Development of multi-voice and multi-language TTS synthesizer (languages: Belarussian, Polish, Russian)

Boris Lobanov and Liliya Tsirulnik

United Institute of Informatics Problems of National Academy of Science
Minsk, Belarus
lobanov@newman.bas-net.by, liliya_tsirulnik@ssrlab.com

Abstract

The paper describes some results of the research which aiming at filling the gap in introducing and promoting computerized speech technology for Slavonic languages, in particular, a technology of TTS synthesis for Belarusian, Polish and Russian. A typological analysis of the peculiarities of phonemic and allophonic systems of Belarussian, Polish and Russian languages is given. Based on the results of this study, an approach to making a unified phonetic-acoustical database for multi-language Slavonic TTS synthesis is proposed. The results of the quantitative analysis of the pitch contours for some Slavonic languages and, besides, the peculiarities of speakers individual intonation are presented. The general structure of multi-language and multi-voice TTS system is described.

1. Introduction

Slavonic languages and speech systems, in particular, those of Belarusian, Polish and Russian, have very much in common. This is true of their phonetic, lexical, morphological and syntactic structure. This fact enables the researchers to set as an objective the creation of an integrated algorithm of multi-language TTS conversion system common for all these languages. One may expect that the main features of such a system will be also applicable to other Slavonic languages, such as Ukrainian, Czech, Slovak, Serbo-Croatian, Slovenian, Bulgarian and Macedonian. At present, only a few TTS systems for Slavonic speech generation are available. However, the quality of the synthesized speech is still far from natural, and the number of synthetic voices is very restricted. Belarussian TTS systems do not exist at all.

The TTS conversion system under discussion has a common structure for all Slavonic languages concerned but it uses different linguistic and acoustical resources for each language. The objective is the development of a high-quality multi-lingual and multi-voice TTS-system on a common platform. This objective is achieved by the evolution of original algorithms of multi-language and multi-voice TTS synthesis, which were developed earlier [1-4]. The speech signal synthesis is based on Allophone and Multi-Allophone Natural Waves (ANW and MANW) method of speech signal concatenation (see section 2). The speech prosody synthesis is based on Accentual Units Portrait (AUP) method of stylization entire tonal, rhythmical and dynamic contours of a phrase and an utterance as a whole (see section 3). In order to synthesize prosodic features the system will also resort to a deep morphological and syntactical analysis of a sentence [5]. The two modules operating jointly are expected to achieve a high quality of synthesized speech. The quality of TTS synthesis largely depends on how close the model of human voice and pronunciation can be made. The voice "cloning" technology ensures a high quality of speech imitation for a specific individual by means of TTS synthesis [6,7]. The cloning procedure is based on two types of text-corpus and corresponding to them audio-data from a speaker: a) for data-driven 'cloning' of an individual voice and phonemic peculiarities, b) for data-driven 'cloning' of individual features of the prosodic organization of speech.

The paper consists of three parts:

- Study and modeling of language and speaker specific *phonemic* peculiarities (section 2);
- Study and modeling of language and speaker specific *prosodic* peculiarities (section 3);
- General description of the Slavonic *multi-language* and *multi-voice* TTS system (section 4).

2. Study and modeling of language and speaker specific *phonemic* peculiarities

2.1. Belarussian, Polish and Russian phonetic systems

Phonetic systems of the Slavonic languages group have much in common among themselves, however each of them possesses also specific features, sometimes significant. Investigated phonetic systems of the Belarussian, Polish and Russian languages are rather close, especially Russian and Belarussian. In the Belarussian language there are 41 phonemes, of which there are 6 vowels and 35 consonants, and in Russian of the whole number – 42 – there are 6 vowels and 36 consonants. The phonemic system of the Polish is more varied. In it there are 51 phonemes: 8 vowels and 43 consonants. Table 1 presents generalized information on the phonemic systems of three languages and about the distinctions of manner and place of articulation in them. In each cell of the table there are names of the phonemes, described by the certain manner and a place of articulation, for the Belarussian, Polish and Russian languages in the order "of top to bottom". For the designation of phonemes letters of the alphabet traditional for each language are used.

In table 1 the cells are blacked out, when the phonetic quality of sounds is practically identical for each of the languages. As it is apparent from the table, the number of such cells in percent relation to all the cells used is rather considerable – 66 %.

The distinctive features of the phonetic systems of Belarussian and Russian consist in the fact, that some of the

Manner of articulation / Place of articulation		Consonants										Vowels	Front(1) / Back(0)	High(1) / Low(0)	Labialized	Nasalized
		Unvoiced			Voiced			Sonant								
		Plosive	Affricate	Fricative	Plosive	Affricate	Fricative	Vibrant	Nasal	Lateral	Liquid					
Velar	Soft	к' k' к'	~	х' h' х'	~ g' г'	~	гх' ~ ~	~	~	~	ў j ў	у у у	0	1	1	0
	Hard	к k к	~	х h х	~ g г	~	гх ~ ~	~	~	~	~	о о о	0	0	1	0
Dorsal	Soft	~	~ ć ч'	~ ś ш'	~	~ dź ~	~ ź ~	~ r' p'	~	~	~	а а а	0	0	0	0
	Hard	~	ч cz ~	ш sz ш	~	дж dż	ж ż ж	р r р	~	~	~	э е э	1	0	0	0
Pre-dental	Soft	~ t' m'	ц' c' ~	с' s' с'	~ d' д'	дз' ~ ~	з' z' з'	~	н' n' н'	л' l' л'	~	ы у ы	0	1	0	0
	Hard	m t m	ц c ц	с s с	д d д	~ dz ~	з z з	~	н n н	л l л	~	і і и	1	1	0	0
Labial	Soft	п' р' п'	~	ф' f' ф'	б' b' б'	~	в' w' в'	~	м' m' м'	~	~	~ ą ~	0	0	1	1
	Hard	п р п	~	ф f ф	б b б	~	в w в	~	м m м	~	ў ł ~	~ ę ~	1	0	0	1

Table 1. The phonetic systems of Belarussian, Polish and Russian languages

phonemes found in Russian are missing in Belarussian, namely:

- Soft consonants ***Т', Д', Ш', Ч', Р';***
- Hard ***Г*** and soft ***Г'*** consonants.

