Integral Robot Technologies and Speech Behavior

Integral Robot Technologies and Speech Behavior

Edited by Alexander A. Kharlamov and Maria Pilgun

This book first published 2024

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-5436-8
ISBN (13): 978-1-5275-5436-8

CHAPTER EIGHT

RULE-BASED MULTI-WAVE SPEECH SYNTHESIS

BORIS LOBANOV

Introduction

The model of speech synthesis described in this chapter is based on the results of multi-year research aimed at the creation of linguo-acoustic foundations for speech synthesis from the text (Lobanov, 1983; Lobanov, 1987; Lobanov and Tsirulnik, 2006a). The model accumulates theoretical and experimental information about the specifics of linguistic processing of texts, the phonetic and prosodic structure of Russian speech, and the articulatory-acoustic phenomena of the speech formation process.

A distinctive feature of the described model, reflected in the name "multi-wave synthesis", is the use of natural speech wave segments as elements of speech compilation in correlation with elements of various phonetic lengths: allophones, diallophones and allosyllables.

8.1. Text-to-speech synthesizer structure

The synthesis of spoken language from text is implemented on the basis of the lexical and grammatical analysis of the input text by modeling the processes of speech formation, taking into account the rules of pronunciation for sounds and intonation inherent in this language. The orthographic form of a document (book, article, web-page, etc.) arrives at the input of the synthesizer and is then subjected to sequential processing by a number of specialized processors in accordance with the general structure of the text-to-speech synthesizer shown in Fig. 8-1. The synthesizer includes four main modules: a text processor, a prosodic processor, a phonetic processor, and an acoustic processor. Each of these modules is supported by sets of corresponding databases (DBs) and rules.

Consider the main functions of these modules.

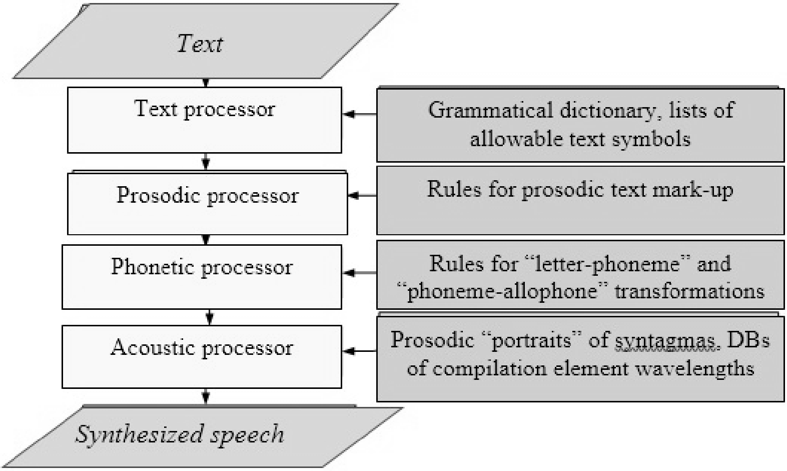

Text
Text processor
Grammatical dictionary, lists of allowable text symbols
Prosodic processor
Rules for prosodic text mark-up
Phonetic processor
Rules for "letter-phoneme" and "phoneme-allophone" transformations
Acoustic processor
Prosodic "portraits" of syntagmas, DBs of compilation element wavelengths
Synthesized speech

Figure 8-1. Text-to-speech synthesis system structure

8.2. Text processor

The text processor (Fig. 8-2) includes two main units supported by the corresponding databases, dictionaries and rules. It performs pre-processing of the input text, as well as morphological and accentual marking of the words in the text.

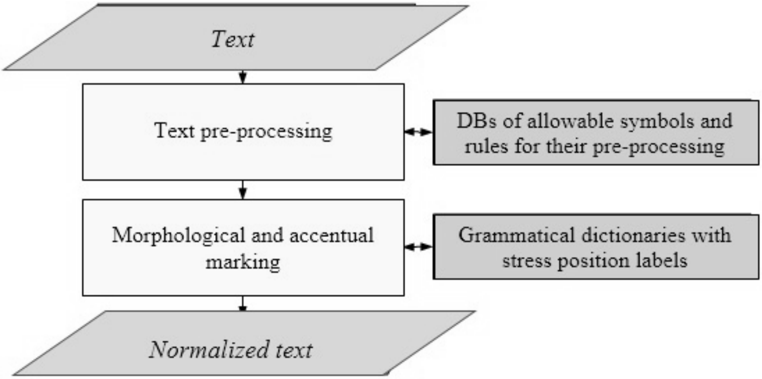

Text
Text pre-processing
DBs of allowable symbols and rules for their pre-processing
Morphological and accentual marking
Grammatical dictionaries with stress position labels
Normalized text

Figure 8-2. Text processor structure

The input of a speech synthesis system can receive texts taken from various sources and often containing graphic objects, links, numbers, formulas, as well as other objects and symbols unsuitable for speech synthesis. The main task of the first unit (the preprocessor) is text normalization, i.e. bringing it to such a form when the text consists of a sequence of words in the Russian language. The next unit (the unit of morpho-phonetic marking) marks each word of the input text, which is necessary for proper synthesis of sounds and intonation of speech. For such marking, a grammatical dictionary is used with stress positions marked in each word.

8.2.1. Preprocessor

The preprocessor structure is shown in fig. 8-3.

Text cleaning. Text cleaning is performed in order to remove graphic objects, links, various markers and other symbols from the input text that are not informative for speech synthesis. To implement this task, it is necessary to have a DB of valid symbols and objects containing Russian and Latin letters, punctuation marks, numbers, mathematical symbols, as well as special characters such as "@", "^", etc. In general, this DB should contain only those symbols that can be "voiced" by a speech synthesizer. For example, if the DB contains Roman numerals or complex mathematical symbols such as "Σ", "∫", then the subsequent stages of text processing should include units that convert sequences of these symbols into words.

It is noteworthy that this unit, from the developer's standpoint, does not present any particular difficulty or particular interest, and in most cases, when developing speech synthesis systems, it is implemented last. For users of the speech synthesis system, this unit, on the contrary, is very important, since the completeness of the "voicing" of the input text depends on the algorithms of its operation.

Number decoding. The task of this unit is to convert the numbers found in the text into numerals. It should be taken into account that the numbers encountered in the text can denote integer, decimal and fractional cardinal numbers, ordinal numbers (which can be written in both Arabic and Roman numerals), date, time, phone numbers, etc. For the correct conversion of numbers, it is necessary to use the rules for "number – numeral" conversion, taking into account not only the number, but also the words and abbreviations in its neighborhood that help determine the characteristics of the number.

In addition, it should be borne in mind that the symbols "." and "," can be used both to separate digits in integers, and to separate the integer part from the fractional part. For example, in the record of the number 53,45, the comma separates the integer part from the fractional part, and in the record of 378,812,547 it serves to separate digits.

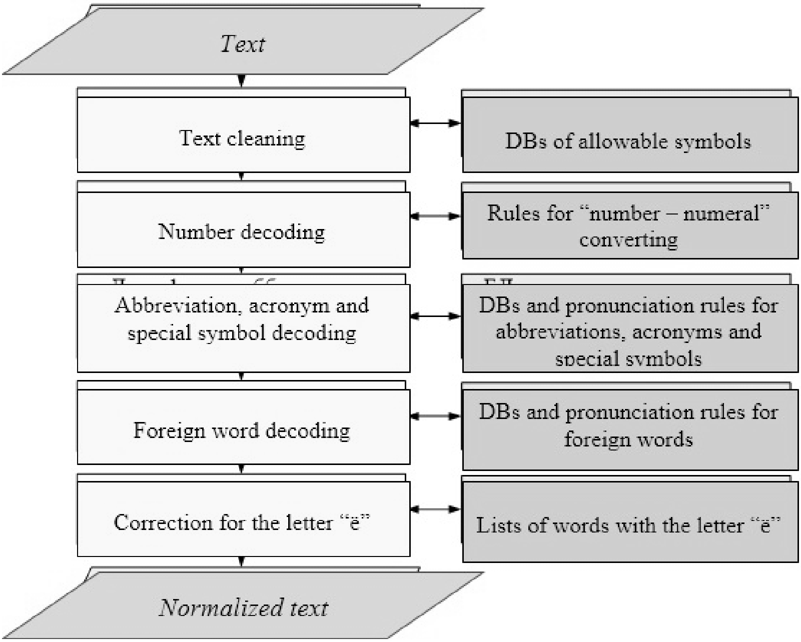

Text
Text cleaning
DBs of allowable symbols
Number decoding
Rules for "number – numeral" converting
Abbreviation, acronym and special symbol decoding
DBs and pronunciation rules for abbreviations, acronyms and special symbols
Foreign word decoding
DBs and pronunciation rules for foreign words
Correction for the letter "ë"
Lists of words with the letter "ë"
Normalized text

Figure 8-3. Text preprocessor structure

Abbreviation, acronym and special symbol decoding. In speech synthesis, it should be taken into account that the rules for reading abbreviations, acronyms and special symbols differ from the corresponding rules for words in the Russian language. To solve this problem, it is necessary to convert abbreviations, acronyms and special symbols into words, for which the standard rules used in phonetic and prosodic text processing are applicable. In this process of decoding, the following factors must be taken into account:

1. Abbreviations in texts are not always written in capital letters. This is typical primarily for the texts of e-mails, blogs and other texts

obtained from various Internet resources.

2. Some abbreviations and acronyms can be decoded differently depending on the subject domain (the context), for example, “г.” can mean “город” or “год”, “т.” can stand for “товарищ” or “тонн”.
3. Some abbreviations are not read in accordance with the standard decoding rules, for example: “США” is decoded as “эс-ше-а” according to the rules, however, the generally accepted pronunciation is “сэ-ше-а”.
4. Special symbols can be converted in different ways, for example, “%” can mean “процент”, “процента” or “процент”, and “$” can mean “доллар”, “доллара”, “долларов”.

To solve these problems, it is necessary to use DBs and the rules for pronunciation of abbreviations, acronyms and special symbols. The list of abbreviations of the Russian language contained in a DB will make it possible to detect an abbreviation in the text even if it is written in capital letters. The list of abbreviations and options for their interpretation (decoding), as well as the analysis of the context of the abbreviation, will ensure correct conversion of the abbreviation into a word.

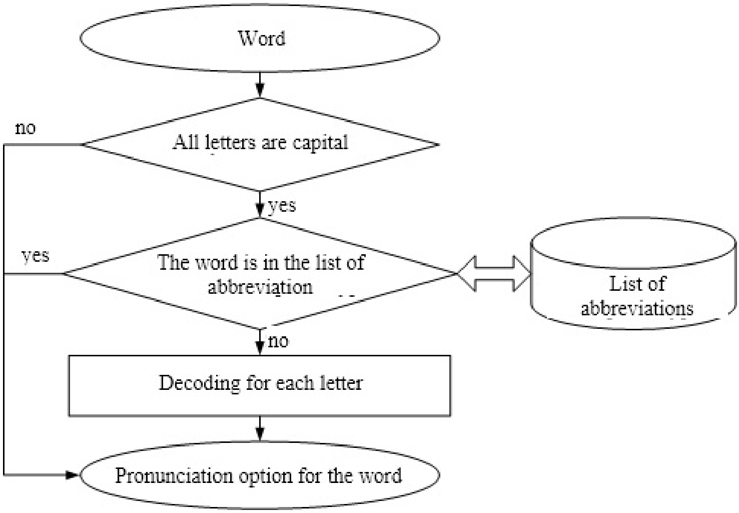
Abbreviations are usually pronounced by letters, for example, “КГБ” is pronounced as “ка-гэ-бэ”, “ФРГ” as “эф-эр-гэ”, with each syllable being stressed. However, the most common abbreviations, as well as abbreviations containing a large number of vowels, are usually pronounced in one word, for example, “ЮНЕСКО” is pronounced as “юнэ́ско”. This should be taken into account by the rules for pronunciation of abbreviations and acronyms. For correct conversion of special symbols, the rules for pronouncing must take into account the context of the symbol.

The scheme of the algorithm for decoding abbreviations is shown in Fig. 8-4a.

Foreign word decoding. Texts in Russian may contain Internet addresses, e-mail addresses, names of organizations written in Latin letters. To convert such words into a sequence of Russian letters read according to general rules, a unit for decoding foreign words is used. This unit uses the database and the rules for decoding Latin letters. The database should contain the most common foreign words, as well as their equivalents in Russian, for example, “Microsoft” – “ма́йкросо́фт”, “www” – “три да́блъю”. In addition, the rules for decoding Latin letters must contain Russian equivalents for each Latin letter. Then, if the word encountered in the text and written in Latin letters is not found in the DB, each letter will be converted according to the appropriate rules.

