UDC 631. 391

B.M. Lobanov

ALLOPHONIC TEXT-TO-SPEECH SYNTHESIZER: GENERAL STRUCTURE AND DESCRIPTION

A structure of the text-to-speech conversion algorithm based on an allophonic representation of the text is described. The serially connected textual, prosodic, phonemic and signal processors organize text-to-speech conversion. The output of the textual processor is prosodically marked phonemic text. The role of the prosodic processor is the generation of pitch, duration and intensity contours of synthetic sounds. The phonemic processor transforms a prosodically labeled phonemic text in to acoustical parameters of the allophones. This information comes to the signal processor that generates the synthetic speech.

Introduction

There are several approaches to the building of text-to-speech systems. The main point is: what kind of minimal elements should be taken for the proper approximation of the phonemic structure of the reading text? Nowadays two kind of minimal elements most commonly used for the approximation of the phonemic structure: diphones [1] and microwaves [2]. Also in [3] allophonic units were suggested as minimal elements. The advantage of allophonic elements usage is the possibility to reach a compromise between the potential quality of the synthetic speech and the volume of the memory required for the storage of the speech units. This paper describes in details a structure of text-to-speech conversion algorithm based on an allophonic representation of the speech units that was developed in the Laboratory [4].

1. General structure of text-to-speech conversion algorithm

A general structure of the text-to-speech conversion algorithm is shown in fig. 1. The input of the synthesizer is an orthographic text while the output is a synthetic speech signal. Text-to-speech conversion is organized by the serially connected textual, prosodic, phonemic and signal processors. The output of the textual processor, as well as the input of the prosodic processor, is prosodically marked phonemic text that generated by textual processor using Knowledge Base (KB) such as dictionary, morphological and syntactic knowledge. The role of the prosodic processor is the generation of pitch, duration and intensity contours of synthetic sounds according with the given text and the KB which contained the pattern contours of pith, duration and intensity as well as the rules for there modification. Next comes the phonemic processor that transforms the prosodically labeled phonemic text in to acoustical parameters of the allophones

