The Algorithms of Tajik Speech Synthesis by Syllable

Khurshed A. Khudoyberdiev[1*]

[1]Khujand Polytechnic institute of Tajik technical university named after academician M.S. Osimi, Khujand, 735700, Tajikistan

Abstract. This article is devoted to the development of a prototype of a computer synthesizer of Tajik speech by the text. The need for such a synthesizer is caused by the fact that its analogues for other languages not only help people with visual and speech defects, but also find more and more application in communication technology, information and reference systems. In the future, such programs will take their proper place in the broad acoustic dialogue of humans with automatic machines and robotics in various fields of human activity. The article describes the prototype of the Tajik computer synthesizer by the text developed by the author, which is constructed on the principle of a concatenative synthesizer, in which the syllable is chosen as the speech unit, which in turn, indicates the need for the most complete description of the variety of Tajik language syllables. To study the patterns of the Tajik language associated with the concept of syllable, it was introduced the concept of "syllabic structure of the word". It is obtained the statistical distribution of structures, i.e. a correspondence is established between the syllabic structures of words and the frequencies of their occurrence in texts in the Tajik language. It is proposed an algorithm for breaking Tajik words into syllables, implemented as a computer program. A solution to the problem of Tajik speech synthesis from an arbitrary text is proposed. The article describes the computer implementation of the algorithm for syncronization of words, numbers, characters and text. For each syllable the corresponding sound realization is extracted from the "syllable-sound" database, then the sound of the word is synthesized from the extracted elements.

1 Introduction

Today speech synthesis is implemented by various methods that have both certain advantages and disadvantages. Speech synthesis is evaluated according to two characteristics - the naturalness of sound and the intelligibility of the speech it reproduces. Some speech synthesizers better convey the naturalness of sound, others - intelligibility. Depending on the purpose for which they are intended, various methods of speech synthesis are laid at the heart of their design. These methods are usually divided into three groups.

* Corresponding author: tajlingvo@gmail.com

1. Articulation synthesis is considered one of the most difficult methods. Its representatives [1-3] try to numerically simulate the work of the human larynx and the articulatory processes occurring in it as accurately as possible in order to reproduce high-quality synthetic speech. Until recently, articulatory synthesis developed mainly for scientific purposes and did not attract much attention from commercial organizations. And only recently, some of the developed models began to appear in speech synthesized systems. A definite idea of earlier and later models of articulation synthesis can be obtained from [4–5].

2. Formant synthesis, without using any samples of human speech, imitates it, producing artificial spectrograms. The speech message of synthesized speech is created by him using an acoustic model. Parameters such as natural frequency, sonication and noise levels vary over time and create a waveform of artificial speech. Many systems, which are based on formant synthesis technologies, generate artificial speech with a "robot-like" sound, so the synthesized speech message cannot be confused with natural human speech. Formant synthesis systems have some advantages over concatenative systems because, firstly, formant-synthesized speech can be very understandable in them because there are no acoustic noises inherent in concatenative systems. Secondly, formant synthesizers are often programs that are smaller in size than concatenative systems, since they do not have a base for speech samples. They can be used in embedded computer systems that require minimal memory and processor power. And finally, since formant synthesis exercises general control over all aspects of the created speech message, its achievement can be a wide variety of prosody (pronunciation systems of stressed and unstressed, long and short syllables in speech) or intonation, which conveys not only questions and statements, but and a spectrum of emotions and tones of voice. The most famous of the formant synthesizers are associated with the name of Klatt (D. H. Klatt [6 - 10]).

3. Concatenative (concatenation) synthesis uses pre-recorded segments of natural speech. Such a synthesis is probably the easiest way to reproduce understandable and naturally-sounding synthetic speech. In it, one of the most important points is the selection of sound bites of suitable length. This choice is made between short and long-sounding units. With longer units, good articulation and a high degree of naturalness of speech are achieved, the number of required connections at the docking points of sound units is reduced. At the same time, a drawback also appears - the inevitable increase in the initially reserved computer memory. Working with shorter sound units (fragments) requires less memory, however, the process of automatically synthesizing them becomes more difficult and complex. Existing concatenative synthesizers use phonemes, diphones, syllables, morphemes, words, phrases, and even sentences as sound units. At first glance, it might seem that in comparison with others, a word should be given preference, however, due to the presence in each language of an immense set of different words and proper names, and also because of the uneven sound of the word in continuous speech and in isolation, one cannot recognize such a choice is acceptable.

The ideas underlying the concatenative synthesis, apparently, were first expressed by Harris (S.M. Harris) in his article on the building blocks of colloquial speech, see [11]. The current status of the issue can be obtained from the works of Potapova R.K. [12 - 13].