Correction for the letter “ё”. The problem of placing dots over the letter “ё” is, perhaps, the problem exiting only in the Russian language. Interestingly, when reading a text, a person does not think about how to read the word correctly, with the letter “ё” or “е”, using their knowledge of the language to correct the word. If, for example, during synthesis, the word “ёлка” is pronounced as “елка”, or the word “весёлый” is read as “веселый”, such an inaccuracy will be immediately noticed by the user. In the overwhelming majority of cases, lexical information is enough to correct the letter “ё”, namely, a DB containing the most complete list of words with the letter “ё” in Russian. Then in each word of the text containing one or more letters “е”, each of them is successively replaced by “ё” and the corresponding word is searched in the DB. However, in some cases, such information is insufficient, as, for example, for the word “все”: “Все в машине?” or “Всё в машине?”. Obviously, in this case, not only lexical and syntactic analyses are needed, but also semantic and pragmatic analyses as well. However, such situations are quite rare in texts.

The scheme of the algorithm for replacing the letter “Е” with “Ё” is shown in Fig. 8-4b.



a)

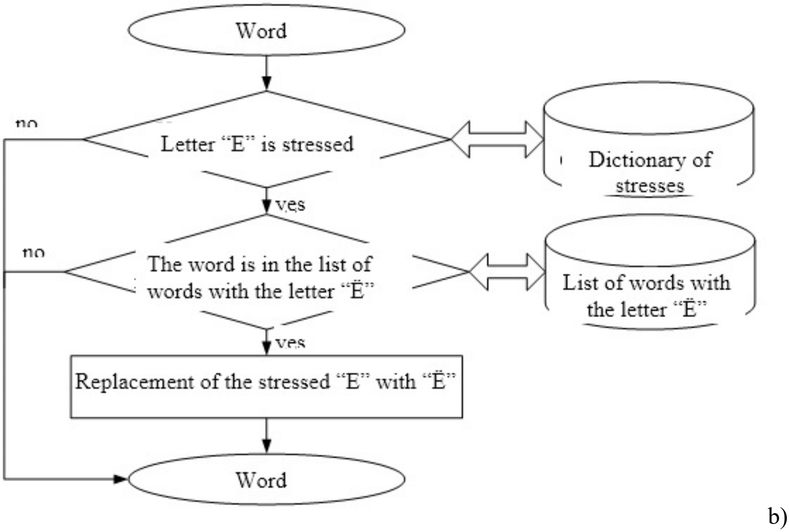

Word
Letter "E" is stressed
no
yes
Dictionary of stresses
The word is in the list of words with the letter "Ë"
no
yes
List of words with the letter "Ë"
Replacement of the stressed "E" with "Ë"
Word
b)

Figure 8-4. Diagrams for decoding abbreviations (a) and identifying the letter "ё" (b)

8.2.2. Unit for morphological and accentual marking of words

The structure of the unit for morphological and accentual marking of words is shown in Fig. 8-5.

Morphological marking means indicating for each word of the input text its belonging to a particular group of parts of speech, as well as additional morphological characteristics defined for this part of speech.

Accentual marking consists in marking stresses in words with strong or weak stress, followed by the addition of unstressed words to them, thus forming one phonetic word together with the stressed word.

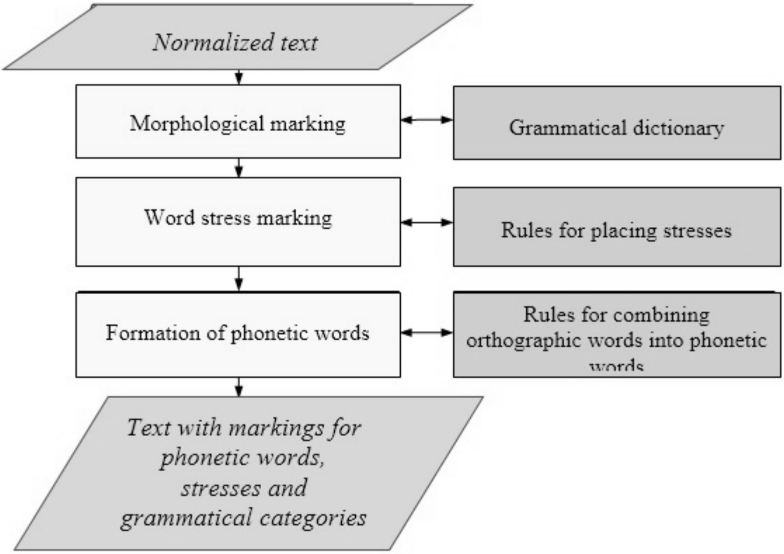

Normalized text
Morphological marking
Grammatical dictionary
Word stress marking
Rules for placing stresses
Formation of phonetic words
Rules for combining orthographic words into phonetic words
Text with markings for phonetic words, stresses and grammatical categories

Figure 8-5. Structure of the unit for morphological marking, placement of stresses and formation of phonetic words

Morphological marking. When morphological marking is used, each word of the text should have labels indicating the name of the part of speech, as well as additional grammatical categories characteristic of this part of speech. A grammar dictionary is used to define this information.

There are ten main parts of speech in Russian: noun, pronoun-noun, adjective, numeral, verb, adverb, preposition, conjunction, particle, interjection. Some verb forms, such as participles and gerunds, can also be reasonably categorized as content (notional) words.

Additional grammatical categories for parts of speech can be *inflectional* when the members of these categories can be represented by forms of the same word, for example: the category of case and number of the noun, the category of person, number, tense and mood of the verb, the category of the degree of comparison for adverbs, etc. Categories for parts of speech can be non-inflective when the members of these categories cannot be represented by forms of the same word, for example: the category of noun gender and the category of verb aspect. Content words: noun, pronoun-noun, adjective, numeral, verb and adverb have both inflectional and non-inflectional categories. Auxiliary parts of speech: preposition, conjunction, particle, interjection have non-inflective categories only. The definition of grammatical categories of parts of speech is necessary in the future for proper prosodic marking of the text.

All the categorical characteristics and properties of lexical units (about 100 thousand characters) for the Russian language are most fully reflected in the “Grammar Dictionary of the Russian Language” by A.A. Zaliznyak (Zaliznyak, 1987). Here, in unity, all sets of forms of a particular word are presented, which makes it possible to find the necessary information about the variability of a single lexeme. In this dictionary, information about the grammatical paradigm of a word (where a paradigm is understood as the totality of all grammatical forms of a certain word) is given using a system of conventions and indices.

For illustrative purposes, we present the complete inflectional paradigm of the noun <***конкурс*** **м 1а**> and the verb <***выбрасывать*** **нсв 1а**>. The symbol <**м**> indicates a number of non-inflectional features that characterize the word “***конкурс***”, namely: a noun, inanimate, masculine, substantive declension; <**1**> denotes the type of declension, depending on the end of the word stem; <**а**> indicates the stress pattern (constant stress on the stem). For the verb выбрасывать, <**нсв**> acts as a characteristic of the aspect (an imperfective verb); the index number indicates the type of conjugation, depending on which the ways of constructing the forms of the verbal paradigm are chosen. In this case, <**1**> shows that the infinitive ends with ***-ать*****,** ***-ять*** or ***-еть***, and the forms of the 1st and 3rd person singular verb of the present tense, respectively: ***-аю, -ает; -яю, -яет; -ею, -еет***. Finally, by the index <**а**> we recognize the stress pattern (constant stress on the stem). In addition to this information, the verb also has specific forms of participles, gerunds and a whole set of categorical meanings, which has a significant impact on the inflectional characteristics of words.

The inflectional categories of the noun and adjective are shown in Fig. 8-6, 8-7 respectively. The number of words in the grammar dictionary by A.A. Zaliznyak per each part of speech, as well as the number of word forms in the paradigms are presented in Table. 8-1.

The total number of words in the original dictionary is 98,222, of which more than two million word forms are generated in the Russian language.

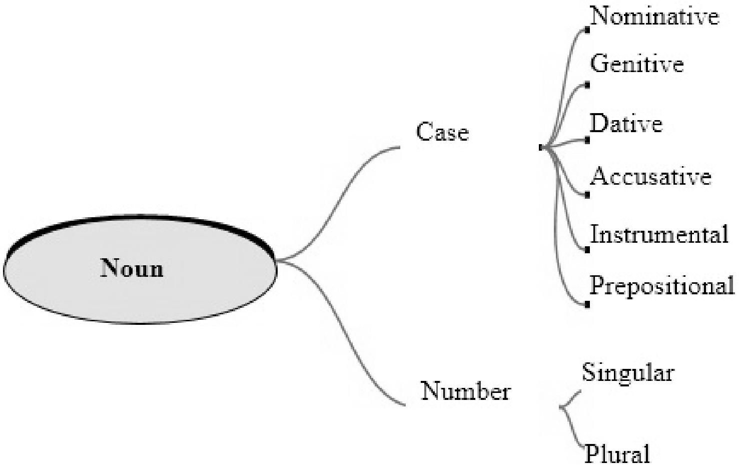


Figure 8-6. Inflectional categories of nouns

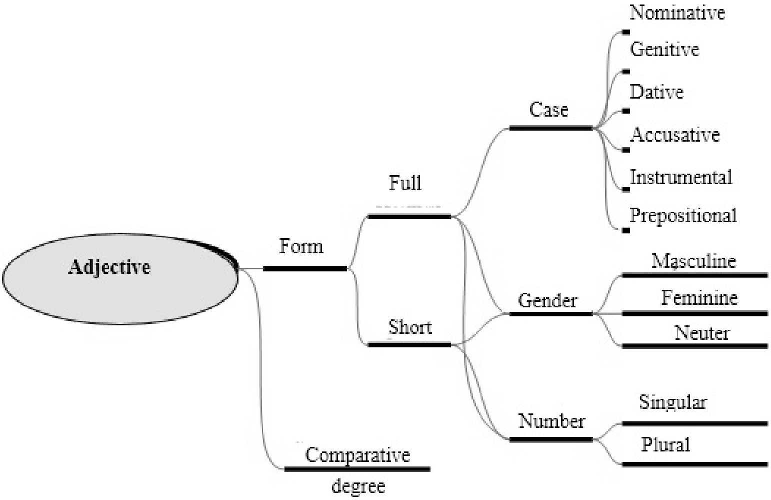


Figure 8-7. Inflectional categories of adjectives

Table 8-1 Number of words and word forms in the grammar dictionary by A.A. Zalizniak

Part of speech	Number of lexemes in the dictionary	Number of word forms in the paradigm	TOTAL word forms
Noun	46523	12	558276
Verb	27474	35	961590
Conjunction	68	1	68
Adverb	1353	1	1353
Interjection	180	1	180
Adjective	20622	29	599198
Preposition	87	1	87
Particle	86	1	86
Numeral	100	10	1000
TOTAL:	98 222	-	2 121 838

Marking of word stresses. To place the stresses, a grammar dictionary is used with marks of the stress position in the word form. Content words, as a rule, refer to fully-stressed words with one stress. However, some fully-stressed words may have, along with one strong (full) stress marked with a sign (+), one or more weak (partial) ones marked with a sign (=). Such words include, in particular, compound adjectives and nouns, for example, “ра=диолокацио+нный”, “мо=тове=лозаво+д”.

It should be noted that a certain number of content words – *homographs* – can have different stresses in different grammatical categories with the same letter composition. For example, some nouns have the same spelling in the genitive singular and in the nominative plural: “руки+”, “ру+ки”. To avoid such situations, i.e. to determine the exact grammatical category of a word, a deeper analysis of the text is needed.

At the stage of placing stresses, it is also necessary to take into account that no matter how large the dictionary is, the text may well contain a word that is not in it. To mark the stress in such a word (since it cannot be voiced without stress), one of two methods can be used: place the full stress based on statistical information about word stresses, or place a partial stress on each syllable. In the second case, the word will be “read” by syllables. With the first method, the stress can be possibly placed incorrectly, and such a word will not be perceived correctly by the listener.

Therefore, the second method seems to be more reasonable.

Formation of phonetic words. Many auxiliary words can be pronounced without explicit stress. Unstressed word forms include non-syllable prepositions ***в, к, с*** and particles ***б, ж, ль***, as well as monosyllabic prepositions ***без***, ***во***, ***для***, ***за***, etc. and particles ***де, ка, ан, бы***, etc. For example, in phrases "*доехать* ***до*** *Киева*", "*прибыл* ***бы*** *вовремя*", the preposition ***до*** and the particle ***бы*** are pronounced, as a rule, without any stress and are attached to the following and preceding words, respectively.

After marking word stresses, it is necessary to attach each unstressed word to a nearby (previous or subsequent) stressed word. The operation of attaching unstressed words to stressed words is performed by the unit for forming phonetic words. A phonetic word is one or more orthographic words that have one common stress. To attach unstressed words to partially- or fully-stressed words, a set of rules is used that takes into account the grammatical characteristics of the unstressed word, as well as the neighboring words. Moreover, only the particles "***бы***", "***-де***", "***дескать***", "***-ли***", "***-же***", "***мол***", "***-то***", "***-ка***", "***-либо***", " ***-нибудь***".

For example, in the sentence: "*Мальчик успел* ***бы*** *вовремя, если* ***бы*** *не остановился поболтать* ***с*** *другом*," the preposition "***с***" is attached to the next word (noun): "*сЪдругом*", and the particle "***бы***" is attached to the previous words (verb and conjunction): "*успелЪбы*", "*еслиЪбы*". Here the letter *Ъ* is used as a symbol of attachment.

