Using text-to-speech synthesis algorithms for solving a task of automatic generation of orthoepic dictionary of Belarusian language

Yu. Hetsevich, S. Lysy, A. Hiuntar, V. Mandzik

United Institute of Informatics Problems, Minsk, Belarus
yury.hetsevich@gmail.com

Abstract

The paper describes the development of the transcription generation system for input electronic texts in Belarusian, as well as creation of orthoepic dictionary generation system for the Belarusian language which uses text-to-speech synthesis algorithms. Schemes and working algorithms for these systems are provided.

1. Introduction

Development of the TTS synthesis system covers numerous problems to be solved. Since speech synthesis is closely related to linguistic, some tasks need to be referred to as a separate class of applied linguistic problems. Therefore, solving of the problem of speech synthesis itself implies creation of algorithms, which can be used for solving to address a number of other applications, in particular through the processing and use of intermediate and final data of the speech synthesizer in the text.

The authors of the paper set a goal to develop an automatic transcription generation system for input electronic texts in Belarusian, and to create on its basis an orthoepic dictionary generation system. This problem is actual and new, because there are no orthoepic dictionaries for the Belarusian language at the moment, and also, there is no information about existing of specialized algorithms for automatic generation of transcriptions [1].

In order to achieve the intended goal the following tasks were solved:

- development of a knowledge base with correspondence "allophone - symbol of transcription";
- implementation of program algorithms to generate three types of transcription on the base of orthographic text ;
- implementation of online access to a component for transcription generation;
- creation of the general scheme of the system for other languages.

The knowledge base for Cyrillic transcription was compiled by expert linguists using with the use of materials on the theory of the Belarusian language – [2, 3]. To compile the knowledge base of the simplified Latin transcription the paper [4], and for the international phonetic transcription the resource [5] were used.

This paper and its description are based on the results presented in the [6].

2. Description and the working algorithm for the transcription generation system

The transcription generation system takes as input any text in the Belarusian language in the orthographic form with labels of the main and side accents. We use the symbol '+' after a vowel to mark the main accent, ‘=’ to mark the side accent, and the character ‘Ъ’ to mark two actual words as one phonetic word. Apart from that, there are special buttons to input standard symbols of accent. The user can select one or several types of transcription and a number of peculiarities to present final data. The system currently supports three types of transcription: transcription with Cyrillic symbols (mark this one as Tr_1), transcription with Latin symbols according to [4] (Tr_2), and transcription in the international format (Tr_3) (International Phonetic Alphabet). The results of the algorithm are given to the user as the original text in the transcribed form with selected types of transcription and the form of the output, and also to an expert linguist and to a software engineer – in the form of all the entered input, output, and analyses (the size of the text, IP-address, time, etc.) data for prompt correction of errors of transcription and generation of general statistical analysis of the system usage. On the figure 1 the general scheme of the transcription generation system work is shown. Let’s take a closer look to the algorithm.

The following data are submitted to the input of the algorithm: a text T and a target type of the transcription T_{tr}. The algorithm performs the following steps:

Step 1. *Text pre-processing*. A certain normalization of the original text T is made (for example, the replacement of the various characters, which are often used as an apostrophe, by its standard symbol). At this stage tokenization and selection of the punctuation marks is performed.

Step 2. *Prosodic processing of the text*. Punctuation marks in the text T are replaced by intonation labels. The resulting text T_p is transmitted to the block of phonetic processing.

Step 3. *Phonetic processing of the text*. Text T_p is served to the phonetic processor TTS synthesizer, where happens the conversion of orthographic form of words into an allophonic based on formalized phonetic rules of the Belarusian language.

Step 4. *Forming of the rules “allophone – symbol of transcription”*. A query is being sent to the database, which contain correspondence “allophone – symbol of transcription”, and a set of rules $R = <R_1,\dots,R_n>$ is being formed, where $R_i = <a_i,tr_{i1},tr_{i2},tr_{i3}>$, a_i – is a code of the allophone, tr_{i1}, tr_{i2}, tr_{i3} – respective symbols of the three allophone a_i transcriptions, i = 1...n, n – number of the rules.