The most common variants of concatenative synthesis are **parametric** synthesis and synthesis **according to the rules**. The first of them is more flexible due to the parameterization based on small phonetic units (allophones, diphons, syllables ...). It allows you to manipulate the parameters that are responsible for the quality of speech (formant value, bandwidth, fundamental frequency, signal amplitude). This makes it possible to glue the signals, so that the transitions at the borders become invisible. Varying parameters such as the frequency of the fundamental tone throughout the message make it possible to significantly change the intonation and temporal characteristics of the message. For the

synthesis, speech units of various lengths are used: paragraphs, sentences, phrases, words, syllables, half syllables, diphons. The smaller the unit of synthesis, the smaller their number is required for synthesis. This requires more computation, and there are difficulties in co-articulation at the joints. The advantages of this method: flexibility, a little memory for storing the source material, preserving the individual characteristics of the speaker.

Synthesis **according to the rules** works with the so-called "unlimited dictionary". Its elements are phonemes or syllables, which are connected according to well-defined rules. It was found that for high-quality speech synthesis it is necessary to have several different pronunciations of the synthesis unit (for example, a syllable), which leads to an increase in the dictionary of the original units without any information about the context situation. For this reason, the synthesis process acquires an abstract character and moves from a parametric representation to the development of a set of rules by which the necessary parameters are calculated based on an introductory phonetic description. This introductory presentation contains little information per se. These are usually the names of phonetic segments (for example, vowels and consonants) with accent marks, tone designations, and temporal characteristics. This method provides freedom for modeling parameters, although the modeling rules themselves remain imperfect. Synthesized speech is worse than natural, however, it satisfies the tests of intelligibility and comprehensibility.

It should be noted that among the syntheses mentioned, formant and concatenative have found widespread use, the first of which has dominated for a long time in the past, but concatenative synthesis is becoming more popular today. Against this background, articulatory synthesis seems too complicated for high-quality reproduction, but it is possible that it may turn out to be a particularly promising method in the near future.

Other less popular speech syntheses are hybrid and HMM-based synthesis (HMM). Hybrid synthesis combines the features of formant and concatenative synthesis in order to minimize acoustic noise in the process of sounding speech segments. In a synthesis system based on HMM, the speech frequency spectrum (speech path), natural frequency (speech synthesizer) and duration (prosody) are simulated simultaneously using hidden Markov models. Speech waveforms are generated from hidden Markov models, which in turn are based on maximum likelihood criteria.

In Russia, the most notable achievements in the field of automatic speech synthesis are associated with the Computing Center of the Russian Academy of Sciences (Yu. I. Zhuravlev [14], V. Ya. Chuchupal [15]); Institute of Information Transmission Problems of the Russian Academy of Sciences (V. N. Sorokin [16]), Institute of Mathematics Siberian branch of RAS and Novosibirsk State University (Velichko V. M., N. G. Zagoruiko [17]), Moscow State University named after M.V. Lomonosov (L.V. Zlatoustova, S.V. Kodzasov, O.F. Krivnova, I.G. Frolova [18]), BMSTU (Kharlamov A.A., Zhigulevtsev Yu.N. [19]). In Belarus, certain achievements are presented in the works of Lobanov B.M. et al., [20 - 22].

Various methods of speech synthesis are the basis of computer programs - speech synthesizers. At the request of the user, such programs belonging to the category "text-to-speech" can read texts recorded in electronic memory by male or female, make intonation pauses, change the tone and timbre of speech during listening, and transmit voiced texts through the network. Here is a list of the most famous computer speech synthesizers: Reader TTS, Govorilka, ToM Reader, Sakrament, Talk-To-Me, Text Aloud MP3, SNAT, Book Reader, Speech2, Phonemafon, MP3book2005, Sakrament Talker, Infovox, DECTalk, Bell Labs Text- to-Speech, Laureate, SoftVoice, CNET PSOLA, ORATOR, Eurovocs, Lernout & Hauspies, Apple Plain Talk, Acu Voice, CyberTalk, ETI Eloquence, Festival TTS System, ModelTalker, MBROLA, Whistler, NeuroTalker, Listen2, SPRUCE, HADIFIX, SVOX Pfister 1995. SYNTE2 and SYNTE3, Timehouse Mikropuhe, Sanosse, Speaking Mouse, ARGUS, AGAFON.

Some programs, such as Sakrament Talker, Govorilka, Talk-To-Me, Text Aloud, Speech2, are reportedly adapted to read texts in any language aloud. However, when working directly with them, it is discovered that the skill attributed to them is not actually confirmed, since the high quality of the synthesized speech is directly related to the specifics of the spoken language, as a result of which the software system developed for a particular language cannot be equally successful its functions in relation to any other language. However, not only this, but also significant shortcomings, determined either by the unnatural sound, or insufficient intelligibility of messages, determines the relevance of further research on the design of speech synthesizers for natural languages.

2 The syllabic structure of the words of the Tajik language

A syllable, by definition, is called a minimal pronunciation unit of speech, consisting of one or more sounds that form a close phonetic unity. According to a slightly different equivalent interpretation, *a syllable* is a sound or a combination of sounds in a word, pronounced with one push of exhaled air.