8.3. Prosodic processor

Text-to-speech synthesis assumes the presence of an automatic procedure for the formation of current melodic (pitch) patterns, sound intensity, phonemic duration and pause duration based on the analysis of certain properties of the input text and its prosodic markup. The prosodic markup of the text consists in its segmentation into syntagmas (or otherwise - into phrases), the markup of syntagmas into accent units and the marking of the intonational type of syntagmas in accordance with certain rules.

A syntagma is understood as a part of a sentence, independent in the intonational sense, or the whole sentence. Setting the boundaries of syntagmas affects the way intonational characteristics and semantic content are conveyed in speech synthesis. When segmenting the text into syntagmas, it is important not to place the boundary of the syntagma where it can disrupt the semantic perception of speech (or the conveying of the semantic content of the text), for example, between an object and its attribute. To set the boundaries of syntagmas, certain rules of syntagmatic segmentation are used, based on punctuation, morphological and syntactic analysis of the text, as well as on the statistical analysis of syntagmatic segmentation in natural speech.

Syntagmas in speech are usually separated by pauses. Pauses take part in the transmission of certain syntactic and semantic relations. In addition, the time intervals created by pauses allow the listener to perform linguistic processing of the text, memorize its results and build the semantic structure necessary for the perception of the text. In natural speech, there are grammatical pauses separating the intonation-shaped parts of the phrase from each other; emphatic pauses and hesitation (uncertainty) pauses. The syntagma boundary can be marked not only by a physical break in the speech signal, but also by a sharp change in pitch and (or) other prosodic characteristics that are perceived as a violation of the smooth flow of speech.

It is important to note that the process of syntagmatic segmentation should be the solution of two main tasks: to establish the boundaries of syntagmas in those places where they must be present, and not to establish the boundary of a syntagma where it can violate the semantic perception of speech.

The structure of the prosodic processor is shown in Fig. 8-8.

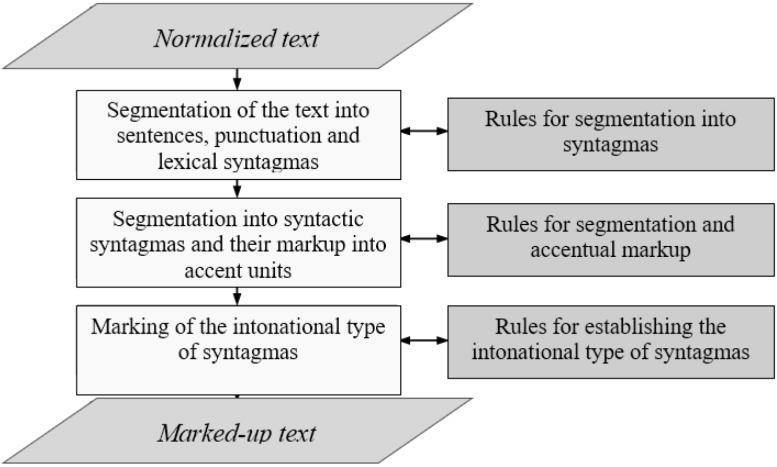

Normalized text
Segmentation of the text into sentences, punctuation and lexical syntagmas
Rules for segmentation into syntagmas
Segmentation into syntactic syntagmas and their markup into accent units
Rules for segmentation and accentual markup
Marking of the intonational type of syntagmas
Rules for establishing the intonational type of syntagmas
Marked-up text

Figure 8-8. Prosodic processor structure

8.3.1. Unit for segmenting the text into sentences, punctuation and lexical syntagmas

The structure of the unit for segmenting the input text into sentences, punctuation and lexical syntagmas is shown in Fig. 8-9.

Segmentation of the text into sentences. Speech synthesis is performed according to sentences that are characterized by a sufficient degree of intonational autonomy in the text and allow for a sufficiently long pause between them (0.5 - 1.5 seconds). A sentence is a minimal unit of speech, which is a grammatically organized combination of words (or only one word), which has certain semantic and intonational completeness.

A *sentence* is a segment of text delimited by the signs [.], [?], [?!], [!], [!!!]. The *end of a sentence* can also be marked with [**…**], provided that the word following it begins with a capital letter.

A *sentence* will also be considered the heading of the entire text or part of it, at the end of which there may be no sign [.]. The end of such a sentence is denoted by the sign [*]. In addition, a *sentence* limited by a point at the end of a paragraph is considered as a separate type of sentences. The end of a paragraph is denoted by the sign [#].

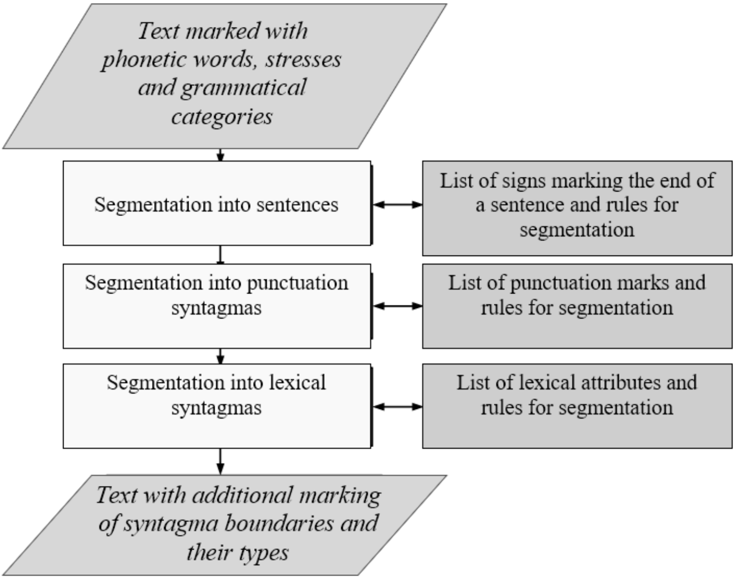

Text marked with phonetic words, stresses and grammatical categories
Segmentation into sentences
List of signs marking the end of a sentence and rules for segmentation
Segmentation into punctuation syntagmas
List of punctuation marks and rules for segmentation
Segmentation into lexical syntagmas
List of lexical attributes and rules for segmentation
Text with additional marking of syntagma boundaries and their types

Figure 8-9. Structure of the unit for segmenting the text into sentences, punctuation and lexical syntagmas

Segmentation of the sentence into punctuation syntagmas. Punctuation marks are indicators of punctuation syntagmas (PSs). We will consider as *punctuation syntagmas* a sentence (in the absence of punctuation marks in it) or parts of a sentence limited by the following signs:

- semicolon [;],
- colon [:],
- comma [,],
- dash [–],
- opening bracket [( ],
- closing bracket [ )],
- combination of symbols [,–],
- opening quotation marks [«], [“],
- closing quotation marks [»], [”]

Thus, if a sentence includes *n* punctuation marks (including the end-of-sentence sign), then it is divided into *n* punctuation syntagmas (*n*=1,2,3,…). A definite exception to this rule can be the situation when the punctuation mark is found after a coordinating conjunction: *и, да, но и, так и, а, но, однако, зато, или, либо, то,* etc. In this case, it would be preferable not to establish a syntagmatic boundary in place of this punctuation mark, although it is acceptable for some individual style of speech.

8.3.2. Unit for segmenting PSs and LSs into syntactic syntagmas and their dividing into accentual units

Even after splitting the sentence into PSs and LSs, their length may be too long.

Example: “*Но молодая жена упорно продолжала отстирывать белую в кровавых пятнах рубаху мужа посиневшими от холода руками в железном тазике с ледяной водой*”.

In the above sentence, there are no signs of the presence of a PS or LS in it. Obviously, in the absence of a mechanism for further segmenting of such sentences into smaller syntactic syntagmas (SSs), difficulties will inevitably arise in understanding the meaning of synthesized speech. An ideal solution to the problem of further segmentation of this kind of PSs or LSs into SSs would be the use of a set of rules for their deep syntactic parsing. However, in view of the complexity and insufficient degree of development of such rules, it is still necessary to limit ourselves to the use of a surface syntactic analysis procedure based on the available morphosyntactic information about the phrases that make up a particular PS or LS.

A phrase is considered as a pair of semantically and grammatically related words that are singled out in a sentence. Being, along with the word, an element of sentence construction, the phrase acts as one of the main syntactic units. The immediate goal of the considered procedure of surface syntactic analysis is the preliminary segmentation of a PS or LS into a sequence of phrases of 2 types: fixed collocations (FCs) and grammatical-semantic phrases (GSPs).

The structure of the unit for segmenting PSs and LSs into syntactic syntagmas and their dividing into accentual units, based on the analysis of phrases, is shown in fig. 8-10.

Identification of collocations. In the analyzed syntagma, collocations (FCs) found in the dictionary of set phrases (fixed collocations) are marked up. Fixed collocations include:

- phraseological fusions: "*попасть впросак", "бить баклуши", "ничтоже сумняшеся", "собаку съесть"*, etc.
- phraseological unities: "*зайти в тупик", "бить ключом", "плыть по течению", "брать в свои руки", "прикусить язык"*, etc.
- phraseological combinations: "*потупить взор", "щекотливый вопрос", "бархатный сезон"*, "*поголовные аресты"*, etc.

The following types of component composition of phraseological units are distinguished:

- a combination of an adjective with a noun: *краеугольный камень, заколдованный круг, лебединая песня*;
- a combination of a noun in the nominative case with a noun in the genitive case: *точка зрения, камень преткновения, бразды правления, яблоко раздора*;
- a combination of a noun in the nominative case with nouns in oblique cases with a preposition: *кровь с молоком, душа в душу, дело в шляпе*;
- a combination of the prepositional case of a noun with an adjective: *на живую нитку, по старой памяти, на короткой ноге*;
- a combination of a verb with a noun (with and without a preposition): *окинуть взором, посеять сомнения, взять в руки, взяться за ум, водить за нос*;
- a combination of a verb with an adverb: *попасть впросак, ходить*

босяком, видеть насквозь;

- a combination of a gerund with a noun: *спустя рукава, скрепя сердце, сломя голову*.

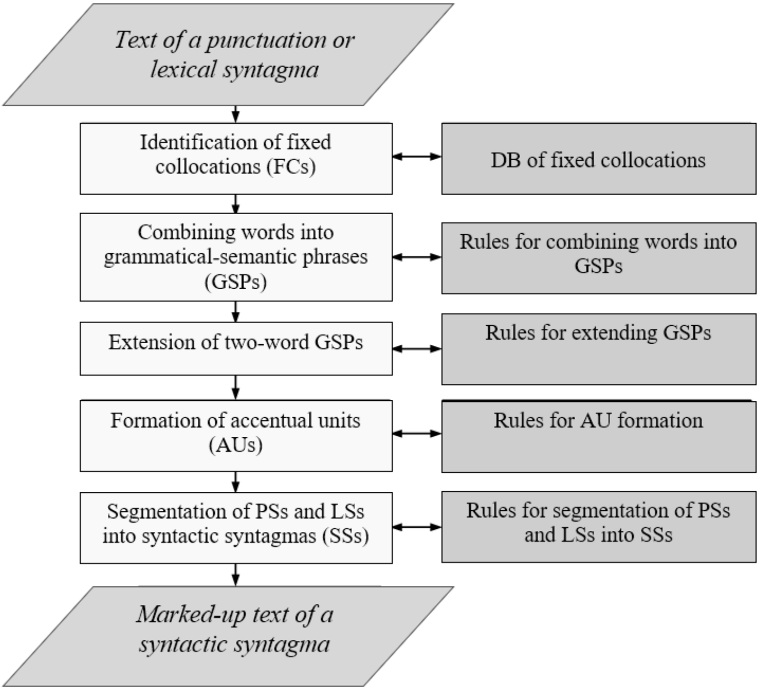

Text of a punctuation or lexical syntagma
Identification of fixed collocations (FCs)
DB of fixed collocations
Combining words into grammatical-semantic phrases (GSPs)
Rules for combining words into GSPs
Extension of two-word GSPs
Rules for extending GSPs
Formation of accentual units (AUs)
Rules for AU formation
Segmentation of PSs and LSs into syntactic syntagmas (SSs)
Rules for segmentation of PSs and LSs into SSs
Marked-up text of a syntactic syntagma

Figure 8-10. Structure of the unit for segmentation of PSs and LSs into syntactic syntagmas

The positions of weak and strong word stresses in fixed collocations can be determined in the dictionary of combinations, while one of the words has a strong stress mandatorily. In the absence of marks for weak and strong stresses, it is quite acceptable to place a strong stress on each of the words of the fixed collocation.

Combining words into grammatical-semantic phrases. A grammatical-semantic phrase (GSP) is a pair of semantically and grammatically related words extracted from a sentence. For example, *"нужная книга", "лекция по литературе", "бежать опрометью", "два студента", "несколько книг"*.

