosodic model [2] to calculates the target prosodic values: $F_0$, $A$, $T$ for each element. According to the AUPs model, accentual units are split into three parts: pre-nucleus (all the allophones, preceding the fully stressed vowel), nucleus (the fully stressed vowel) and post-nucleus (all the allophones, following the fully stressed vowel). Then the allophones natural waves, corresponding to allophone names in input sequence, are retrieved from ANWs DB and the Accent Unit Portrait for appropriate type of phrase are retrieved from AUPs prosodic elements DB, and target values of $F_0$, $A$, $T$ are calculated. Because the target $F_0$ values should be calculated for every pitch in vocalized allophones, consequently, it is necessary to know the number of pitches in each allophone, the calculation are performed after retrieving the ANWs from the ANWs DB.

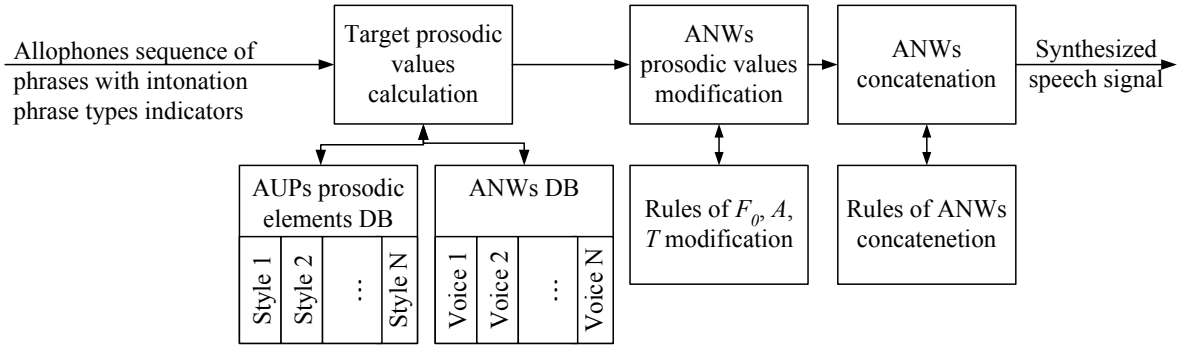


**Figure 3** – Structure of speech signal processing module

The rules of ANWs prosodic values modification are implemented as acoustic module resources. After the modification of $F_0$, $A$ and $T$ values the ANWs are concatenated, forming the synthesized speech signal.

## 3 Peculiarities of TTS-synthesis system realization for Russian and mobile telephones restrictions

Russian is an inflectional language and characterized by non-regular word stress position. For example, nouns in Russian have 12 wordforms, adjectives have 30 wordforms; both nouns and adjectives have more than 10 different schemes of word stress placing in wordforms. Obviously, to place word stress position in TTS-synthesis it is essential to use grammatical dictionary, contains maximum possible number of wordforms with indicators of word stress positions. The basic Russian dictionary [3] contains 100 000 paradigms, that correspond to about 2 000 000 wordforms. As the experience shows, this number of paradigms is not enough to process unrestricted Russian text. The dictionary of about 3 500 000 entries is sufficient to process most of present texts. Representation of dictionary as two-level structure [4], where the first level contains invariable parts of words and second level – flexions, provides minimization of utilized disc space and requires 70 Mb for 3 500 000 wordforms. Dictionary search operations use hash-tables and provide computational complexity equal to $O(n)$, where $n$ is the number of dictionary entries.

All the operations, performed on linguistic, prosodic and phonetic processing of text, including dictionary search operations, have computational complexity that is equal to $O(m)*O(n)$, where $m$ is the number of words in input text.

But in estimation of TTS-synthesis algorithms very important to take into account the clock frequency of the device, because the average time of processing of one phrase should be much less, than time of playing a synthesized phrase, which lies in the range from 1 to 10 seconds. The time of one phrase processing on personal computer with CPU clock frequency of 1,3 GHz is about 0.4-0.5 seconds.

Most of current mobile telephones have the following characteristics: available memory capacity from 128 Kb to 4 Mb, 32-bits RISC-processor with CPU clock frequency from 50 MGz and higher, Java ME programming language and CLDC configuration support. This characteristics are insufficient for implementation of the full-fledge TTS-synthesis system, but enough for implementation of the speech signal processing module.

## 4 Modified TTS-synthesis system architecture

Subject to characteristics of mobile telephones, described above, and taking into account the main purpose of development of TTS-synthesis system, which is creating and playing audio-books, the TTS-synthesis system modules were divided into two parts: the text processing module is implemented on server, and the speech signal processing module is implemented on mobile telephone (fig. 4).

The major alteration in the architecture is the calculation of target prosodic values on text processing stage, that performed on server. Because target prosodic values depend on the particular ANWs DB and the ANWs DB itself is placed on the mobile telephone, the text processing module resources are complemented by list of ANWs DB entries with duration of every allophone and number of pitches in every vocalized allophone.