To study the patterns of the Tajik language associated with the concept of a syllable, we introduce an additional concept of the *syllable structure of a word.*

Let W be a word representing a certain sequence of letters. Replacing vowels in it with the number 1, and consonants with the number 0 (we consider the letter “й” to be consonant), we thereby transform the word W into an ordered collection $W_{0,1}^{*}$ of zeros and ones. We call such a transformation the encoding of the word W, and the result obtained, i.e. notation $W_{0,1}^{*}$, - *the syllabic structure of the word* W.

The dimension of the structure $W_{0,1}^{*}$ is the number of letters that make up the word W, or the number of characters (binary characters) that are used in the record $W_{0,1}^{*}$. The structures of two words are called *the same* if their representations in binary notation are identical, otherwise they are *different*. It is clear that the structures can be the same only if they have the same dimension. It is also obvious that for every word W one and only one image is associated with $W_{0,1}^{*}$. In turn, essentially for any natural language, to any $W_{0,1}^{*}$ several words W simultaneously corresponds. This means that different words with the same number of letters can have the same syllabic structure. For example, the words "дилшод", "кардам", etc. corresponds to the same structure “010010”.

The results formulated hereinafter are based on the statistical processing of a representative sample composed of fragments of the works, which amounted to 1800000 words. In the future, the images of these words, i.e. the corresponding syllabic structures represented by the set of $W_{0,1}^{*}$ became the object of statistical analysis.

1. On the set $W_{0,1}^{*}$ 2978 different syllabic structures of Tajik words were found, with 1 and 14 being the dimensions of the minimum and maximum word structures, respectively.

2. The statistical distribution of structures is obtained, that is, a correspondence is established between the syllabic structures of words and the frequencies of their occurrence in texts in the Tajik language.

3. It was found that 8 structures provide 50% and 23 structures - 75% coverage of Tajik texts (see table 1, part 1).

These data are presented as follows, the first column gives the number of the structure in decreasing order of frequency of its occurrence, in the second - the record of the structure itself and in the third - the percentage of its occurrence in the texts.

Table 1. Part 1. The frequency of occurrence of words in the form of syllabic structures.

№	$W_{0,1}^{*}$	%		№	$W_{0,1}^{*}$	%		№	$W_{0,1}^{*}$	%
1	01	11,006		9	010010	3,684		17	1010	1,192
2	010	8,849		10	0101010	3,258		18	01001010	1,142
3	01010	6,781		11	0100	2,799		19	010100	1,087
4	01001	5,486		12	01010101	1,735		20	01001011	1,053
5	10	5,096		13	01011	1,711		21	100	0,986
6	0101	5,066		14	1001	1,280		22	10101	0,960
7	010101	4,773		15	010011	1,226		23	10010	0,957
8	0100101	3,787		16	0101001	1,218				

4. It was also established that 51 structures carry out 90%, and 76 structures - 95% coverage of Tajik texts (see table 1, part 2).

Table 1. Part 2.

№	$W_{0,1}^{*}$	%		№	$W_{0,1}^{*}$	%		№	$W_{0,1}^{*}$	%
24	0101011	0,923		42	011	0,421		60	0110101	0,190
25	01010010	0,895		43	010101011	0,404		61	0101001010	0,189
26	1	0,875		44	0101101	0,366		62	010111	0,189
27	010010101	0,869		45	010010011	0,348		63	0100101001	0,189
28	100101	0,810		46	101010	0,321		64	101001	0,188
29	01010100	0,734		47	0101100	0,317		65	0100110	0,187
30	010100101	0,717		48	1010101	0,306		66	1001011	0,185
31	0110	0,716		49	010101001	0,289		67	01001101	0,173
32	01001001	0,660		50	011010	0,281		68	01010010101	0,172
33	010101010	0,601		51	0100100101	0,279		69	1001001	0,170
34	101	0,556		52	0101010101	0,278		70	01001100	0,166
35	01101	0,554		53	101011	0,257		71	0101010100	0,164
36	010110	0,549		54	01010110	0,254		72	010001	0,160
37	0100100	0,533		55	010010010	0,248		73	0101001011	0,158
38	10010101	0,468		56	0100101011	0,244		74	010101101	0,144
39	0101010010	0,443		57	010100100	0,223		75	01000101	0,143
40	1001010	0,438		58	10011	0,195		76	0100101010	0,141
41	01010011	0,432		59	0100011	0,193				

5. Each of the 274 discovered syllable structures of Tajik words was divided into syllables “manually” (in accordance with the division into syllables of those Tajik words that fell under one or another structure). As a result, only 9 different syllable structures were discovered -

1, 10, 01, 010, 100, 0100

and

001, 0010, 00100.

Of these, the first six are inherent in the nature of the Tajik language, and the last three are borrowed from other languages.