The main point of singling out phrases such as FCs and GSPs in PSs or LSs is that now the freedom of segmenting PSs or LSs into SSs is limited. The boundary of a syntagma can only be outside a FC or GSP, but not inside them.

Extension of two-word GSPs to three- or more word combinations. If a syntagma processed in accordance with the sequence of actions for segmenting syntagmas into phrases indicated above contains words that are not included in the identified two-word combinations, then the possibility of extending them to three- or more word combinations is considered.

Formation of accentual units (AUs). In each word of the sentence, the position of the full (+) and partial (=) stresses must be indicated. For two-word combinations, the location of the preferred installation of strong and weak stresses is usually known. It characterizes the average trend for a fairly wide range of different texts. Under certain conditions (individual reading style, pursuance of a certain rhythmic structure, etc.), the signs (+) (=) for a given phrase can change places, or both signs can indicate strong stress, that is, (+) (+). An important role may also be played by the presence of some indicators of the potential "weakness" or "strength" of any of the words in the phrase. In particular, the indicator of "weakness" may include the fact that a word belongs to a group of potentially weakly stressed words, such as polysyllabic prepositions and particles, conjunctions and pronouns. The presence of an intensifying or negative particle before the word can be attributed to the indicator of "strength". The marking of the sequence of words in a sentence into accent units is carried out according to the following rules:

- Marking on AUs is carried out separately for each FC or GSP.
- If there are words with weak stress in FC or GSP, then each of them is combined into one AU with a strongly stressed word to the left or right of it.
- The remaining heavily stressed words are marked as separate AUs.

Segmentation of PSs and LSs into syntactic syntagmas. As already mentioned, the main meaning of the preliminary division of PSs or LSs into FCs and GSPs is that now the freedom of their division into SS is limited, because the border between the SSs cannot be inside the FC or GSP. In the simplest case, the boundaries of the FC or GSP can serve as the boundary of each SS. In this case, each SS will include a different number of AEs: from 1 to 2 or more. If the required reading style suggests that the SS should include, if possible, more than one AE, then in this case it is allowed to include more than one FC or GSP in this SS.

8.3.3. Unit for marking an intonational type of syntagmas

The structure of the unit for marking an intonational type of syntagmas in a sentence is shown in Fig. 8-11.

Marking an intonational type of a sentence. As mentioned above, in text-to-speech synthesis, a sentence can be identified by one of the following eight signs: [.], [?], [?!], [!], [!!!], […], [*], [#]. The end-of-a-sentence signs define its three main intonational categories:

- narrative – ***Nr*** (Narration),
- interrogative – ***In*** (Interrogation),
- exclamation (incentive) – ***Ex*** (Exclamation).

The category of ***narrative sentences*** – ***Nr*** – is characterized by a narrative or, otherwise, finalized intonation – ***F*** (Finality). This category is recognized by the signs [.], […], [*], [#] that determine its intonational type indicated during the text processing, respectively, by the following symbols:

- ***F0*** – “point” intonation - [.],
- ***F1*** – “ellipsis” intonation - […],
- ***F2*** – “header” intonation - [*],
- ***F3*** – “paragraph” intonation - [#].

In addition to the four main intonational types of narrative sentences, identified by punctuation, it is possible to expand their number, for example, by supplementing them with subtypes that characterize direct or indirect speech, etc.

The category of ***interrogative sentences*** – ***In*** – is recognized by the signs: [?], [?!] and is indicated during the text processing, respectively, by symbols of intonational types:

- ***I0*** – “question” intonation - [?],
- ***I1*** – “question-exclamation” intonation - [?!].

In addition to the mentioned 2 types of intonation for interrogative sentences, identified by punctuation, each of them can have several subtypes, such as intonations of:

- a general question,
- a special question,
- a disjunctive question,
- an echo-question,
- a negation question, etc.

The category of ***exclamatory and incentive sentences*** – ***Ex*** – is recognized by the signs [!], [!!!] and is indicated, respectively, by symbols of intonational types:

- ***E0*** – exclamation-incentive intonation - [!],
- ***E1*** – emotional exclamation intonation - [!!!].

In addition to the indicated 2 types of intonation of exclamatory sentences, identified by punctuation, each of them can have several subtypes that express certain feelings and motivations using various kinds of interjections.

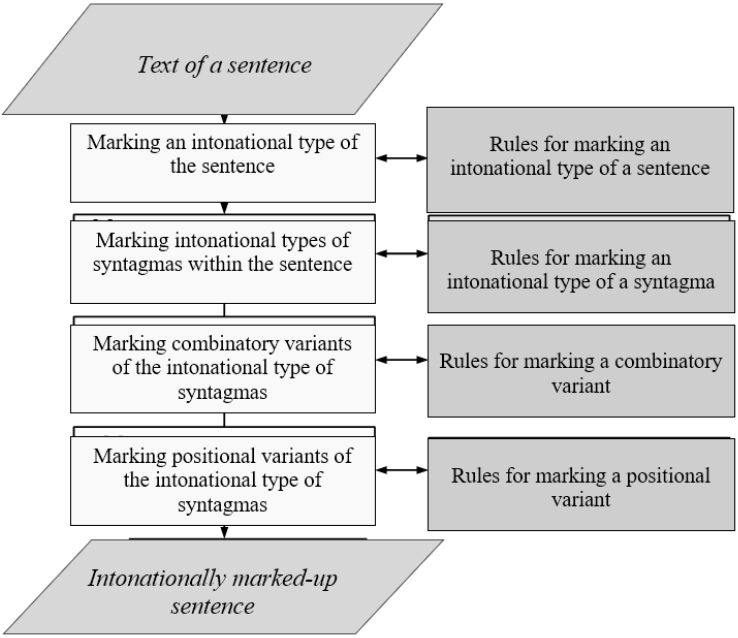

Text of a sentence
Marking an intonational type of the sentence
Rules for marking an intonational type of a sentence
Marking intonational types of syntagmas within the sentence
Rules for marking an intonational type of a syntagma
Marking combinatory variants of the intonational type of syntagmas
Rules for marking a combinatory variant
Marking positional variants of the intonational type of syntagmas
Rules for marking a positional variant
Intonationally marked-up sentence

Figure 8-11. Structure of the unit for intonational mark-up of syntagmas in a sentence

Marking intonational types of syntagmas within a sentence. We will consider the features of marking intonational types of syntagmas within a sentence using the example of narrative sentences. In addition to the main punctuation types of finalized intonation listed above, which are realized in the last PS or LS of a sentence, two additional punctuation types of intonation can also be present inside it, characterized by varying degrees of finalization:

- ***F4*** – “semicolon” intonation – [;],
- ***F5*** – “introductoriness” intonation – [ )], [,–], [–].

The “introductoriness” intonation is realized under the condition that the indicated signs are immediately preceded, respectively, by the signs [( ], [,–], [–].

Within a sentence, there may also be 5 punctuation subtypes of intonation, characterized by varying degrees of non-finalization:

- ***N0*** – “comma” intonation - [,],
- ***N1*** – “dash” intonation - [-],
- ***N2*** – “colon” intonation - [:],
- ***N3*** – “pre-introductoriness” intonation - [( ], [,–], [–].

The intonation of “pre-introductoriness” is realized under the condition that the indicated signs in the sentence text are followed, respectively, by the signs [ )], [,–], [–].

Punctuation syntagmas, in turn, can contain lexical syntagmas with the intonation of non-finalization of the following 3 types (see section 8.3.1):

- ***N4*** – “conjunction AND” intonation,
- ***N5*** – “conjunction OR” intonation,
- ***N6*** – intonation of lexical syntagmas.

Further, both the sentence itself and the punctuation and lexical syntagmas within it can contain an indefinite number of syntactic syntagmas with their distinctive intonation of non-finalization:

- ***N7*** – intonation of syntactic syntagmas.

8.4. Phonetic processor

The purpose of the phonetic processor is to convert orthographic texts into a sequence of allophones, which is used at the stage of acoustic processing in the speech signal synthesis.

The phonetic processor contains rules for converting orthographic texts into a sequence of phonemes (letter-phoneme conversion) and rules for converting a sequence of phonemes into an allophone sequence (phoneme-allophone conversion). The general structure of the phonetic processor is shown in Fig. 8-12.

8.4.1. Unit for conversion of words-phonetic exceptions

The input of the processor receives the orthographic text of a syntagma with marked word stresses and boundaries of accentual units. At the initial stage, each word of the syntagma is searched in the database of words-phonetic exceptions. If the word is found, it is replaced with the corresponding equivalent.

Words-phonetic exceptions include a large number of foreign words or words with a foreign root, for example, “*ателье*”, “*варьете*”, “*декольте*”, “*интервью*”, “*кабаре*”, “*кафе*”, “*кашне*”, “*моделировать*”, “*филателист*”, where the consonant before “*е*” is not softened. Exception words are converted to their equivalents, following the standard letter-phoneme conversion rules. For the examples given, the equivalents will be, respectively: “*атэлье*”, “*варьетэ*”, “*декольтэ*”, “*интэрвью*”, “*кабарэ*”, “*кафэ*”, “*кашнэ*”, “*модэлировать*”, “*филатэлист*”. Words-phonetic exceptions also include commonly used words, such as “*пожалуйста*” (the equivalent is “*пожалуста*”), “*здравствуйте*” (“*здраствуйте*”), “*что*” (“*што*”), “*чувства*” (“*чуства*”).

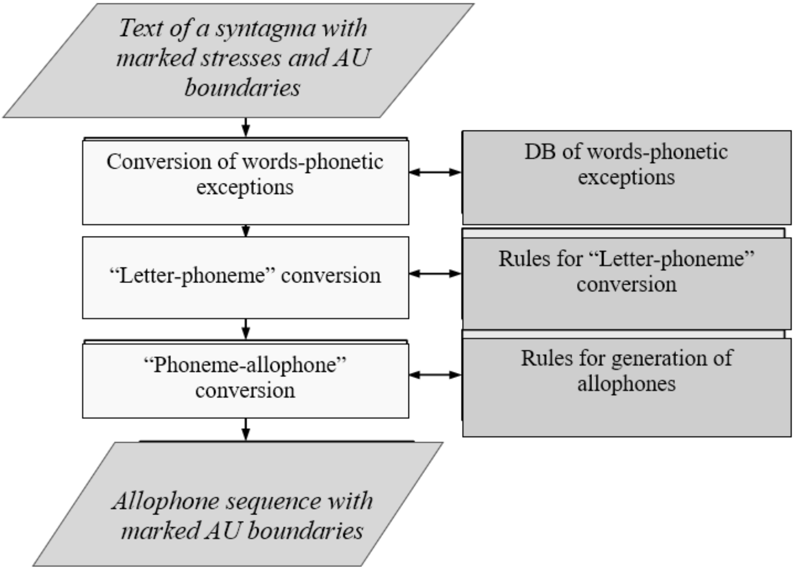

Text of a syntagma with marked stresses and AU boundaries
Conversion of words-phonetic exceptions
DB of words-phonetic exceptions
"Letter-phoneme" conversion
Rules for "Letter-phoneme" conversion
"Phoneme-allophone" conversion
Rules for generation of allophones
Allophone sequence with marked AU boundaries

Figure 8-12. Phonetic processor structure

8.4.2. Unit for letter-phoneme conversion

At the next stage, according to the standard rules, a letter-phoneme conversion is performed, taking into account pronunciation features for the Russian language. The basic regular rules for letter-phoneme conversion have been described in (Lobanov, B.M. 1983). Here we note some features of the conversion for consonant letters that are not reflected in the above regular rules, and also describe the rules for letter-phoneme conversion at the junction of words.

Additional intra-word rules for letter-phoneme conversion. In Russian, there are three-term consonant combinations, where one of the consonants is not pronounced. Such combinations include "стн", "стл", "нтг", in which "т" is not pronounced; "здн", "здц", "ндц", "рдц", "ндш", "гдт", in which "д" is not pronounced; a combination of "лнц", in which "л" is not pronounced.

The rules for converting a letter-phoneme inside a significant word, at the junction of a functional and significant word, and at the junction of two phonetic words are shown in Fig. 8-13a for consonants and in Fig. 8-13b for vowels.