Step 5. *Generation of transcriptions*. Allophonic text T_a is divided into an array of allophones. For every sequentially selected in the text T_a allophone a_i generates symbol of transcription of one or another type tr_{ij} according to the set of rules R formed before. As a result of the sequential processing the whole text T_a we obtain the final transcription T_r of the selected type Tr_t.

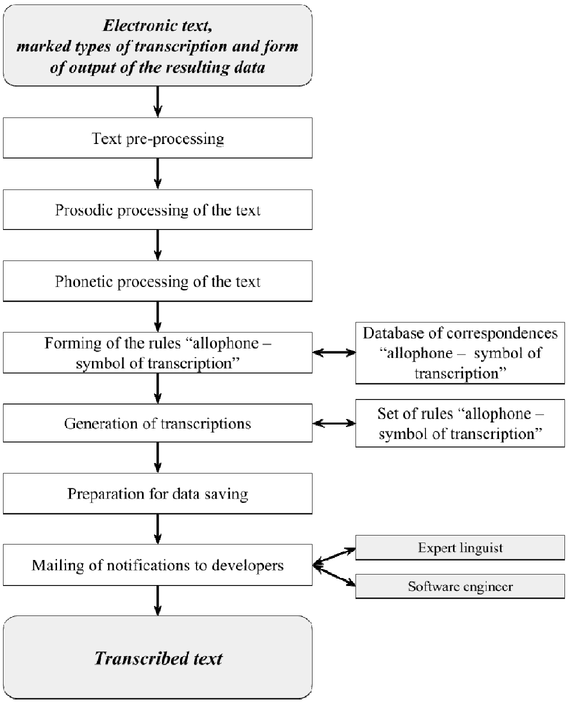

Electronic text,
marked types of transcription and form of output of the resulting data
Text pre-processing
Prosodic processing of the text
Phonetic processing of the text
Forming of the rules "allophone – symbol of transcription"
Database of correspondences "allophone – symbol of transcription"
Generation of transcriptions
Set of rules "allophone – symbol of transcription"
Preparation for data saving
Mailing of notifications to developers
Expert linguist
Software engineer
Transcribed text

Figure 1: The general scheme of the transcription generation system work

Step 6. *Preparation for data saving*. In the process of generation of transcription information about input data (original text T, the target type of transcription Tr_t), output data (resulting transcriptions Tr) and analytical data (the size of the text, IP-address, time, etc.) are simultaneously collected for prompt correction of errors of transcription generation. Also, the information about the absence of some elements in the base "allophone – symbol of transcription" is collected.

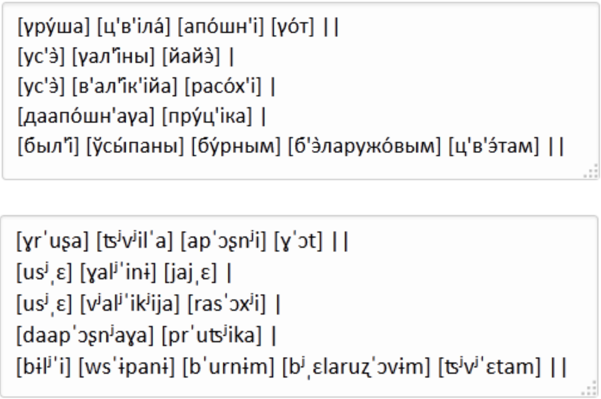
Step 7. *Mailing of notifications to developers.* The above-mentioned data are attached to an e-mail letter which is sent to developers – an expert linguist and a software engineer – for determining the type of a problematic situation (linguistic mistakes, errors of algorithms and program codes), and the best way to solve it.

Step 8. *The end of work of the algorithm.*

As a result of the algorithm described above user receives the transcribed text. For example, a user has submitted to the input the following text:

Гру+ша цвіла+ апо+шні го+д. Усе= галі+ны яе=, усе= вялі+кія расо+хі, даЪапо+шняга пру+ціка, былі+ ўсы+паны бу+рным бе=ла-ружо+вым цве+там.

An example of the processing of the following text is shown in the figure 2.