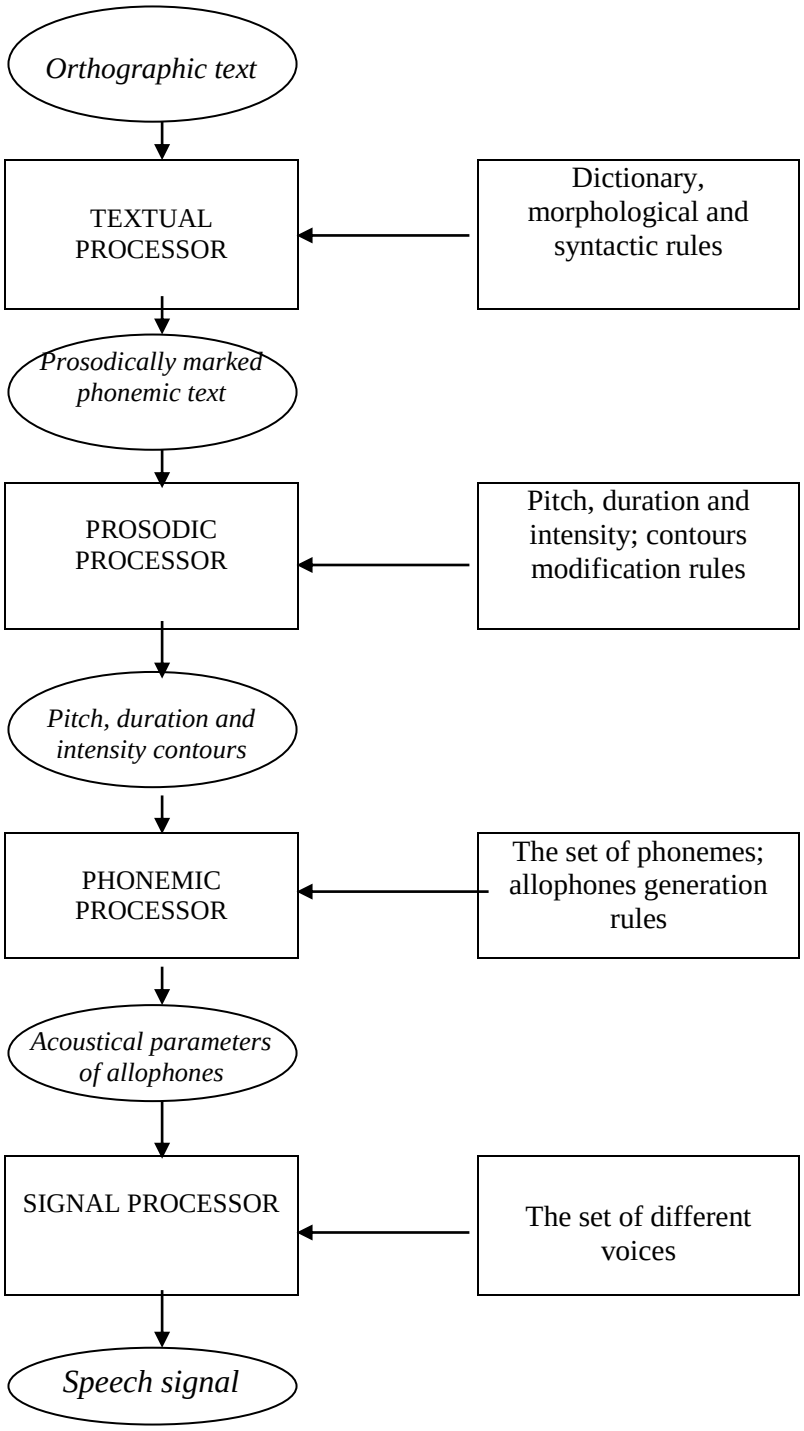

Fig. 1. General structure of allophonic text-to-speech synthesizer

by using KB: the set of phonemes of the given language and the allophone generation rules. This information comes to the signal processor which generates the synthetic speech with a certain voice chosen from the database. More details of the functioning of each processor are to be found in fig. 2-8.

2. Textual processor

A general configuration of the textual processor is shown in fig. 2. The central part of Fig. 2 is a sequence of different stages of the text analysis. On the left, the types of KDB used for each stage are shown. On the right, the types of information provided by each stage of analysis are listed. Examples of the results of textual processor function are shown in fig. 3. The brackets // are used to mark phrase (sentence) boundaries, the brackets / show the positions of syntagmatic boundaries, the number of the signs # indicates the relative duration of the pauses, different punctuation marks show different possible types of intonation, the signs ' and " show the accent type (weak or strong) and its position in the word.

3. Prosodic processor

A configuration of the prosodic processor is shown in fig. 4. Prosodically marked phonemic text coming from textual processor as syntagm by syntagm is first labeled for accent groups (AG). The labeling procedure is based on a set of special rules which take into account the distribution of word accent type in the syntagm being analyzed. At the next step each ERG is divided into pre-nucleus, nucleus and post-nucleus parts using special rules and the information on the strong accent position in the ERG being analyzed. At the final stage all of the parts of the ERG are provided with certain values of the pitch (F0), duration (T) and the intensity of sounds (I). These values are determined in accordance with the intonation type of the syntagm, the ERG position in the syntagm and the given pattern contours of the F0, T, and I.

An example of the results of the prosodic processor function is presented in fig 5. The brackets [ ] show the AG boundaries and the brackets ( ) - the boundaries of the pre-nucleus, nucleus and post-nucleus. The results of the prosodic processor function include the initial and final values of the pitch F0, intensity I, and the values of the duration T for each parts of AG mentioned above.