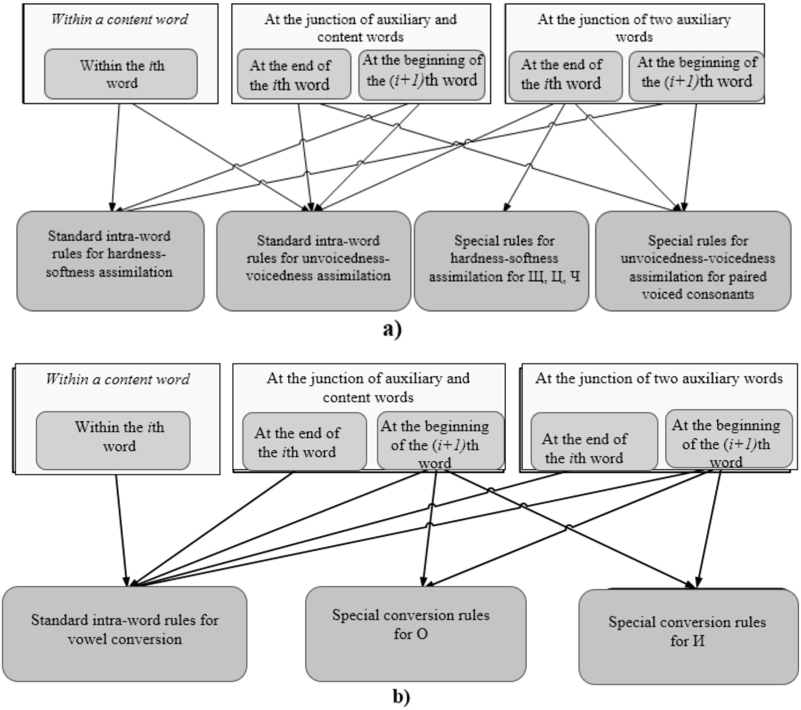

Within a content word
Within the ith word
At the junction of auxiliary and content words
At the end of the ith word
At the beginning of the (i+1)th word
At the junction of two auxiliary words
At the end of the ith word
At the beginning of the (i+1)th word
Standard intra-word rules for hardness-softness assimilation
Standard intra-word rules for unvoicedness-voicedness assimilation
Special rules for hardness-softness assimilation for Щ, Ц, Ч
Special rules for unvoicedness-voicedness assimilation for paired voiced consonants
a)
Within a content word
Within the ith word
At the junction of auxiliary and content words
At the end of the ith word
At the beginning of the (i+1)th word
At the junction of two auxiliary words
At the end of the ith word
At the beginning of the (i+1)th word
Standard intra-word rules for vowel conversion
Special conversion rules for O
Special conversion rules for И
b)

Figure 8-13. Features of the letter-phoneme conversion within and at the junctions of words: a) for consonants, b) for vowels.

8.4.3. Unit for phoneme-allophone conversion

Phoneme-allophone conversion is performed in two stages. At the first stage, phonemes are converted into positional allophones, and at the second stage, positional allophones are converted into positional-combinatorial ones.

The conversion of phonemes into positional allophones (which show the position of the phoneme in relation to word stress) is especially important for vowels, since they are subject to significant quantitative and qualitative reduction. Strongly stressed vowels have the greatest duration and strength of sound; weakly stressed vowels have smaller duration and strength of sound. The next in terms of duration and strength of sound are the vowels of the first degree of reduction, and, finally, the least emphasized are the vowels of the second degree of reduction.

As studies have shown (Lobanov and Tsirulnik, 2006b), the first degree of reduction includes vowels that are immediately before the stressed vowel (i.e., the first pre-stressed vowels) in the phonetic word or are the first or last sound of the phonetic word. Vowels of the second degree of reduction do not include the first pre-stressed and post-stressed vowels (if they are not the first or last sound of the phonetic word).

These factors are taken into account by the rules for generating positional allophones of vowels, shown in Fig. 8-14. Denoting the positional allophone with an index *i* following the phoneme name, we will have the following *i* values for vowels: 0 - strongly stressed, 1 - weakly stressed, 2 - vowel of the first degree of reduction, 3 - vowel of the second degree of reduction.

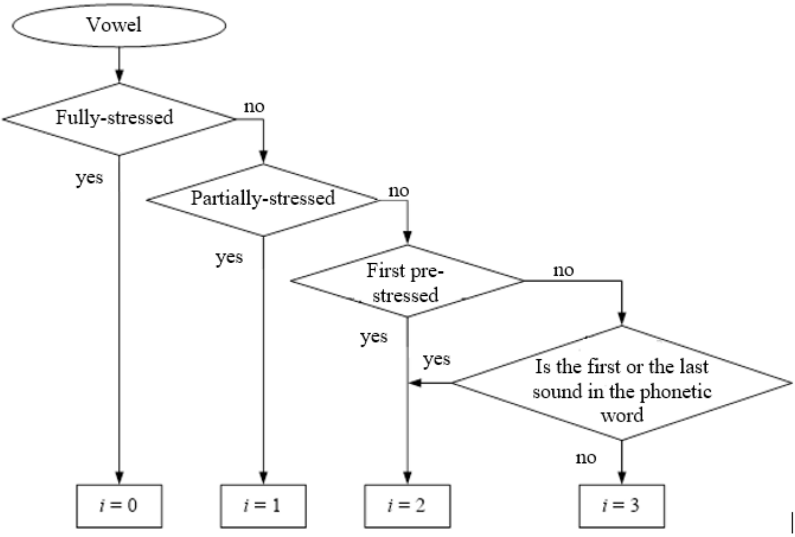


Figure 8-14. Rules for generating the positional index for vowels

The reduction of consonants in natural speech is small compared to vowels, and here, taking into account auditory perception, two situations must be distinguished: the consonant may be in a stressed syllable and in an unstressed syllable. Such a division, however, doubles the number of consonant allophones, which leads to an increase in the size of the speech corpus, requires more time to record it and more time to prepare the phonetic-acoustic DB. With these factors in mind, it is possible to discard the positional allophony of consonants. This simplification is compensated to some extent at subsequent stages of synthesis, when allosyllabic segments containing the positional allophones of consonants required for synthesis are selected from the phonetic-acoustic DB.

For the identity of the designation of consonant allophones, a positional index is also introduced, but unlike vowels, it shows the doubling of the phoneme and means the following: 0 is a usual phoneme, 1 is a doubled phoneme. The rules for generating the positional index for consonants are shown in Fig. 8-15.

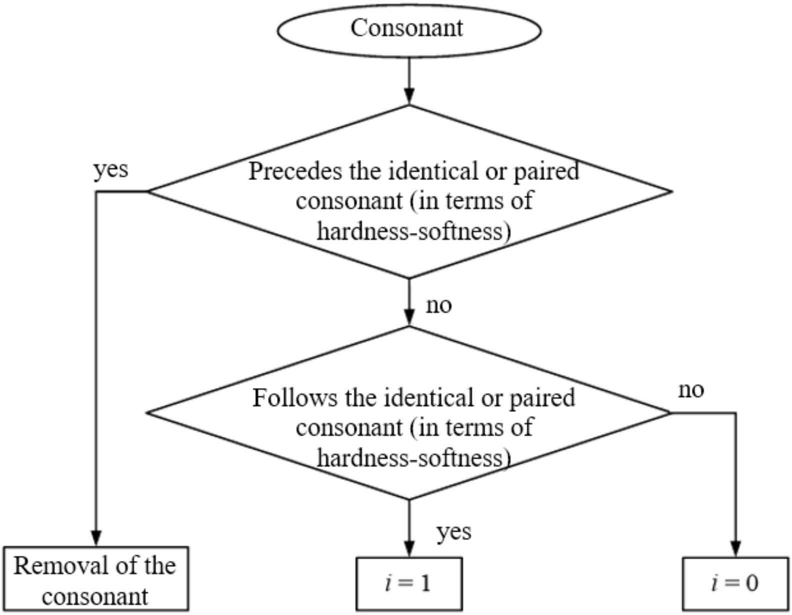

Consonant
Precedes the identical or paired consonant (in terms of hardness-softness)
yes
no
Follows the identical or paired consonant (in terms of hardness-softness)
no
yes
Removal of the consonant
$i = 1$
$i = 0$

Figure 8-15. Rules for generating the positional index for consonants

It is noteworthy that when such rules are used, the positional allophone formed by two identical consonant phonemes will coincide with the positional allophone formed by two consonant phonemes paired in terms of hardness-softness. For example, in the phoneme sequence "*p, a, d, d', e, r', e, v, a, m*" (orthographic text "под деревом"), a pair of phonemes *d, d'* (hard and soft) is converted to the positional allophone d'_1; in the phoneme sequence "*h, o, d', d', e, r', e, v, a*" (orthographic text "хоть дерево"), a pair of phonemes *d'*, *d'* (both soft) is also converted into the positional allophone d'_1. In such situations, the hardness-softness of the phoneme is taken into account at the next stage (described below), when converting positional allophones into combinatorial ones, and the combinatorial indices of the allophone that precedes a pair of consonants that are identical or paired in terms of hardness-softness will be different.

The next stage of the unit for phoneme-allophone conversion is the conversion to positional-combinatorial allophones. The combinatorial factor takes into account the immediate neighborhood of the phoneme, i.e. the left context that is the phoneme immediately preceding the given one, and the right context that is the phoneme immediately following the given phoneme. Combinatorial characteristics are denoted by indices *j* and *k*, while the index *j* indicates the group of the left context, the index *k* the group of the right context.

As studies have, the rules for generating combinatorial allophones are different for phonemes that differ *in the way they are generated*. To generate combinatorial allophones, the entire set of phonemes is divided into the following classes:

- non-labial vowels {*a, i, e, y*},
- labial vowels {*u, o*},
- most unvoiced consonants {*p, p', t, t', k', c, ch', f, f', s, s', sh, sh', h'*},
- hard aspirative consonants {*k, g, h*},
- voiced obstruent, fricative and sonorant consonants { *b, b', d, d', g', z, z', zh, l, l', m, m', n, n', r, r'*},
- liquid sonorant consonants {*v, v', j'*}.

The left and right phonemic contexts are grouped *according to the place of generation*, but they are different for different classes of phonemes. Vowels have the largest number of groups of left contexts (six), and all consonants, except for liquid sonorants, have the smallest number (one left context). The largest number of right contexts (four) are attributed to vowels, as well as to voiced and liquid sonorants; and the smallest (two contexts) to unvoiced consonants.

8.5. Acoustic processor

The general structure of the acoustic processor is shown in Fig. 8-16.

The purpose of the first unit of the acoustic processor is to convert the prosodically marked sequence of allophones of the syntagma into a sequence of their sound waves with the values of pitch frequency F_0, amplitude A and duration T specified by DBs of prosodic portraits. In the second unit, the synthesis of a speech signal is performed by selecting from the DB of sound waves of multiphones (single allophones, dialophones, allosyllables) corresponding to the input allophone text, and their concatenation (combination).

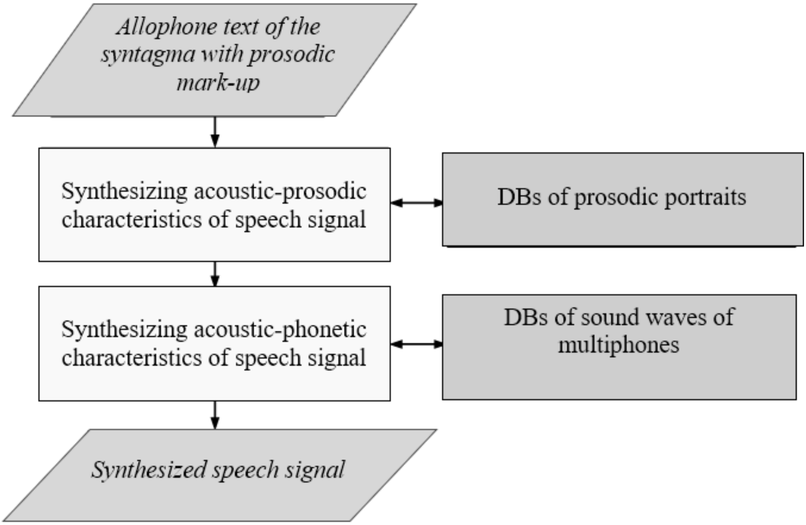

Allophone text of the syntagma with prosodic mark-up
Synthesizing acoustic-prosodic characteristics of speech signal
DBs of prosodic portraits
Synthesizing acoustic-phonetic characteristics of speech signal
DBs of sound waves of multiphones
Synthesized speech signal

Figure 8-16. Acoustic processor structure

8.5.1. Unit for synthesizing acoustic-prosodic characteristics of speech signal

The functional diagram of the unit for synthesizing prosodic characteristics of speech is shown in Fig. 8-17. Prosodic characteristics are synthesized sequentially for each syntagma. At the first stage, each syntagma is marked-up into AUs, each AU into accentual unit elements (AUE): pre-core, core, and post-core. The AU core, according to the rules used, is a fully-stressed vowel; all allophones preceding a fully-stressed vowel are a pre-core segment; all allophones following a fully-stressed vowel are a post-core segment (Lobanov et al., 2006).

Then, for each syntagma, it is necessary to choose the prosodic patterns corresponding to its intonational type: intonational, rhythmic, dynamic. For this, a DB of prosodic portraits of accentual units (AUPs) is used, which contains prosodic "portraits" for each intonational type used. The AUP DB can contain several sets of prosodic portraits, each of which characterizes a certain style of the "voiced" text (scientific, official, journalistic, fiction, colloquial), individual prosodic characteristics of a particular speaker, expression of various emotions in speech and etc. In prosodic portraits, the duration of the inter-syntagmal pause is also preserved.