The frequency (in percent) of the mentioned structures in the processed text information is shown in table 2.

Table 2. Frequency of syllables in a character record (in%).

Syllables	1	10	01	100	010	0100	001	0010	00100
Frequency	8.10	5.74	56.56	0.78	25.75	2.95	0,05	0,06	0,01

It can be stated that the two-letter syllables are of the type 'да', 'ба', 'ро', 'на', 'ни', 'та', 'ме', 'ва', 'ки' (in the symbolic notation - 01) and etc. are the most common, and three-letter syllables like ‘абр’, ‘илм’, ‘ишк’, ‘умр’, ‘орд’ (in a symbolic record - 100), etc. - especially rare. In addition, syllables 001, 0010 and 00100, borrowed from other languages, occasionally appear in Tajik texts (in total - 0.12%).

3 Automatic word decomposition

This article provides a conceptual description of the sequence of procedures, the implementation of which in the form of a computer program allows automatic separation of an arbitrary Tajik word into syllables. The separation process is based on the concept of the syllable structure of a word and essentially uses 6 syllable structures.

Let W - be a Tajik word representing a certain sequence of letters of the Tajik alphabet, and $W_{0,1}^{*}$ - be a syllable structure of the word W, i.e. encoded record W as a set of zeros and ones. Recall that $W_{0,1}^{*}$ is obtained from W by replacing in W consonants with 0 and vowels with 1.

The proposed algorithm consists of two parts: in the first part, the division $W_{0,1}^{*}$ into syllable structures is carried out, in the second part the result obtained is used directly to represent the original word W in the form of an ordered collection of syllables.

Part 1. So, in the Tajik language there are 6 syllable structures - 1; 10; 01; 010; 100; 0100. In the first part of the algorithm, which divides $W_{0,1}^{*}$ into syllable structures, the following procedures are performed.

1. Getting started.
2. Enter the word W.
3. Performing conversion $W \to W_{0,1}^{*}$.
4. Counting the number k units in a record $W_{0,1}^{*}$. Since vowels are encoded with the number 1, the number k essentially indicates the number of syllables making up the word W.

5. If k = 1, then, obviously, the entry $W_{0,1}^{*}$ consists of one syllable, and this syllable is identified by identifying $W_{0,1}^{*}$ with one of the 6 previously mentioned syllables. Next, go to step 9. If, however, $k \neq 1$, then go to step 6.

6. If k = 2, then the entry $W_{0,1}^{*}$ consists of two syllables. Which syllables make up $W_{0,1}^{*}$ is determined by identifying $W_{0,1}^{*}$ with one of the various entries made up of two syllable structures and obtained by attaching one of 6 structures to each of the 6 syllable structures. Obviously, out of 6 syllable structures, 36 such paired combinations can be composed. Next, go to step 9. If, however, $k \neq 2$, then go to step 7.

7. If k = 3, then the entry $W_{0,1}^{*}$ consists of three syllables. Which syllables make up $W_{0,1}^{*}$ is determined by identifying $W_{0,1}^{*}$ with one of all possible entries made up of three syllable structures and obtained by attaching one of 6 structures to each of the 6 syllable structures and then another of 6 structures to the resulting record. Obviously, from 6 syllable structures, 216 such three-syllable combinations can be composed. Then go to step 9. If $k \neq 3$, then go to step 8.

8. The syllabic composition of entries $W_{0,1}^{*}$ for which $k > 3$ but at the same time $k \leq 8$ is recognized in the same way, because it is currently known that the Tajik language does not contain words containing more than 8 syllables.

9. The end.

The word “хуршед”, chosen by us as an example, in coding with the help of zeros and ones is identified with the 9th record of table 1. Therefore, in encoded form this word gets a syllable representation

$$W_{0,1}^{*} \text{ (“хуршед”)} = 010 \oplus 010, \quad (1)$$

where $\oplus$- is the sign of agglutination, i.e. joining (gluing) one syllable structure to another without a space.

Part 2. After decomposition $W_{0,1}^{*}$ into syllable structures, splitting the source word W is very simple. From the first part of the algorithm, it suffices to store in memory the number of letters that make up the 1st syllable, 2nd syllable, etc. These numbers are used to highlight syllables already in the original word W.

So, in the above example, when separating $W_{0,1}^{*}$ (“хуршед”), 2 syllables were obtained, and the first and second syllables consisted of 3 letters. Therefore, when dividing the word W = “хуршед” itself, we get the result “хур - шед”.

4 The variety of the types of the Tajik language

Based on the author's algorithm and a computer program developed on its basis, statistical studies on the variety of syllables of the Tajik language were carried out.