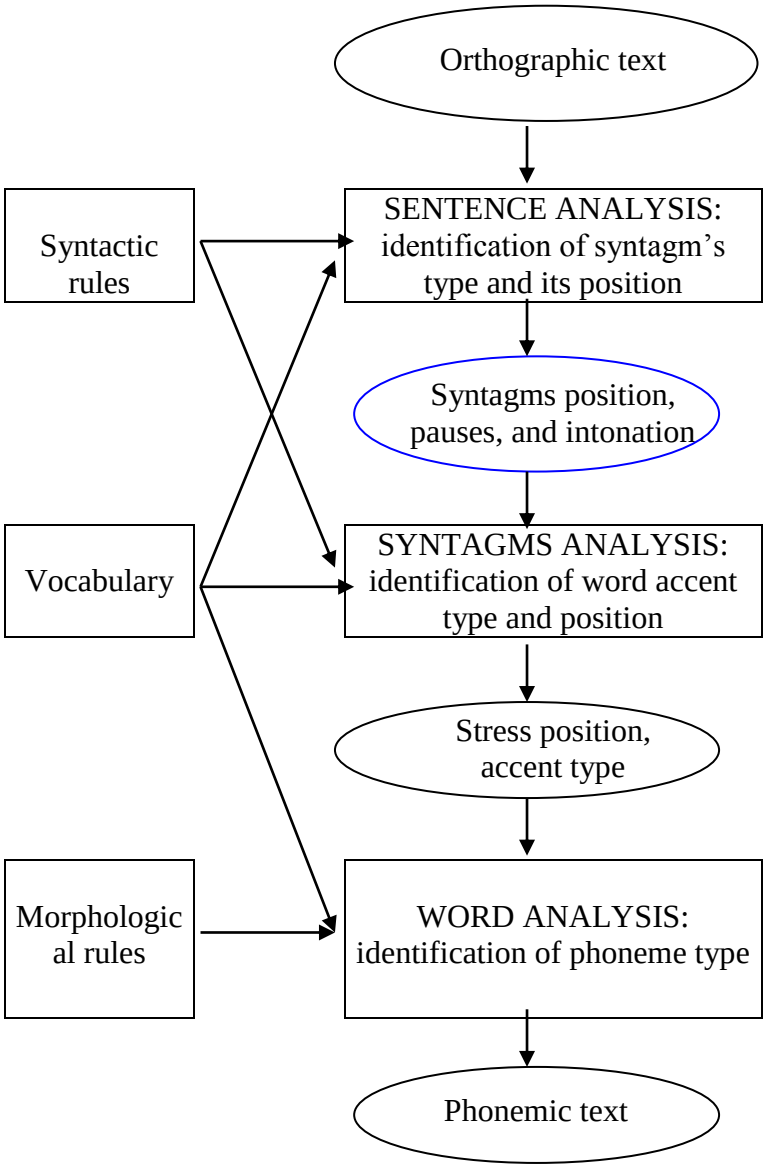

Fig. 2. Configuration of the textual processor

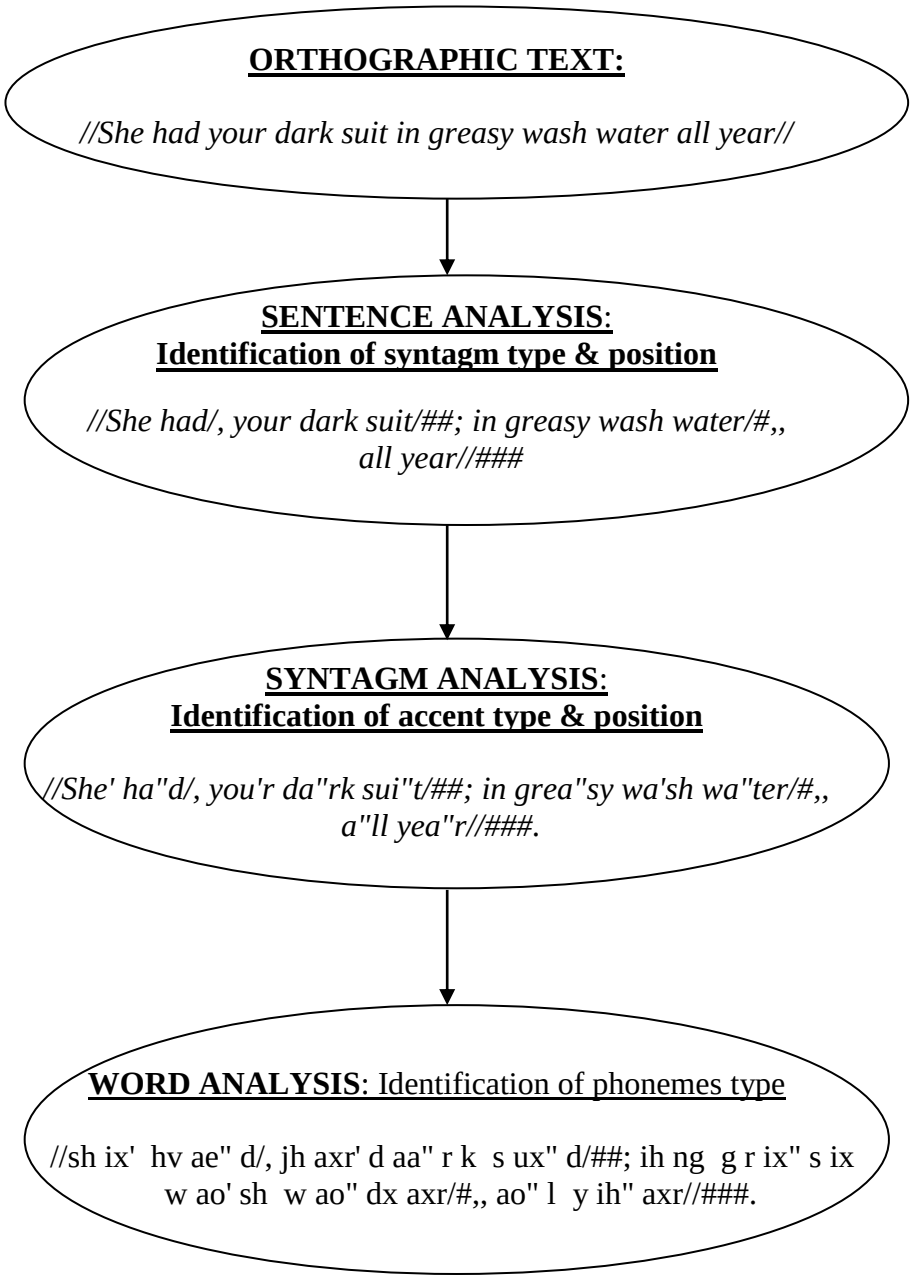

Fig. 3. An example of textual processor operation

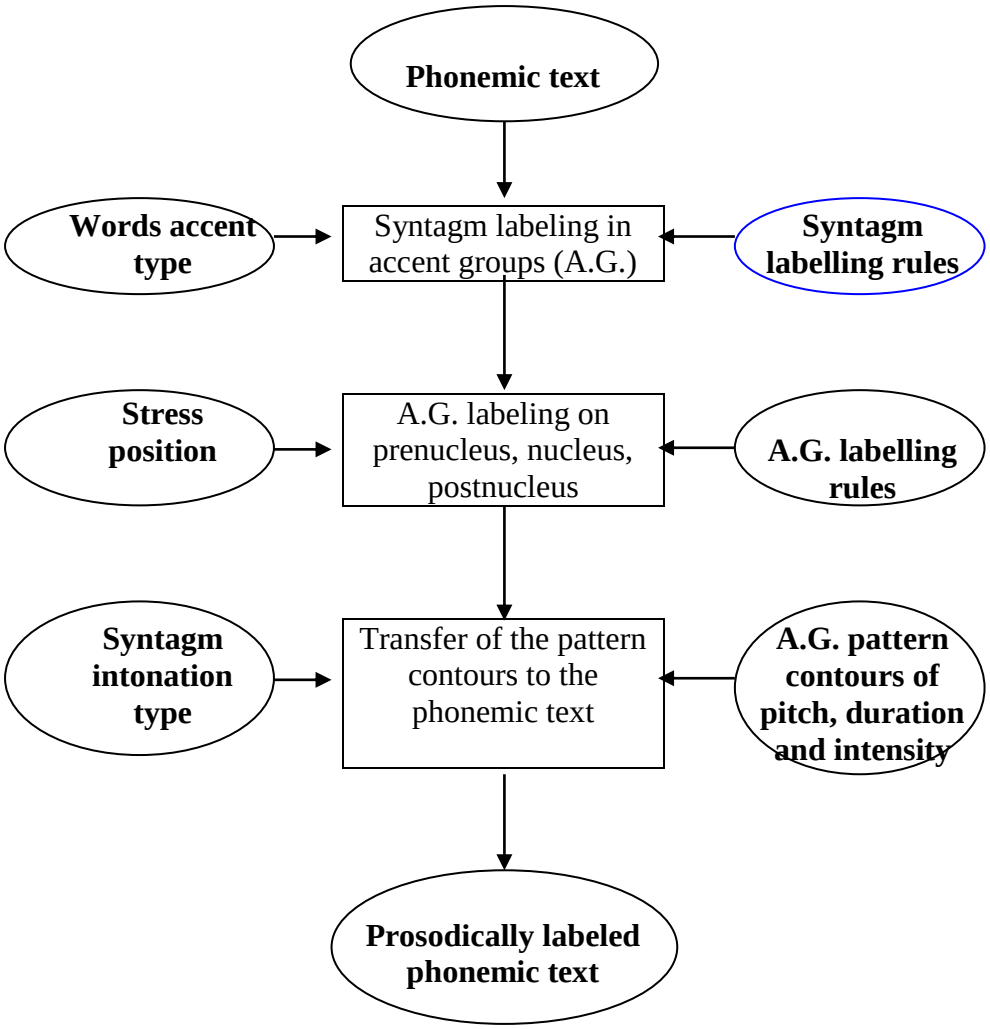

Fig. 4. Configuration of the prosodic processor

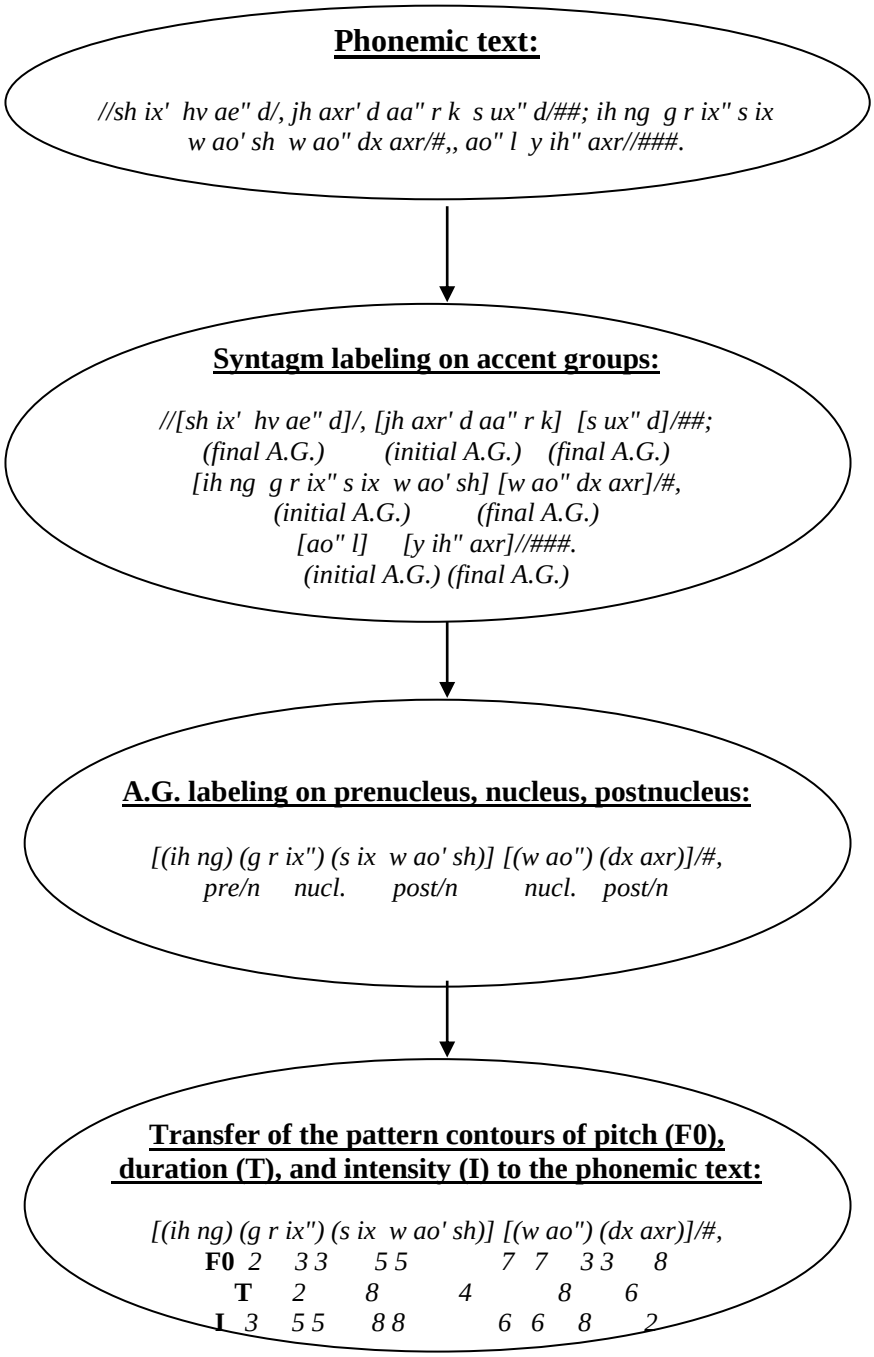