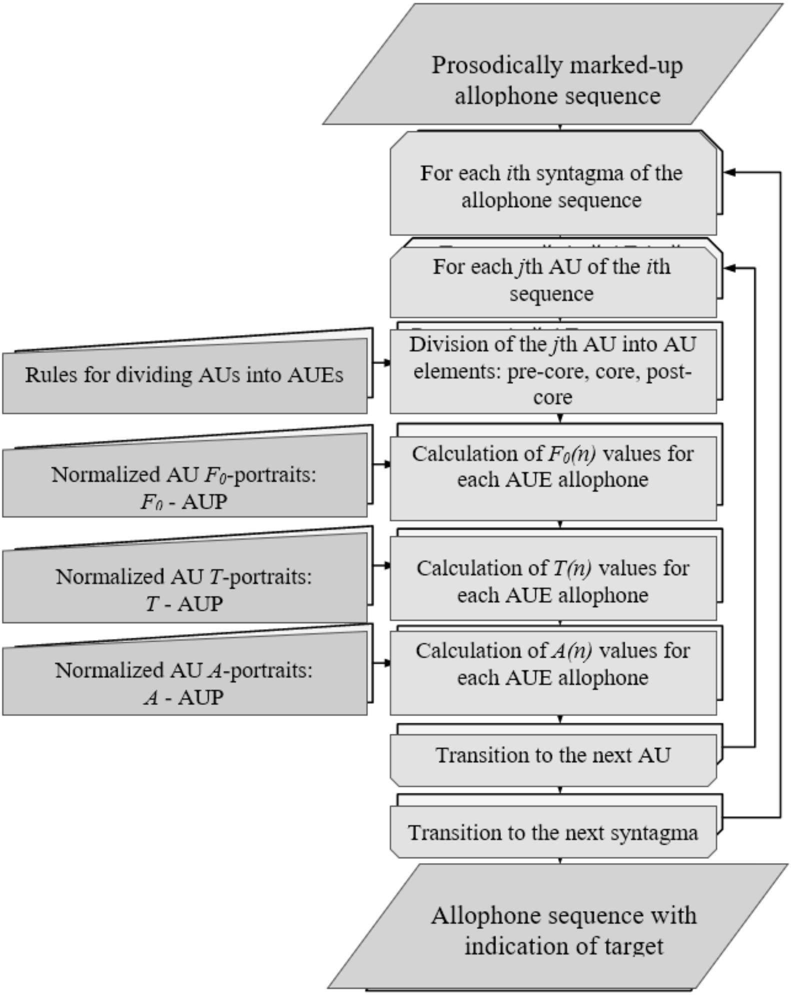

Prosodically marked-up allophone sequence
For each ith syntagma of the allophone sequence
For each jth AU of the ith sequence
Rules for dividing AUs into AUEs
Division of the jth AU into AU elements: pre-core, core, post-core
Normalized AU F_0-portraits: F_0 - AUP
Calculation of $F_0(n)$ values for each AUE allophone
Normalized AU T-portraits: T - AUP
Calculation of T(n) values for each AUE allophone
Normalized AU A-portraits: A - AUP
Calculation of A(n) values for each AUE allophone
Transition to the next AU
Transition to the next syntagma
Allophone sequence with indication of target

Figure 8-17. Functional diagram of synthesis acoustic-prosodic characteristics

Further, using the normalized portraits *F_0-AUP*, *A-AUP*, *T-AUP* for the syntagma of the corresponding intonational type, the values *$F_0(n)$*, *A(n)*, *T(n)* are calculated for each *n*th allophone of the elements of the pre-core, core and post-core of the *j*th AU.

The process of calculating the absolute values of prosodic parameters for each allophone of the syntagma is shown in Fig. 8-18 on the example of calculating *F_0* for the phrase "*Мариана приехала?*", the phonemic record of which is "*m a r' i a n a p r' i j' e h a l a*". This phrase is an interrogative syntagma consisting of two AUs. The corresponding pitch portrait selected from the DB of prosodic portraits is shown in Fig. 8-18a, where the abscissa axis *T_N* corresponds to the normalized time, and the ordinate axis *F_N* to the normalized value *F_0*.

The next stage – the mark-up of each syntagma AU into pre-core, core, post-core and the division of the pitch portrait in accordance with the number of phonemes in the pre-core and post-core segments – is shown in Fig. 8-18b. The syntagma under consideration consists of 2 AUs: "***m a r' i a n a***" and "***p r' i j' e h a l a***". The pre-core of the first AU contains four phonemes: "m, a, r', i"; the core of the AU, as mentioned above, is the stressed vowel, in this case it is "***a***"; the post-core of the first AU contains the phonemes "***n, a***". The pre-core, core and post-core of the second AU have, respectively, the following composition: ***"p, r', i, j'", "e", "h, a, l, a"***. In Fig. 8-18b, the abscissa axis corresponds to the so-called "phonemic" time ***T_{Ph}***, when all phonemes of the syntagma are assumed to be of the same duration.

At this mark-up stage, it is necessary to take into account such situations as the absence of a pre-core or post-core in an AU, as well as the absence of voiced phonemes within the pre-core and post-core. Indeed, if in such cases a "truncated" portrait is used, i.e. without a pre-core or without a post-core, the pitch pattern will not be fully realized, and intonation distortion will occur. To avoid this, it is necessary to mark the initial or final part of the stressed vowel as the pre-core or the post-core, respectively.

The next step (see Fig. 8-18c) is the mark-up of the intonational portrait in accordance with the proper duration of the phonemes. As can be seen from the figure, the AU cores (stressed vowels), as well as the final vowel of the syntagma, have a relatively large proper duration; in this case, the consonant "*j'*" has the smallest proper duration. The abscissa axis in Fig. 8-18c corresponds to real time *T*.

At the next stage shown in Fig. 8-18d, the proper lengths of phonemes are corrected in accordance with the rhythmic portrait of a syntagma of this type, selected from the DB of prosodic portraits. At the top of Fig. 8-

18e, the proper durations of the phonemes of the syntagma are shown; in the lower part, there are the fractions that determine the correction process. As can be seen from the figure, the duration of the core of the first AU of a two-accent syntagma of interrogative type increases, while the duration of the core of the second AU decreases. Along the abscissa in Fig. 8-18d, time T_P is shown, normalized in accordance with the rhythmic portrait of the syntagma.

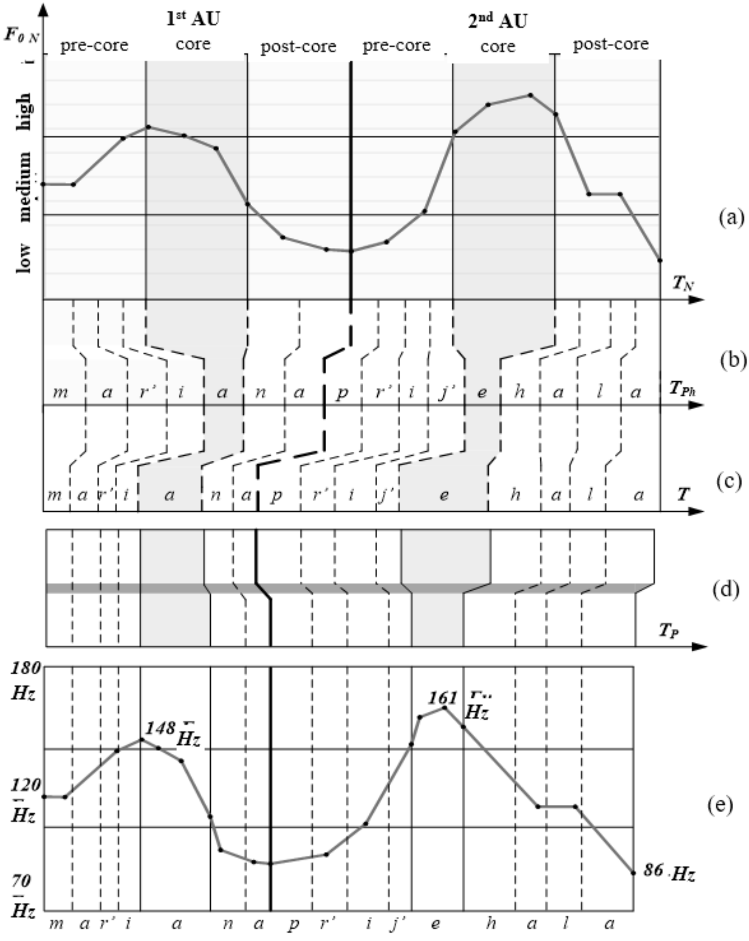


Figure 8-18. Process of calculating the absolute values of the sound duration *T* and the pitch frequency F_0

The last stage (see Fig. 8-18e) is the calculation of the absolute values of the pitch frequency F_0 for each phoneme of the syntagma. Since the pitch portrait determines normalized values, in order to calculate the absolute values, it is necessary to set the variation range for F_0. At the same time, the type of synthesized voice will affect the variation range of prosodic parameters: female, male or child; text style, expression of emotions, etc. The absolute value of F_0 is calculated by the formula:

$$F_0 = F_{0\,min} + F_{0\,norm}(F_{0\,max} - F_{0\,min}) \tag{8.1}$$

In the case under consideration, $F_{0\,min}$ = 70 Hz, $F_{0\,max}$ = 180 Hz. The obtained absolute values for F_0 are shown in Fig. 8-18e; the maximum value achieved in the core of the second AU turned out to be 161 Hz; the minimum value achieved at the end of the final vowel of the syntagma was 86 Hz.

Similar conversions are performed in the prosodic processor to calculate the absolute values of the amplitude and duration of each sound of the syntagma.

8.5.2. Unit for synthesizing acoustic-phonetic characteristics of speech signal

The structure of the unit for synthesizing acoustic-phonetic characteristics of speech signal is shown in Fig. 8-19.

The purpose of this unit is to synthesize a speech signal in accordance with the output data of the unit for generating acoustic-prosodic characteristics of speech (see Fig. 8-17) by selecting AU-elements from the DB of sound waves, their concatenation, the synthesis of F_0-modified sound waves and the formation of the speech tempo.

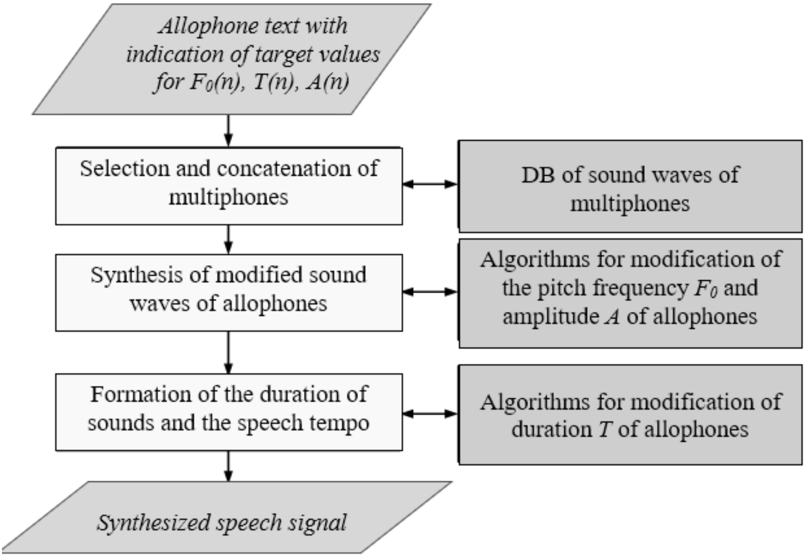

Allophone text with indication of target values for $F_0(n)$, $T(n)$, $A(n)$
Selection and concatenation of multiphones
DB of sound waves of multiphones
Synthesis of modified sound waves of allophones
Algorithms for modification of the pitch frequency F_0 and amplitude A of allophones
Formation of the duration of sounds and the speech tempo
Algorithms for modification of duration T of allophones
Synthesized speech signal

Figure 8-19. Structure of the unit for synthesizing acoustic-phonetic characteristics

Selection and concatenation of allophones and multiphones. The use of a basic set of allophones provides the synthesis of quite intelligible speech from an arbitrary text; however, the quality of speech may not be high enough. This is explained by the fact that the real variety of overtones of phonemes during their interaction in the speech flow is disproportionately greater than is provided by the minimal set of allophones used. In addition, the mutual influence of neighboring allophones in some cases may be so strong that it is often simply impossible to draw a clear line between them. Such cases include, in particular, combinations of allophones vowel-vowel, vowel-sonorant, sonorant-sonorant. A significant increase in the quality and naturalness of speech can be achieved if not only allophones, but also larger phonetic segments are used as compilation elements, that is, diallophones (a sequence of two consecutive allophones) and allosyllables (syllabic segments, taking into account positional and combinatorial variability). However, it should be borne in mind that a sharp increase in the volume of the phonetic-acoustic DB can become the price for achieving higher quality. Indeed, a rough calculation of the potential number of diallophones is estimated by the following number: $N_{da} = N^{2}_{a} = 561^2 =$ 314 721. Not all combinations of allophones are possible, but experience shows that their number in continuous speech can reach tens of thousands.