1. 3259 different syllables are extracted.
2. The statistical distribution of syllables in texts in the Tajik language is obtained, i.e. an empirical correspondence ν = ν (n) was established between the number of each of 3259 different syllables arranged in decreasing order of their occurrence frequencies and the frequency ν (in percent) of occurrence of the corresponding syllable corresponding to this number.
3. It is established that 41 syllables cover 50% of the Tajik text:

Table 3. Frequency of occurrence of Tajik syllables (in extended Cyrillic Unicode).

N	Syllable	ν	N	Syllable	ν	N	Syllable	ν
1	и	4,210	15	ди	1,277	29	ми	0,760
2	да	2,447	16	ки	1,189	30	би	0,727
3	ро	2,347	17	о	1,156	31	то	0,722
4	ба	2,235	18	мо	1,149	32	я	0,697
5	ҳо	2,022	19	до	1,112	33	ин	0,693
6	ни	1,827	20	ра	1,077	34	ҳа	0,673
7	на	1,796	21	ма	1,071	35	са	0,647
8	ти	1,665	22	аз	0,986	36	за	0,611
9	ри	1,612	23	му	0,968	37	ло	0,602
10	та	1,552	24	ли	0,951	38	во	0,562
11	ме	1,508	25	а	0,914	39	ла	0,552
12	ва	1,500	26	со	0,833	40	ё	0,548
13	бо	1,355	27	си	0,823	41	хо	0,523
14	дар	1,325	28	но	0,766			

4. It was also found that 148 syllables cover 75% of the Tajik text. The relevant data are given in the continuation of the table.

5. It was also found that 204 syllables cover 80% of the Tajik text, 418 syllables - 90%, and 683 syllables - 95% of the text. Note that all other syllables in the aggregate (from 684 to 3259) cover only 5% of the text. Consequently, the appearance of each individual syllable from such an aggregate is an extremely rare occurrence.

5 Structure of the software complex

This article sets out the basic idea of implementing a Tajik speech synthesizer in the text. A text is a sequence of sentences constructed according to the rules of a given language and a given sign system and forming a message. In turn, the proposal will be considered as a set of ordered 7 elements: words, numbers, special characters, space. Also, punctuation marks inside the sentence (comma, colon, semicolon, dash), external punctuation marks (dot, ellipsis, question and exclamation mark). And finally, the end of the paragraph (it is not in the written text, but it is present in the electronic text as an unprintable ¶ sign).

Note that the meaning that we include in the names of the elements should be understood in generally accepted meanings. We also emphasize that in a particular sentence some elements may be absent (for example, numbers, symbols, internal punctuation marks,

etc.) while the presence of others is mandatory (for example, an external punctuation mark). We will need 5 types of pauses used in speech:

p_s - pause between syllables when pronouncing a word;

p_w - a pause between words when reading a sentence (corresponds to a space between words);

p_i - a pause marking the internal punctuation mark;

p_e - a pause marking the external punctuation mark;

p_e - a pause marking the end of a paragraph.

A conceptual model for synthesizing speech in the text in the form of a flowchart is shown in Figure 1. The synthesizer operates as follows. After entering the next sentence, it is analyzed by the composition of its elements. If the next element is a word, then in block 1 it is divided into syllables indicating the stressed syllable and then it is voiced using the syllable-sound base.

If the next element is a number, then it in block 2 is converted into text and then its sounding occurs through block 1.

If the next element is a symbol, then its scoring takes place in block 3 by extracting the corresponding sound from the "symbol-sound" base.

If the next element is a space, an internal or external punctuation mark, or the end of a paragraph mark, then the corresponding pause is extracted from the corresponding block.

The speech synthesizer, presented in the form of a block diagram, suggests that it is based on the principle of concatenation of voiced syllables.

Since the syllable acts as the main sound unit of speech, for the implementation of the *synthesizer it is required to describe the variety of all syllables of the corresponding natural language.*

Since each syllable, presented in the form of a chain of letters, needs its sound image, *the creation of a "syllable-sound" base is required.*