Fig. 5. An example of prosodic processor operation

4. Phonemic and acoustic processors

A configuration of the phonemic processor is shown in fig. 6. The main function of this processor is a transformation of a phonemic text into an allophonic one. At the first stage the word, AG and syntagm position of each phoneme is analyzed and a certain allophone is then chosen with the help of special rules. At the second stage the nearest environment of each phoneme is analyzed and a certain allophone is again chosen by special selection rules.

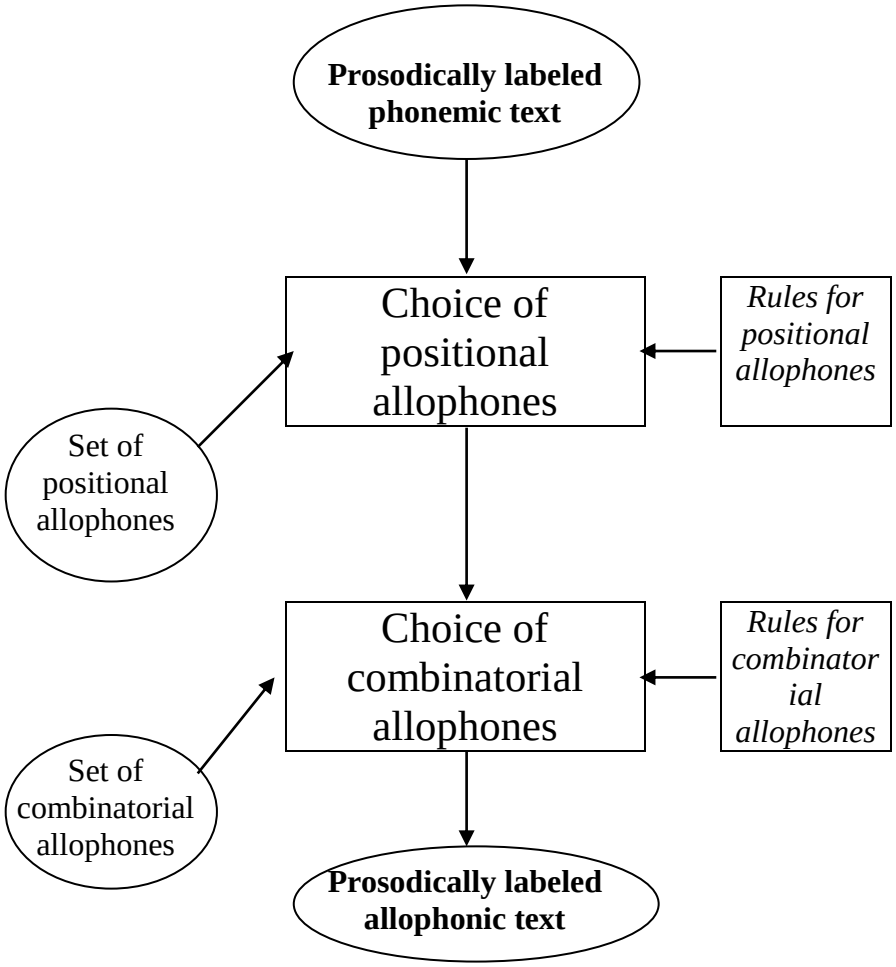

Fig. 6. Configuration of the phonemic processor

An example of phoneme-to-allophone transformation (only for vowels) is shown in fig. 7. The central indexes after the vowel indicate the type of positional allophone. The left and the right indexes indicate the type of the combinatorial allophone.