Synthesis of modified sound waves of allophones. The formation of the target values of the pitch frequency, which entails the modification of the periods of the natural speech signal, should be performed with the maximum possible preservation of the individuality and sound quality of the speech. To form the pitch pattern of *$F_0(t)$*, the SL-algorithm (Lobanov, 1991) is used, which allows for a "sparing" modification of the pitch frequency by "soft lacing" of adjacent periods of the natural signal at the intervals of the open glottis, keeping the speech signal unchanged on the rest intervals.

According to the theory of speech production, the segment of the closed glottis, where the most intense formant vibrations are realized, carries the largest amount of information about sound. Therefore, the modified speech signal (allophone) must be marked-up into pitch frequency periods in such a way that the period boundary indicates the moment of time immediately preceding the beginning of the closure of the vocal cords. With such a mark-up, the second half of the period is used to modify the periods, which corresponds to the segment of the open glottis. The first half remains unchanged.

An example of such an allophone extracted from the DB at the stage of selecting /segments of a natural speech wave is shown in Fig. 8-20. The boundaries of the pitch frequency periods marked in the figure by vertical dashed lines are set at the zero crossing points of the signal, which correspond to the moment the vocal cords start to close. The duration of one period of the pitch frequency T_0 of the presented allophone is 10 ms, and the pitch frequency F_0 is 100 Hz.

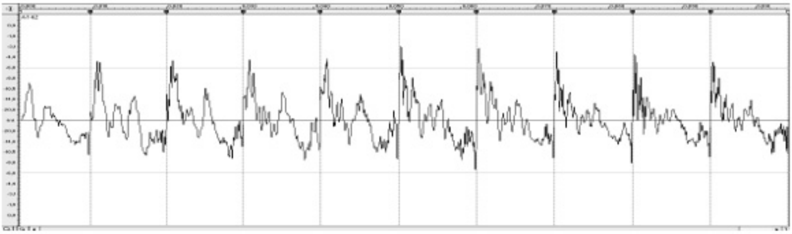


Figure 8-20. Fragment of vocalized allophone *A_{142}* with markers of pitch periods

To change the pitch frequency values, it is necessary to increase or decrease the duration of each period of the allophone.

If the procedure for reducing the duration of the period is performed by simply cutting off the “extra” segment, then signal distortions will occur and the sound quality of the speech signal will degrade significantly. An example of such a change in the duration of the period is shown in Fig. 8-21, 8-22. Fig. 8-21 shows two successive periods of the signal pitch, the proper duration of the pitch period for the signal *T_0*, the target duration of the period *T'_0*, as well as the portion of the signal to be removed. The result of the removal is shown in Fig. 8-22.

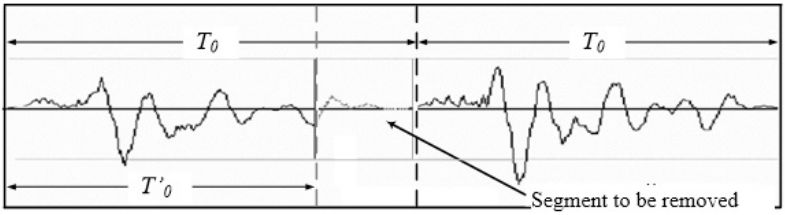


Figure 8-21. Process of reducing the pitch period

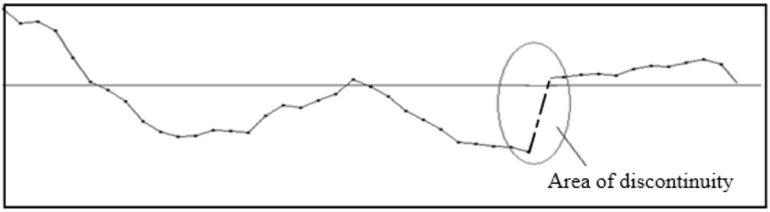


Figure 8-22. Signal discontinuity

Such signal discontinuities are perceived in synthesized speech as distinctive clicks, the presence of which significantly degrades the quality of speech.

To eliminate discontinuities, soft lacing of two adjacent periods is used. In this case, the removed segment “moves” to the left and “overlaps” the previous segment of the same period, as shown in Fig. 8-23, 8-24.

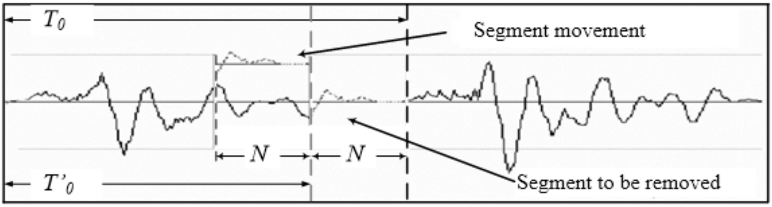


Figure 8-23. Movement of the segment to be removed

"Overlapping" of the two segments (Fig. 8-24) occurs by multiplying each of them by the characterizing lines L_1 and L_2; here the value of L_1 is 1 at the start point and 0 at the end point, and the value of L_2 is 0 at the start point and 1 at the end point.

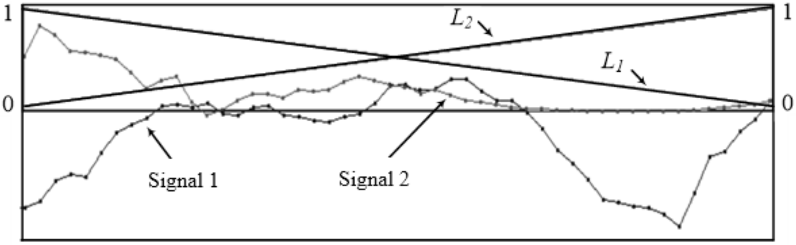


Figure 8-24. "Overlapping" of the segments of the two signals

Formation of the duration of sounds and the speech tempo. The values of the duration of the sound waves ***T(t)*** for allophones are set in accordance with the specified target values of the duration for the AU sound elements and then adjusted taking into account its qualitative and quantitative composition. The speech tempo is modified by adjusting the duration of the AU sound elements and inter-syntagmal pauses, taking into account the coefficient of "susceptivity" of each particular sound to tempo changes.

8.5.3. Algorithm for speech signal prosodic processing in the acoustic processor

In accordance with the text of the current syntagma, the required sequence of allophones and (or) multiphones is selected from the DB of sound waves; then their series combination (concatenation) takes place.

According to the specified prosodic characteristics of the syntagma, the required prosodic portraits of AUs are selected and then the current values *$F_0(t)$, $A(t)$, $T(t)$* of allophone sound waves are formed.

The general structure of the algorithm is shown in Fig. 8-25.

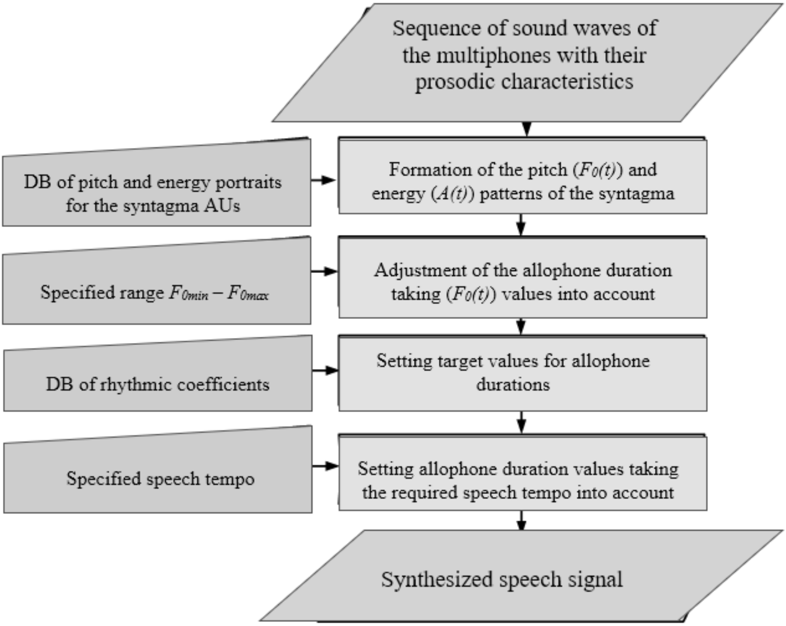

Sequence of sound waves of the multiphones with their prosodic characteristics
DB of pitch and energy portraits for the syntagma AUs
Formation of the pitch ($F_0(t)$) and energy ($A(t)$) patterns of the syntagma
Specified range $F_{0min} - F_{0max}$
Adjustment of the allophone duration taking ($F_0(t)$) values into account
DB of rhythmic coefficients
Setting target values for allophone durations
Specified speech tempo
Setting allophone duration values taking the required speech tempo into account
Synthesized speech signal

Figure 8-25. Structure of the algorithm for speech signal prosodic processing

Formation of the pitch ($F_0(t)$) and energy ($A(t)$) patterns. The operation of the algorithm can be shown by the example of the formation of the pitch ($F_0(t)$) and energy ($A(t)$) patterns for the syntagma of the text "*Машенька открыла глаза*", consisting of three AUs, the intonational type of which is C3.

The allophone notation of this syntagma is as follows:

1st AU: M$_{002}$,**A$_{012}$**,SH$_{001}$,E$_{323}$,N'$_{003}$,K$_{002}$,A$_{232}$
2nd AU: A$_{222}$,T$_{001}$,K$_{002}$,R$_{002}$,**Y$_{021}$**,L$_{002}$,A$_{212}$
3rd AU: G$_{002}$,L$_{002}$,A$_{21}$2,Z$_{002}$,**A$_{020}$**

The core of each AU is marked in bold. The cores are fully-stressed vowels recognized by their presence in the list {A, E, I, O, U} and the first digital index {0}.

The pre-core of each AU are allophones to the left of the core, and the post-core are allophones to the right of the core. In the 3rd AU, there is no post-core.

In each AU, ***noise*** allophones (as opposed to ***voiced*** ones) are marked, which are not processed during the synthesis of the pitch pattern. Such allophones are recognized by their presence in the list: {*b, d, g, p, t, k, z, zh, f, s, sh, h, b', d', g', p', t', k' , z', f', s', sh', h', c, ch'*} regardless of their digital indices.

Figure 8-26 shows a pitch portrait of a three-accent syntagma of the C3 intonational type, taken from the DB of intonational patterns. Each AU portrait contains three segments of the same length: pre-core, core and post-core, each of which consists of 100 equally spaced samples of the pitch (intonation) curve.

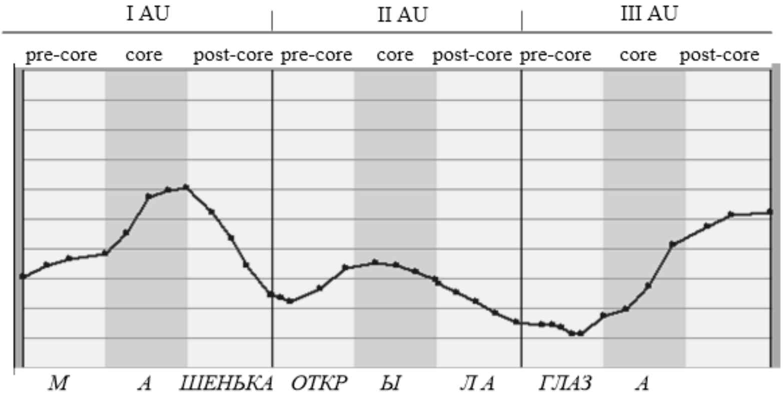


Figure 8-26. Pitch portrait of a three-accent syntagma of the C3 intonational type

The pitch portrait describes the movement of the pitch frequency normalized (F_0^N) from 0 to 1. In order to find the absolute values of the pitch frequency, it is necessary to set the minimum ($F_{0\ min}$) and the maximum ($F_{0\,max}$) values characteristic of the synthesized voice and make calculations by the following formula:

$$F_0 = F_0^N(F_{0\,max} - F_{0\,min}) + F_{0\,min} \tag{8.2}$$

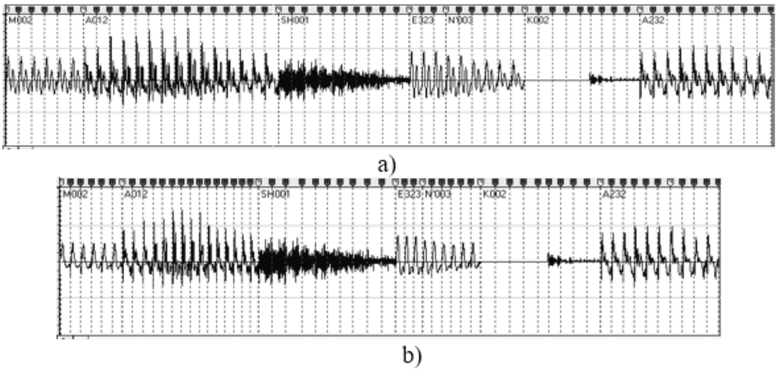
When using the DB of allophones of a male voice, we choose $F_{0\,max}$ = 200 Hz, and $F_{0\,min}$ = 70 Hz. If required, by changing these values, one can change the frequency range for the same voice (up to monotonous speech), or shift the pitch of the voice.