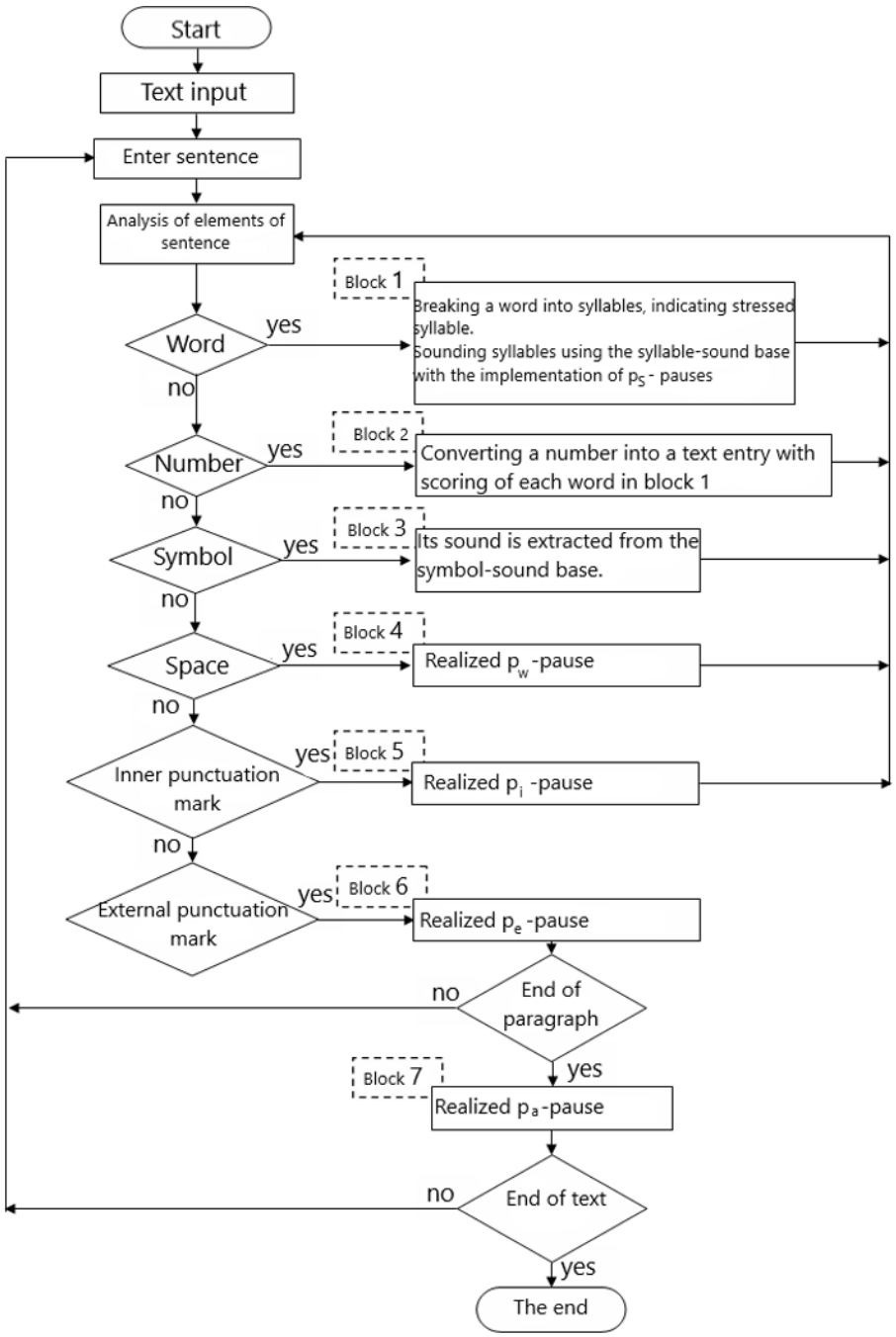

Fig. 1. Schematic diagram of the synthesis of speech in the text.

Since the synthesizer provides for the sounding of numbers and symbols, the corresponding algorithms and programs have been developed for the implementation of the synthesizer to transform the number into text and create a symbol-sound base.

And finally, you need to adjust the duration of the pauses p_s, p_w, p_i, p_e and p_a in such a way as to obtain, as far as possible, a natural and legible synthetic speech

3. Based on detailed elaborations of the conceptual scheme, the Tajik Text-to-Speech synthesizer was created, that is, a set of programs for synthesizing Tajik speech in the text. The synthesizer parameters were tuned by computational experiments. Satisfactory values of pause durations were established:

- for paragraph boundaries p_a = 900 ms,

- for offer boundaries p_e = 600 ms,

- for commas inside sentences p_i = 400 ms,

- for inter-word and inter-syllable pauses, accordingly, p_w = 200 ms и p_s = 20 ms.

To evaluate the synthesizer's performance, experiments were organized to voice a variety of textual information (fragments from novels, novels, scientific articles, textbooks, newspapers, magazines, Internet sites). The assessment of the completeness of the many syllables used to form synthetic speech was associated with the percentage of spoken words in relation to the total number of words within the selected text fragments. The results of the experiment showed quite satisfactory quality of the Tajik Text-to-Speech software package for scoring the Tajik text. The block diagram of the software package is presented in Figure 2.

In the first block, the "User Interface" consists of two components - "Text Entry" and "Speech", which have one-way communication, that is, the user has the opportunity to enter text information and as a result receive a speech version of the input text. To get the results, block 1 is connected with block 2 in two directions - to provide information for linguistic analysis and to obtain the results of scoring. Block 1 also interacts with block 3 directly to use the necessary data about the system settings (male or female voice selection, volume and speed of scoring).