[γрýша] [ц'в'ілá] [апóшн'і] [γóт] ||
[ус'э̀] [γал'̆іны] [йайэ̀] |
[ус'э̀] [в'ал'̆ік'ійа] [расóх'і] |
[даапóшн'аγа] [прýц'іка] |
[был'̆і] [ўсы́паны] [бýрным] [б'э̀ларужóвым] [ц'в'э́там] ||

[ɣrˈuʂa] [ʦʲvʲilˈa] [apˈɔʂnʲi] [ɣˈɔt] ||
[usʲˌɛ] [ɣalʲˈinɨ] [jajˌɛ] |
[usʲˌɛ] [vʲalʲˈikʲija] [rasˈɔxʲi] |
[daapˈɔʂnʲaɣa] [prˈuʦʲika] |
[bɨlʲˈi] [wsˈɨpanɨ] [bˈurnɨm] [bʲˌɛlaruʐˈɔvɨm] [ʦʲvʲˈɛtam] ||

Figure 2: The example of the resulting text which has been transcribed according to the Cyrillic alphabet, and IPA

3. Description and the working algorithm for generator of the orthoepic dictionary

Developers noticed that, if it is possible to get a transcription of the word/text, then it is possible to create a resource, which will give not only the transcription of the submitted text, but combine it with the input text. The whole creation of the "Orthoepic Dictionary Generator" begins with this idea.

To get a transcription after each word one needs to enter a certain part of an orthographic dictionary which is taken as a basis into a special field and press the button "Get text with transcriptions! /Атрымаць тэкст з транскрыпцыямі!". But, as one may know, apart from the registered word, quite frequently there are certain forms of the word in the dictionary, as well as its grammar or stylistic labels, so, after them we can get unnecessary transcription. In order to avoid it, a special field for stop-words was created, and the transcription will not generate after the stop words. As an output of the text processing the Cyrillic transcription which illustrates correct pronunciation is given after a title word and after each of its forms. Let us see how the system works. For instance, user inserted the following fragment of the orthographic dictionary:

сакаляня́ і **сакалянё** *н.* *НВ* сакаляня́ (-нё), *РДМ* сакаляня́ці, сакалянём; *мн.* сакаляня́ты, *РВ* сакаляня́т, сакаляня́там, сакаляня́тамі, сакаляня́тах

After the processing of the text the system displays the following resulting line:

сакаляня́ [сакал'ан'а́] і **сакалянё** [сакал'ан'о́] *н.* *НВ* сакаляня́ [сакал'ан'а́] (-нё), *РДМ* сакаляня́ці [сакал'ан'а́ц'і], сакалянём [сакал'ан'о́м]; *мн.* сакаляня́ты [сакал'ан'а́ты], *РВ* сакаляня́т [сакал'ан'а́т], сакаляня́там [сакал'ан'а́там], сакаляня́тамі [сакал'ан'а́там'і], сакаляня́тах [сакал'ан'а́тах]

As can be seen from the given example, the system generates the transcription after each necessary word form, bypassing the various labels, maintaining the desired format for the convenience of work, and determines the position of the accent in unambiguous situations, where for that reason they are not specified in the orthographic dictionary.

Let us take the closer look to the architecture of the system.

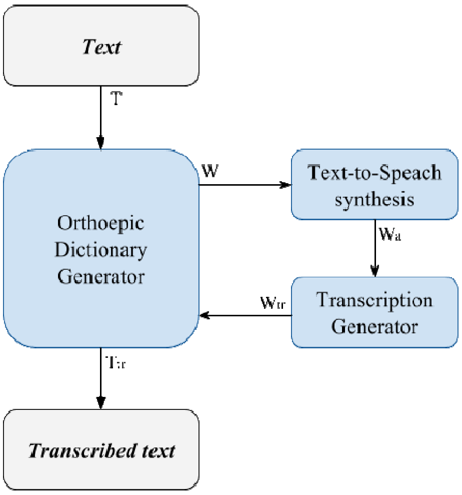


Figure 3: The general scheme of the generator of the orthoepic dictionary work

To the input of the algorithm a text T is submitted. The algorithm performs the following steps:

Step 1. *Tokenization*. The input text is divided into paragraphs, in each all the words W_i are extracted. It should be noted that the extraction of the words does not happen by breaking the line with gaps, but by the search for strings that match a pattern of possible appearance of written words in the Belarusian language. For all the extracted words (except for the stop-words) steps 2 and 3 are sequentially performed.