Formula (8.2) is used further to calculate the duration of each period of voiced (not unvoiced) allophones. The duration of the current period (*N_0* in the number of signal samples) is determined by the following formula:

$$N_0 = \frac{F_d}{F_0} \tag{8.3}$$

where F_d is the speech signal sampling rate.

Adjustment of the allophone duration taking pitch frequency values into account. Fig. 8-27a shows the waveform of the first syntagma AU, obtained by directly compiling it from the DB of allophones, and Fig. 8-27b shows its waveform after modifying the duration of the periods of each of the voiced allophones in accordance with the AU pitch portrait.



a)

b)

Figure 8-27. Illustration of the change in the duration of allophones

As can be seen from the comparison of Fig. 8-27a and 8-27b, the duration of unvoiced allophones during the pitch frequency modification in accordance with the described algorithm remained unchanged; however, the duration of voiced allophones changed quite significantly in some cases.

Setting the required speech tempo. The speech tempo is modified by adjusting the duration of allophones and inter-syntagmal pauses, taking into account the coefficient of “susceptivity” of each particular sound to tempo changes. An experimental assessment of the limits of change in the relative duration of allophones and pauses with a change in tempo is given in Table 8-2. It is believed that the average tempo corresponds to the duration of allophones included in the DB.

Table 8-2. Relative duration of allophones with a change in tempo

No.	Acoustic units	Slow tempo (***K****max*)	Medium tempo	Fast tempo (***K****min*)
1.	Pauses	2.5	1.0	0.2
2.	Stressed vowels (index of 0 or 1)	2.0	1.0	0.5
3.	Pre-stressed vowels (index of 2)	2.0	1.0	0.7
4.	Post-stressed vowels (index of 3)	2.0	1.0	0.8
5.	Consonants	1.3	1.0	0.8

The rhythmic pattern is synthesized on a syntagma-by-syntagma basis, i.e. by buffering a sequence of Wav-files of allophones included in the syntagma.

The rhythmic pattern synthesis procedure is based on the calculation of ***new*** values of allophone durations $T^N a_i$, based on the set of factors listed above. The calculation takes place in accordance with the following formula:

$$T^N a_i = K_p(K_{min} * TMP + (1 - TMP) * K_{max}) * Ta_i \qquad (8.4)$$

where K_p is a prosodic coefficient; ***TMP*** is the desired speech tempo set within the interval of (0 - 1); K_{min} is the coefficient of the minimum possible reduction of the allophone; K_{max} is the coefficient of the maximum possible elongation of the allophone; Ta_i is the duration of the ith allophone.

Zero corresponds to the fastest tempo, unity corresponds to the slowest tempo. The coefficients K_{min} and K_{max} are taken from Table 8-2. They must be different for different classes of allophones. The duration of the ith allophone Ta_i is determined by the DB of allophones. Durations: original Ta_i and prosodically modified $T^N a_i$ are set by the number of samples in the allophone signal.

8.6. Software implementation of the speech synthesis system "Multiphone"

The functional diagram, input and output data, the interaction of units in the speech synthesis system are shown in Figure 8-28.

The system implements the algorithms for processing text and speech signals described above.

System input data – an orthographic text contained in a text file or entered from the keyboard.

System output data – a synthesized speech signal fed to an audio output device or saved to a file in the WAVE PCM format.

At the first stage of synthesis, cleaning and morpho-syntactic processing of the text take place. Also, symbols that are not included in the set of symbols acceptable for the synthesis of Russian speech are removed from the text; abbreviations and acronyms are decoded; word stresses are placed and a morphological category (MC) is indicated for each word of the text. To perform the first stage of text processing, an orthographic dictionary is used, as well as lists of abbreviations and acronyms.

The normalized text with indication of morphological categories and word stresses enters the unit for segmenting into syntagmas and combining words into AUs. In this unit, on the basis of the indicated MCs, enclitics and proclitics are attached to significant parts of speech, the boundaries of phonetic words and AUs are determined, and the text is segmented into syntagmas.

The next unit of the speech synthesizer performs marking of intonational types of syntagmas. The operation of this unit results in a list of syntagmas with indication of the boundaries of phonetic words and AUs, as well as the intonational type of each syntagma; this list is then subjected to phonetic processing.

In phonetic processing, which implements the replacement of words-phonetic exceptions with equivalents, as well as letter-phoneme and phoneme-allophone conversions, a list of allophone sequences is formed, each of which retains the marks of the AU boundaries obtained at the previous stages of processing.

In the next unit (the unit for synthesizing acoustic-prosodic characteristics), based on the data from the prosodic DB, the target values of the prosodic parameters (pitch frequency, amplitude, duration) of each allophone of each syntagma of the input list are calculated. At this stage of processing, the user can select a prosodic DB, as well as specify the required tempo of speech being synthesized.

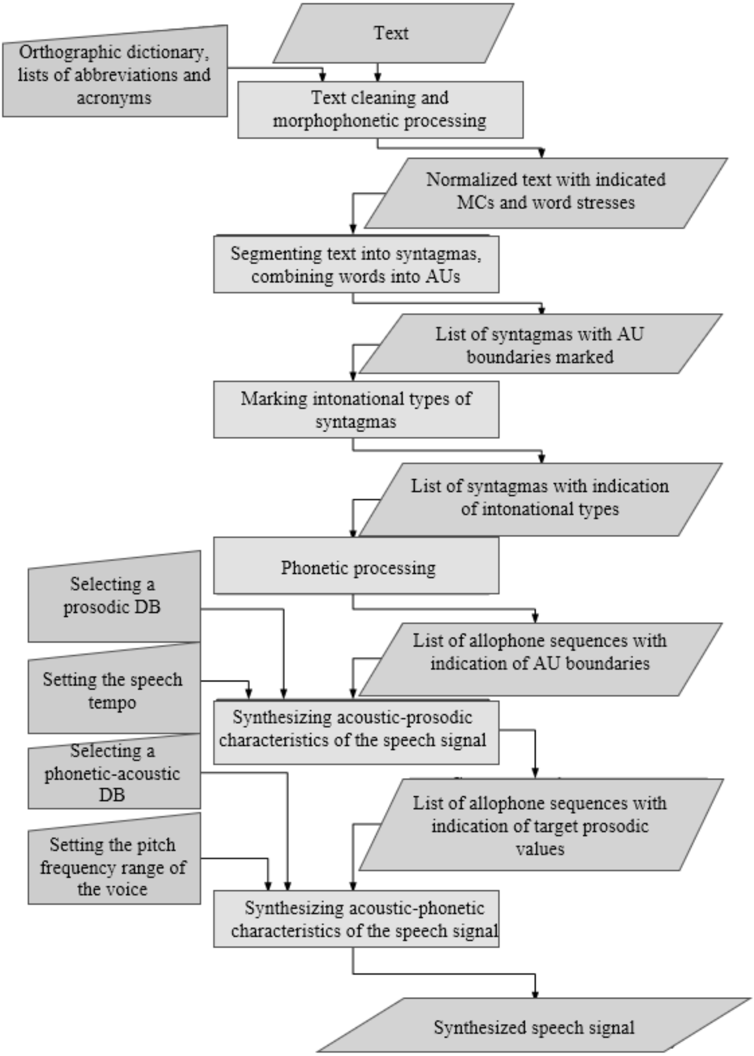

Orthographic dictionary, lists of abbreviations and acronyms
Text
Text cleaning and morphophonetic processing
Normalized text with indicated MCs and word stresses
Segmenting text into syntagmas, combining words into AUs
List of syntagmas with AU boundaries marked
Marking intonational types of syntagmas
List of syntagmas with indication of intonational types
Phonetic processing
Selecting a prosodic DB
List of allophone sequences with indication of AU boundaries
Setting the speech tempo
Synthesizing acoustic-prosodic characteristics of the speech signal
Selecting a phonetic-acoustic DB
List of allophone sequences with indication of target prosodic values
Setting the pitch frequency range of the voice
Synthesizing acoustic-phonetic characteristics of the speech signal
Synthesized speech signal

Figure 8-28. General functional diagram of the text-to-speech system

The acoustic-phonetic characteristics of the speech signal are synthesized by selecting the required allophones or multiphones from the phonetic-acoustic DB, their concatenation and modification of the signal in accordance with the target values of the prosodic parameters. At this stage, the user can select a phonetic-acoustic DB, as well as specify the range of pitch frequency change for the selected DB.

The operation algorithms of the blocks of the speech synthesizer correspond to the main provisions discussed in sections 8.1 - 8.5.

Conclusion

The development of the software model of rule-based multi-wave speech synthesizer described in this chapter dates back to the beginning of the 1st decade of the 21st century. Employees of the Laboratory of Speech Recognition and Synthesis (see: www.ssrlab.by) of the Joint Institute for Informatics Problems of the National Academy of Sciences of Belarus (see: www.uiip.bas-net.by) took part in its creation. The most significant contribution to the development and software implementation was made by Vitaly Kiselev, Dmitry Zhadinets, Andrey Davydov, Lilia Tsirulnik, Yuras Getsevich. With their participation, a new technology and special software tools for cloning the phonetic and intonation characteristics of the speech of a particular person are also being created - "PhonoClonator" and "IntoClonator". The developed software tools made it possible to provide personalization of the sound of synthesized speech. The statement of the problem of speech cloning was previously presented in (Lobanov and Karnevskaya, 2002).

The achieved level of intelligibility and quality of the synthesized speech were quite satisfactory in comparison with the speech of other well-known synthesizers of that time. Below are Internet links for three examples of speech synthesizer sounds:

1. Reading with a synthesizer an excerpt from a humorous story by Semyon Altov “*Who is there*?”: https://storage.googleapis.com/intontrainer.by/files/Story_Altov.mp3
2. Examples of sounding different voice clones synthesized for 2 men and 2 women: https://storage.googleapis.com/intontrainer.by/files/A%2BB%2BO%2BV_Clones.mp3
3. An example of a synthesizer performance based on musical notation and the text of the song “*March of the High-Rises*”: https://storage.googleapis.com/intontrainer.by/files/March%20of%20workers.mp3.wav

The theoretical and experimental results obtained in the early 2000s made it possible to carry out a number of joint scientific and applied projects.

In 2005 - 2007 the international INTAS-project “*Development of a polyphonic and multilingual Text-to-Speech system (TTS) and a Speech-to-Text system (STT) (languages: Belarusian, Polish, Russian*)” was completed. Project participants: Belarus, Germany, Poland, Russia.

In 2006 - 2007 On the basis of the speech synthesizer "Multiphone" (its SAPI-version), together with LLC "INVOSEVIS" (Minsk), the "*Reader system*” was developed, oriented for special schools for the blind in Belarus.

In 2007 - 2008 jointly with Telecontent LLC (Moscow), a speech synthesis system for mobile phones has been developed, which makes it possible to create a new type of mobile service - "*Audio-Book*".

In 2008, together with the Research Institute of the Russian Academy of Sciences (St. Petersburg), a multimedia system for audiovisual speech synthesis was developed - "*Talking Head*".

References

Lobanov, B.M. 1983. Research and development of methods for automatic speech synthesis based on phonemic text. Dr.Sc. diss.: 05.13.01. – Riga, 323 p. (in Russian)

Lobanov, Boris M. 1987. The “Phonemophone” Text-to-Speech System. Proc. of the XI-th International Congress of Phonetic Sciences ICPhS’87,Tallinn, USSR: 120-124.

Lobanov, Boris M. 1991. “MW-TTS-Synthesis”. Proc. of the XII International Congress of Phonetic Sciences ICPhS’91, Aix-en-Provence, France: 128-132.

Lobanov, Boris, Karnevskaya, Helena. 2002. TTS-Synthesizer as a Computer Means for Personal Voice “Cloning”. Phonetics and its Applications. Festschrift for Jens-Peter Köster on the Occasion of his 60th Birthday. – Stuttgart: Franz Steiner Verlag: 445–452.

Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya H. 2006. Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis. Speech Prosody: proceedings of the 3rd International conference. Dresden, Germany, May 2–5, V. 2: 553-556.

Lobanov, Boris M., Tsirulnik, Liliya I. 2006a. Development of Multi-voice and Multi-language TTS Synthesizer (languages: Byelorussian, Polish, Russian)”. Proc. 11-th International Conference “Speech and Computer” SPECOM’2006, June 25-29, Saint-Peterburg, Russia: 274-283. (in Russian)

Lobanov, Boris M., Tsirulnik, Liliya I. 2006b. “Intrawords and Interwords Rules of Phonemic Text Processing for Full and Conversational

Speech Styles". Proc. International Conference "Functional Styles of Sounding Speech", September 5-7, Moscow, Russia: 80-89. (in Russian)

Zaliznyak A.A. 1987. Grammaticheskij slovar' russkogo yazyka. Slovoizmenenie. [Grammatical dictionary of the Russian language. Inflection]. Moscow: Russkij yazyk. 880 p. (in Russian)