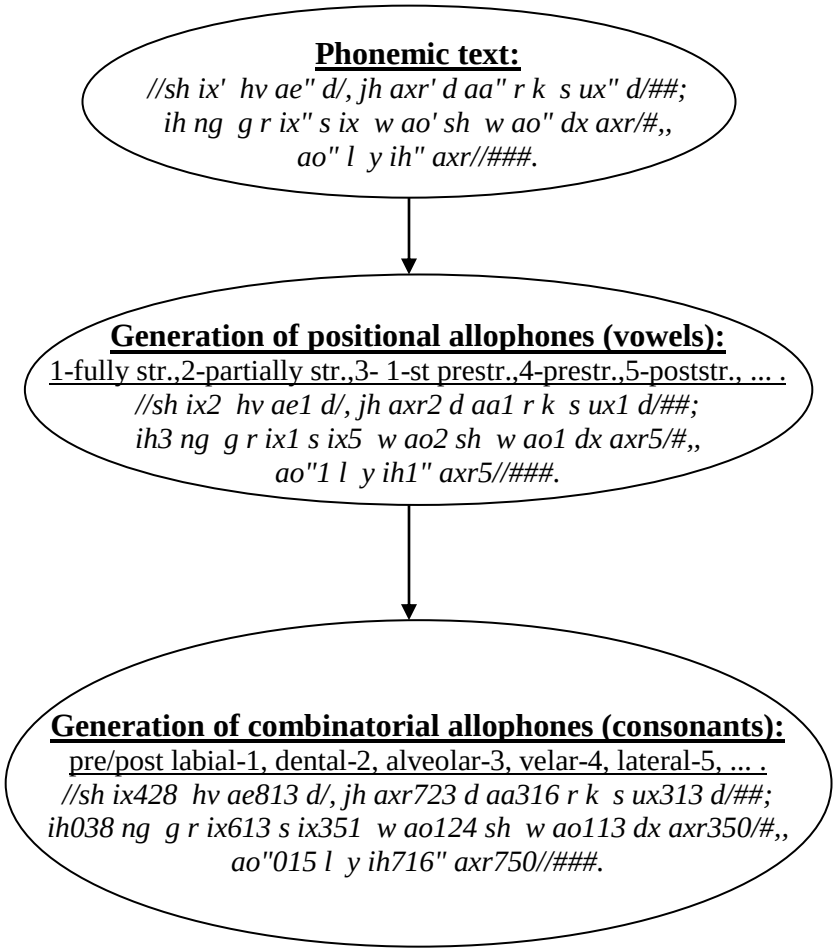

Fig. 7. Anexample of phonemic processor operation

A general structure of the acoustic processor is shown in fig. 8. Its functions are obvious from the figure.