Calculation of target prosodic values on server reduces the demand of mobile telephone resources due to AUP's prosodic elements DB is placed on server, and reduces time necessary for data processing on mobile telephone. On the other hand, the allophones sequence of phrases with target prosodic values, calculated on server, is specific for particular ANW's DB and particular prosodic style. It can be processed and played many times on mobile telephone, but in changing a voice (ANWs DB) or prosodic style the allophone sequence of phrases should be generated and sent to mobile telephone again.

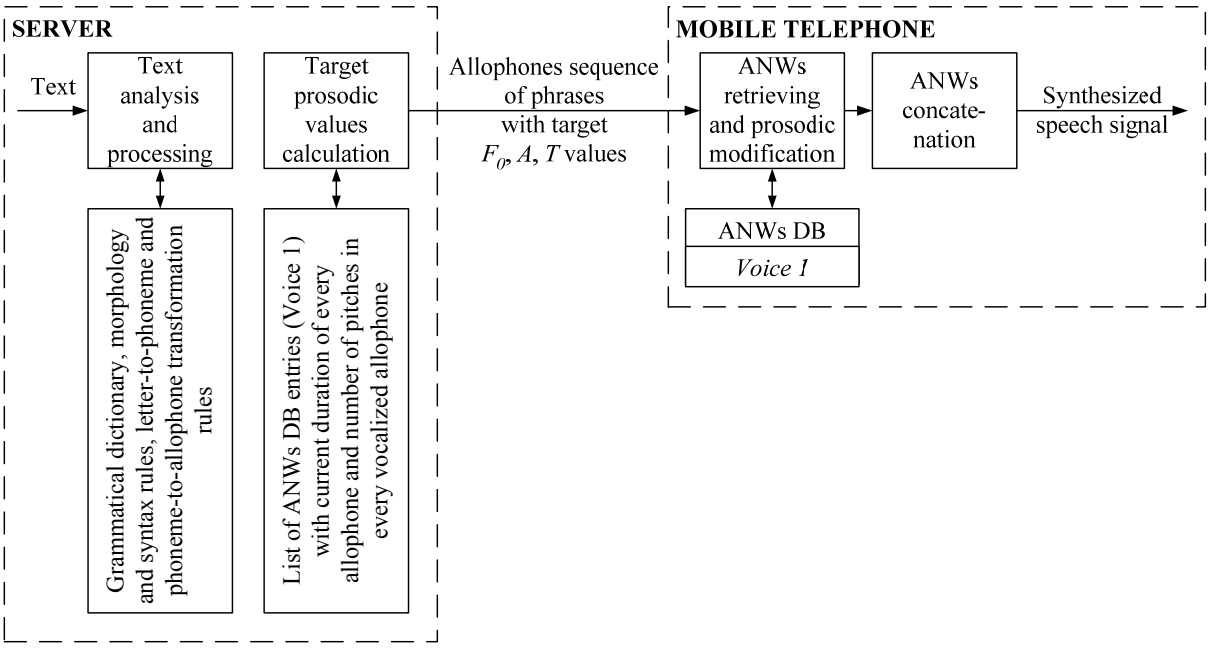


**Figure 4** – "Server – mobile telephone" architecture of TTS-synthesis system

## 5 Peculiarities of acoustical module realization on mobile telephones

The Java MIDlet on mobile telephone is implemented as three-thread application. The main thread is used to read the incoming sequence and synchronize two other threads; first one performs the incoming allophone sequence and speech signal processing, the second one is used to play the synthesized speech. The main thread postpones the second thread until the next phrase is synthesized and resumes it again, as well as postpones the first thread, while the next portion of synthesized speech is taken by second thread.

The processing of text is performed cyclically, phrase-by-phrase. The portion of text is read by main thread and placed to the 3000 bytes buffer, and then the allophonic phrase boundary is defined. The allophonic phrase is sent to the processing thread. On the next iteration the phrase found is deleted from the buffer and the next portion, complemented the buffer to 3000 bytes, is read.

The allophone sequence and speech signal processing thread utilizes the Soft Lacing Method [5] for $F_0$ modification that provides smooth pitch transition and preserves personal voice acoustic characteristics on speech signal modification. The process of period reducing is shown on fig. 5 and fig. 6. The deleted part of the period with length $N = T_0 - T_0'$, where $T_0$ is current period length, $T_0'$ is target period length, is shifted and overlapped to the preceding part of the period (fig.5). The overlapping of two parts is performed by continuous damping the first signal and amplification of the second signal, as it is shown in fig. 6.