The second block "Analytical subsystem" consists of two parts - "Linguistic analysis" and "Sound module". The first of them consists of the submodules "Text Validation", "Text Encoding" and "Separating Words into Syllables". "Text Validation" is used to validate input information, which includes text elements such as words, integers, characters, and punctuation marks. This submodule checks text elements, converts integers and characters into a test case, and then passes them for encoding. The coding process implements the submodule of the same name, which converts each word of the input text into an ordered set of zeros and ones, i.e. all words are represented by their syllabic structures. The encoded text is transmitted to the subdivision "Separation of words into syllables." Syllable words are linguistically analyzed and transmitted to the Sound Module. In this module, the formation of sound information occurs using the base "syllable-sound" of the information subsystem, stressed syllables, inter-syllable and inter-word pauses, as well as pauses marking such punctuation marks as a comma and period. The scoring module is the final stage of the analytical subsystem, and the audio version of the text information is sent to the user interface.

The third block, "Information Subsystem," contains databases called "System Settings" and "Syllable-Sound Base". The first of them is used to store temporary system setup data, the second "syllable-sound" base - to store statistical data on sound files of 3259 Tajik syllables. To work with this database, a module is used to provide access, check and select the necessary data.

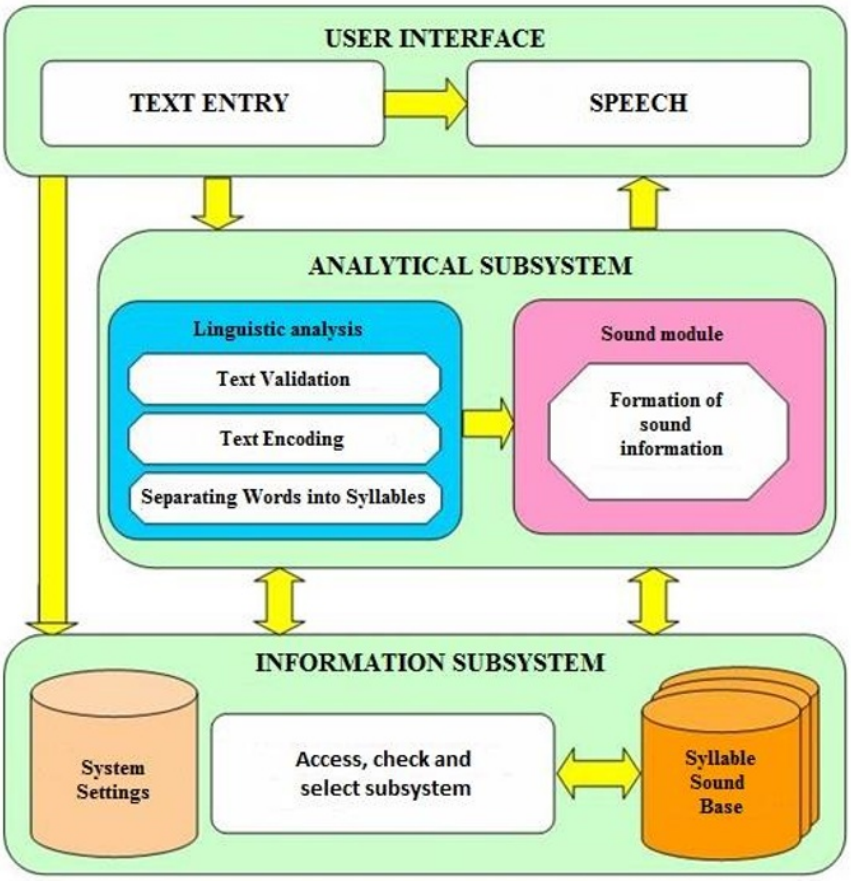

Fig. 2. Block diagram of Tajik Text-to-Speech.

6 Conclusion

Thus, the software package for computer dubbing of the Tajik text Tajik Text-to-Speech [24] and the announcer of the Tajik text Tajik Text-Narrator [25], although they do not completely solve the problem of synthesizing Tajik speech, are still the first software product, satisfactorily performing computer scoring of Tajik texts. At this level of development, the complex can now be used by people with impaired vision. The experiments were carried out at scientific seminars of the Khujand Polytechnic Institute of the Tajik Technical University named after Academician M.S. Osimi. Its participants, at their discretion, entered Tajik texts into the computer and then evaluated the naturalness and intelligibility of the sound of synthetic speech. The general opinion of the seminar - a computer synthesizer, built on the principle of concatenation of 3259 Tajik syllables, quite successfully performs the function of scoring Tajik texts. The synthesizer implements such elements of prosodic synthesis as the arrangement of stresses, taking into account the intonation pause between paragraphs, after the decimal point inside sentences and the point at the end of the sentence. Computational experiments have established the prospect of further development of the Tajik Text-to-Speech and Tajik Text-Narrator software systems into a Tajik speech synthesizer with Russian language.