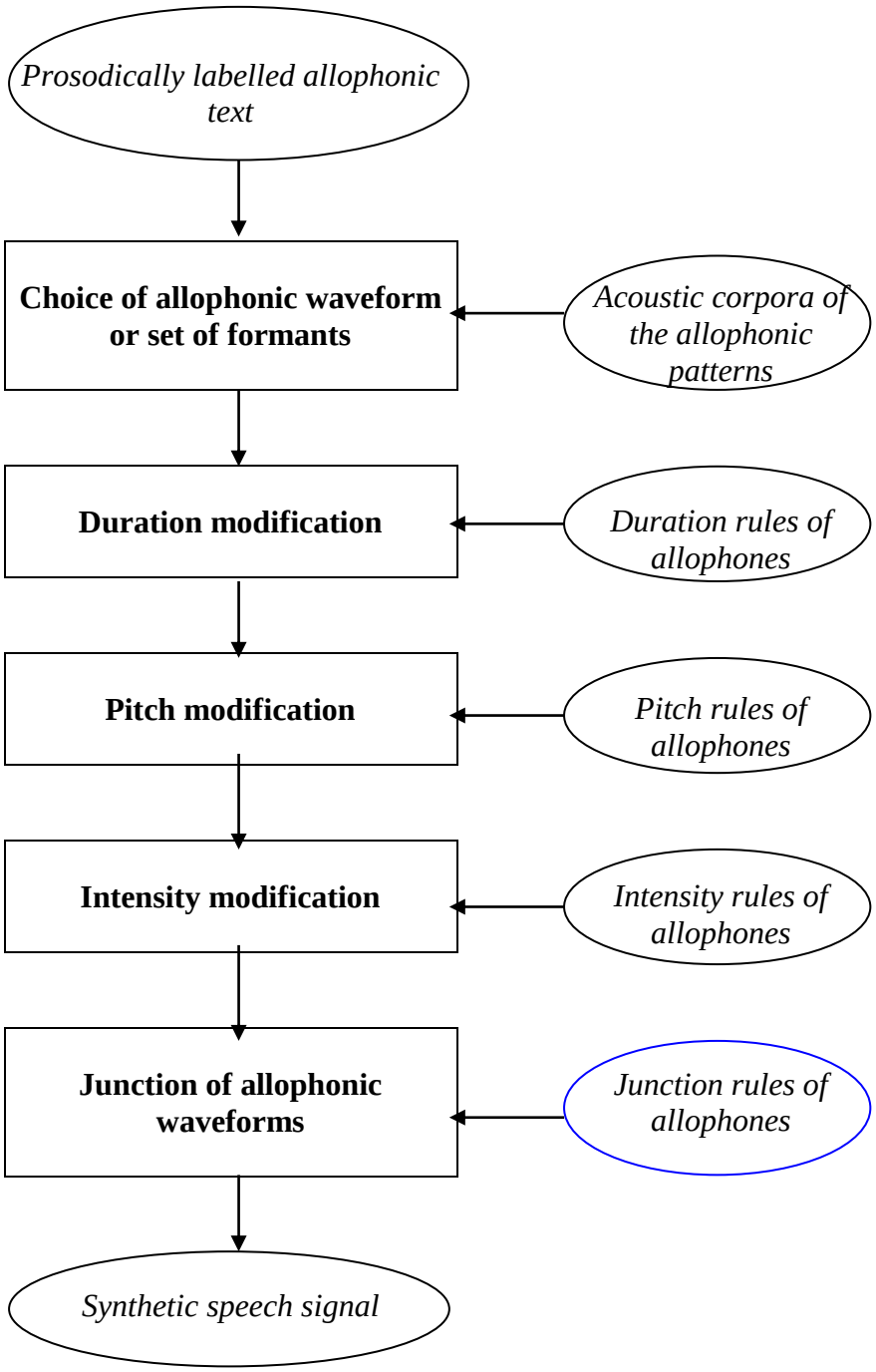

Fig. 8. Configuration of the acoustical processor

Concludion

The algorithms described above are utilized in software model of Russian text-to-speech system [4] as wall as in Russian/German synthesizer [5].

References

1. Hamon C., Muilines E., Charpenties F. A Diphone Synthesis System Based on Time Domine Prosodic Modification of Speech. Proceedings of the ICAASP,Glasgow, 1989, pp. 238-241.

2. Karnevskaya E., Lobanov B., Microwave Speech Synthesis from Text, Proceedings of the 7th International Congress of Phonetic Sciences – ICPhS '91, Volume 5, Aix-en-Provance, 1991, pp 406-409.

3. Zinovieva N. Phonetically Sufficient Allophonic Database for Concatenation Synthesis of Russian Speech. Proceedings of the ICPhS'95, V.2, Stockholm, 1995, pp. 358-361.

4.Кубашин А.В., Шаков А.М. Аллофонный синтез русской речи по орфографическому тексту. //В наст. сборнике.

5. Lobanov B., at al. A Bilingual German/Russian Text-To-Speech System. Proc. Of International Workshop "Speech And Computer". St.-Petersburg, Russia. 26-29 October, 1998, pp. 327-330.

Institute of Engineering Cybernetics
National Academy of Sciences Belarus