Step 2. *Conversion of the orthographic text into allophones.* Word is served to the speech synthesizer, where it is converted into the allophone form.

Step 3. *Generation of the transcriptions*. The word in the allophone form W_a is transmitted to the transcription generation system and is converted into a transcribed form W_{tr}. Transcription is passed back to the automated generator of the orthoepic dictionary.

Step 4. *Collection of the resulting text*. Resulting transcriptions W_{tr} are combined into one text, where after the wordforms the transcription is inserted in square brackets.

$$W_{tr} = U\{W_i, W_{tri}\} \quad (1)$$

Labels of grammatical categories remain unchanged.

Step 5. *Preparation for data saving*.

Step 6. *Mailing of notifications to developers.*

Step 7. *The end of work of the algorithm.*

Thus, created service for the generation of the orthoepic dictionary is intended to represent the original data in the transcribed form, which makes the work of linguists on creation orthoepic dictionary of the Belarusian language much easier. This automatic service can be made for the other languages, including Ukrainian.

4. Conclusions

This article describes the beginning of the work on the development of a system for generating different types of the transcription by an input orthographic text in the Belarusian language, as well as computer-aided system for generation of the orthoepic dictionary of the Belarusian language. Experimental prototype of these systems is implemented as free service which is always available online services at the www.corpus.by/transcriptionGenerator/ and www.corpus.by/orthoepicDictionaryGenerator/ for the resource www.corpus.by [7, 8, and 9]. With the help of the transcription generation system it is possible to automatically generate a transcription of any Belarusian orthographic line in three types.

5. Acknowledgements

We would like to thank Julia Borodina for her help in revising the language of this paper.

6. References

[1] Гецэвіч, Ю.С. Стварэнне сэрвіса арфаэпічнага генератара слоўнікаў / Гецэвіч Ю.С., Гюнтар А.В., Лысы С.І., Русак В.П., Мандзік В.А. // Тези доповідей міжнародної конференції «Діалекти в синхронії та діахронії : загальнослов'янський контекст» (Київ, 2–4 квітня 2014 року) / За ред. П.Ю. Гриценка . Ін-т укр. мови НАН України . Київ : КММ, 2014 . – С. 101–106.

[2] Чахоўскі Г.К., Чахоўская Т.Л. Сучасная беларуская мова. Фанетыка. Фаналогія. Арфаэпія. Мінск: БДУ, філалагічны факультэт, кафедра сучаснай беларускай мовы, 2010. - 110 с.

[3] Гачко, А.К. Метадычныя рэкамендацыі па раздзелу "Фанетыка. Фаналогія" курса "Сучасная беларуская мова": фанетычная транскрыпцыя. [Электронны рэсурс]. – 2012. – Рэжым доступу : http://edu.grsu.by/books/gachko _phonetic/index.php/fanetychnaya-transkryptsyya. – Дата доступу : 09.12.2013.

[4] Беларуска-англійскі размоўнік / уклад. У.А. Кошчанка. – Мн. : Артыя Груп, 2010. – 190 с.

[5] The International Phonetic Associacion [Electronic resource]. – 2005. – Mode of access : http://www.langsci.ucl.ac.uk/ipa/index.html/ – Date of access : 08.12.2013.

[6] Hetsevich, Yu. The system of generation of phonetic transcriptions for input electronic texts in belarusian / Yu. Hetsevich, V. Mandzik, V. Rusak, A. Hiuntar, T. Okrut, B. Lobanov, S. Lysy, Dz. Dzenisiuk // Pattern Recognition and Information Processing : Proceedings Of The 12th International Conference (28-30 May, Minsk, Belarus) — Minsk : UIIP NASB, 2014. — C.81-85.

[7] Transcription Generator [Electronic resource]. – 2014. – Mode of access : http://www.corpus.by/transcriptionGenerator/. – Date of access : 09.10.2014.

[8] Orthoepic Dictionary Generator/ [Electronic resource]. – 2013. – Mode of access : http://www.corpus.by/orthoepicDictionaryGenerator/. – Date of access : 09.10.2014.

[9] Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – 2013. – Mode of access : http://www.corpus.by/tts3/. – Date of access : 09.10.2014.