References

1. J. Beskow, *Talking Heads - Communication, Articulation and animation*, Proceedings of Fonetik-96: pp. 53-56 (1996)
2. M. Cohen, D. Massaro, *Modelling Coarticulation in Synthetic Visual Speech*, Proceedings of Computer Animation 93, Suisse (1993)
3. B. Kröger, *Minimal Rules for Articulatory Speech Synthesis*, Proceedings of EUSIPCO92 (1), pp. 331-334 (1992)
4. J.L. Flanagan, *Speech Analysis, Synthesis, and Perception*, Springer-Verlag, Berlin-Heidelberg-New York (1972)
5. J.L. Flanagan, K. Ishizaka, and K.L. Shipley, *Synthesis of speech from a dynamic model of the vocal cords and vocal tract*, The Bell System Technical Journal, 54(3), pp. 485-506 (1975)
6. D.H. Klatt, *Synthesis by rule of segmental durations in English sentences*, In B.E.F. Lindblom and S. Ohman (Eds.), Frontiers of Speech Communication Research, pp. 287-299, Academic (1979)
7. D. Klatt, *Software for a Cascade/Parallel Formant Synthesizer*, Journal of the Acoustical Society of America, JASA, Vol. 67, pp. 971-995 (1980)
8. D.H. Klatt, *The Klattalk text-to-speech conversion system*, IEEE ICASSP-82, pp. 1589-1592 (1982)
9. D. Klatt, *Review of Text-to-Speech Conversion for English*, Journal of the Acoustical Society of America, JASA, **Vol. 82 (3)**, pp. 737-793 (1987)
10. D. Klatt, L. Klatt, Analysis, *Synthesis, and Perception of Voice Quality Variations Among Female and Male Listeners*, Journal of the Acoustical Society of America, JASA, **Vol. 87 (2)**, pp. 820-857 (1990)
11. C.M. Harris, *A study of the building blocks in speech*, Journal of the Acoustical Society of America, 25(5), pp. 962-969 (1953)
12. R.K. Potapova, *The main modern methods of analysis and speech synthesis*, Moscow (1971)
13. R.K. Potapova, *Syllabic phonetics of Germanic languages*, Moscow (1986)
14. Yu.I. Zhuravlev, *Selected scientific papers*, Master, Moscow (1998)
15. V.Ya. Chuchupal, K.A. Makovkin, A.V. Chichagov, *To the question of the optimal choice of the alphabet of the models of sounds of Russian speech for speech recognition*, Artificial Intelligence, **Vol. 4**, No. 1, Kiev, pp. 575-579 (2002)
16. V.N. Sorokin, I. S. Makarov, The inverse problem for a voice source, Information Processes, **Vol. 6**, No. 4, pp. 375-395 (2006)
17. V.M. Velichko, N.G. Zagoruyko, A.V. Kelmanov, S.A. Khamidullin et al, *Development of algorithmic, technical, and software recognition tools for isolated speech commands in a limited frequency range*, Report of Novosibirsk State University on research on the subject "M-91-86", state number, 01870014595, Novosibirsk, pp. 89 (1987)
18. L.V. Zlatoustova, S.V. Kodzasov, O.F. Krivnova, I.G. Frolova, *Algorithms for converting Russian spelling texts into phonetic recording*, Moscow, Moscow State University (1970)
19. A.A. Kharlamov, Yu.N. Zhigulevtsev, *Microprocessor-based tools for building embedded speech applications "Artificial Intelligence"*, No. 4 (2006)

20. B.M. Lobanov, *Analysis and synthesis of speech. Collection of scientific works*, Academy of Sciences of the BSSR Institute of Technical Cybernetics. Scientific Ed. B. M. Lobanov, Minsk, p. 86 (1991)
21. B.M. Lobanov, E.B. Karnevskaya, T.V. Levkovskaya, Text-to-speech synthesizer as a computer tool for "cloning" a personal voice, Tr. International Conference Dialogue-2001, Moscow, pp. 265-272 (2001)
22. B. M. Lobanov, L. I. Tsirulnik, D. V. Zhadinets, O. G. Sizonov, Algorithms for synthesizing prosodic characteristics of speech by text in the Multifon system, Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus, Minsk (2007)
23. Z.D. Usmanov, H.A. Khudoyberdiev, *The experience of computer synthesis of Tajik speech in the text (scientific monograph)*, Irfon, Dushanbe, p.145 (2010)
24. Z.D. Usmanov, Kh.A. Khudoyberdiev, *Computer dubbing of Tajik text Tajik Text-to-Speech*, A patent (intellectual product) is registered by the National Patent Information Center of Ministry of Economic Development and Trade of the Republic of Tajikistan, 041TJ from 09/04/2007 (2007)
25. Z.D. Usmanov, H.A. Khudoyberdiev, A.A. Khudoyberdiev, *Announcer of the Tajik text Tajik Text-Narrator*, A patent (intellectual product) is registered by the National Patent Information Center of Ministry of Economic Development and Trade of the Republic of Tajikistan, No. 4201800386 from 07/10/2018 (2018)