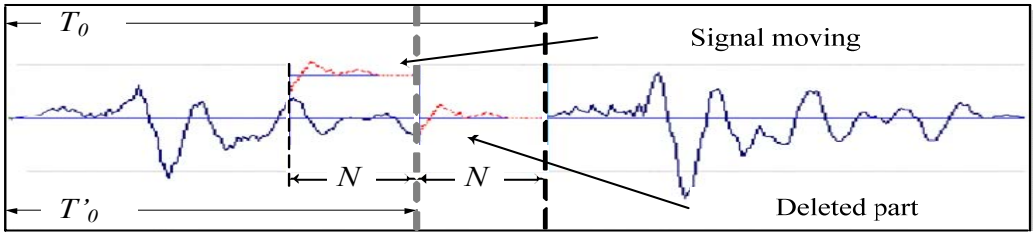


**Fig. 5** – Moving of the deleted part of signal

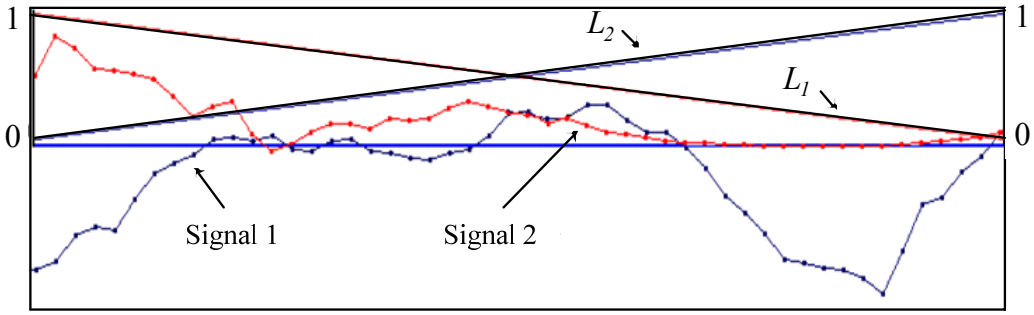


**Fig. 6** – Generation of Soft Lacing-transition sections

Damping of the first signal is calculated using the formula:

$$S'_1(n) = S_1(n)L_1(n), \quad \text{where } L_1(n) = 1 - \frac{n}{N}, \quad 0 \le n \le N \tag{1}$$

where $S'_1(n)$ is the modified signal 1, $S_1(n)$ is the initial signal 1, $N$ denotes the size of smooth transition frame.

Amplification of the second signal is expressed by the formula:

$$S'_2(n) = S_2(n)L_2(n), \quad \text{where } L_2(n) = \frac{n}{N}, \quad 0 \le n \le N \tag{2}$$

where $S'_2(n)$ is the modified signal 2, $S_2(n)$ is the initial signal 2.

For the calculation of the summarized signal of transition section $S'(n)$ the following formula is used:

$$S'(n) = S'_1(n) + S'_2(n), \quad 0 \le n \le N \tag{3}$$

The process of period lengthening is performed similarly.

The synthesized speech playing is performed with help of Java Player class [6]. This class can work with a number of sound formats, including WAVE PCM. This standard is supported by MIDP 2.0 and Sun JavaTM Wireless Toolkit for CLDC development environment, which was used for MIDlet development. All the allophones natural waves are stored in DB in WAVE PCM format with the following characteristics: number of channels – 1 (mono); frequency – 16 KHz, bits – 8.

## 6 Results and Conclusion

The implemented system was successfully tested on mobile telephones of the following models: Motorola, Sony-Ericsson, LG, that are characterized by CPU clock frequency from 68 to 115 MHz and memory capacity from 3 500 to 4 200 Kb. The system can synthesize speech signal in real-time on mobile telephones with ARM-processors of seventh generation. The certain disadvantage of the TTS-system is the necessity of incoming text re-processing on server in changing of prosodic style or voice DB.

## References

[1] Lobanov B.M., Tsirulnik L.I. “TTS-synthesis and Voice Cloning” [in Russian], Minsk, Belorusskaya Nauka, 2008, 342 p.

[2] Lobanov B., Karnevskaya E. “Auditory Estimation of Effectiveness of the AUP-Stylization Model of the Melodic Contour TTS-synthesis and Voice Cloning”, , Proc. 13-th International Conference “Speech and Computer” SPECOM’2009, June 21-25, 2009, St.-Petersburg, Russia, pp. 130-135.

[3] Zalizniak A. “Grammatical dictionary of Russian language” [in Russian], Moscow, Russian Language, 1987, 880 p.

[4] Dmitry V. Zhadinets, Oleg G. Sizonov, Liliya I. Tsirulnik “Russian and Belarusian Electronic Dictionaries for a Bilingual Text-to-Speech synthesis System” [*in Russian*], Proc. Third Scientific Conference “Tanajev’s readings”, March 28, 2007. Minsk, Belarus, pp. 65-69.
[5] Boris M. Lobanov, Liliya I. Tsirulnik, Dmitry V. Zhadinets, Elena B. Karnevskaya “Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis”, Proc. Third International conference “Speech Prosody’2006”, May 2–5, 2006, Dresden, Germany, V. 2, pp. 553-556.
[6] Mobile Media API (JSR-135) [electronic resource] – http://java.sun.com/javame/reference/apis/jsr135/.