On the other side, there are a number of specific consonants in the Belarussian, that are missing in Russian:

- Liquid ***Ў***;
- Hard ***Ч*** and soft ***Ц'***;
- Hard ***Дж*** and soft ***Дз'***;
- Hard ***Гх*** and soft ***Гх'.***

Comparing the phonetic system of the Polish language with Russian, we shall also note some features of similarity and difference. In the Polish language there are all phonemes, characteristic of Russian, however, the pronunciation of soft phonemes ***Ш'*** and ***Ч'*** differs from the Polish soft ***Ś*** and ***Ć,*** the place of articulation of which intermediate between the soft Russian *С'*, *Ш'* and *Ц'*, *Ч'*, accordingly. Besides, in the Polish language there are a number of specific phonemes, which are absent in Russian:

- Liquid ***Ł;***
- Soft ***С', Ć*** and hard - ***Cz***;
- Soft affricate ***Dź,*** and hard ***Dż*** and ***Dz***;
- Nasalized vowels - ***Ą*** and - ***Ę***.

If we compare the phonemic systems of all the languages being considered, and also each pairs of languages, counting up the quantity of occurrence in the cells of table 1 we will receive the following values in percentage to the total of cells used by them:

- Russian – Belarussian – Polish - 66 %;
- Russian – Belarussian - 71 %;
- Russian – Polish - 78 %;
- Polish – Belarussian - 69 %.

However surprising it may seem, but the Belarussian language in its phonetic structure differs almost equally both from Polish, and from Russian. This, certainly, does not concern the statistics of the occurrence of these or those

phonemes in various languages. Thus, as is, similar in sound the Russian and Polish phonemes ***/t ’/, /d ’/, /s’/, /z ’/, /l/*** are used in Russian very frequently, while in Polish they occur much less frequently. In Polish words, similar in their phonetic form in Polish and in Russian, the specific Polish phonemes - ***/ć/, /dź/, /ś/, /ź/, /ł/*** are used, accordingly.

2.2. Mini- and maxi-sets of allophones for TTS synthesis of the Belarus, Polish and Russian

As it is known, in the speech flow phonemes are realized in the form of allophones, or otherwise, in the form of positional and combinatory variants of phonemes. The positional factor considers the position of the given phoneme in relation to the stressed syllable of a word, accentual unit or phrase. The combinatory factor considers the nearest phoneme environment. Generally, it is impossible to give an exact estimation of the quantity of allophones, since it directly depends on the degree of detail in the account of influence of positional and combinatory factors. However, the quality of synthesized speech directly depends on a degree of detail and elaboration. A demand for greater elaboration may lead to a huge quantity of allophones (hundreds of thousands), that makes the problem of creating a DB of allophones insoluble. Experience of creating Russian-speaking TTS [8] has shown, that synthesized speech of a high enough quality can be reached under certain conditions of generating the positional and combinatory allophones. Two types allophone sets have been investigated: the so-called maxi- and mini-sets.

For using a maxi-set allophone base for the synthesis of Russian speech the following positional allophones of *vowels* are created: fully stressed - (0), partially stressed - (1), the first pre-stressed - (2), not the first pre-stressed - (3), post-stressed - (4). In all there are 5 positions. With regard to the left context the following combinatory allophones of vowels are created: after a phrase pause - (0), after the most of the labial - (1), pre-dental and dorsal - (2), velar - (3) hard consonants, after /L / - (4), /R/ - (5), /M/ - (6)/, /N/ - (7), after the most of the soft consonants - (8), after /P’/ - (9), /M’/ - (10), /H’/ - (11), after the vowels /U/ - (12), /O/ - (13), /A/ - (14), /E/ - (15), /Y/ - (16), /I/ - (17). In all there are 18 left contexts. Considering the right context, the following combinatory allophones of vowels are created: before a phrase pause - (0), before labial hard - (1), before pre-dental, dorsal and velar hard consonants - (2), before labial soft - (3) before non-labial soft consonants and the vowel /I/ - (4). In all there are 5 right contexts. Overall, for the 6 vowels we avail of Nv = 5*18*5*6 = 2700 allophones.

Positional allophones of *consonants* for a maxi-set include two positions: in an accented syllable – (0) and in an unaccented syllable – (1). The left context of the consonants includes the following groups: after a pause - (0), after unvoiced - (1) and voiced - (2) consonants, and after vowels - (3). The right context includes the positions: before a pause - (0), before unvoiced - (1) and voiced - (2) consonants, before unstressed - (3) and stressed - (4) vowels. Overall, for 36 consonants we avail of Nc = 2*4*5*36 = 440 allophones. In total for all Russian phonemes we use 2700 + 1440 = 4140 allophones.

For a mini-set of allophones for Russian speech synthesis only 2 types of positional allophones of *vowels* are created: stressed - (0), unstressed - (1). In view of the left context the following combinatory allophones of vowels are created: after a phrase pause - (0), labial - (1), pre-dental - (2), velar - (3) hard consonants, and after soft consonants - (4). In all there are 5 left contexts. In view of the right context there are the following combinatory allophones of vowels: before a phrase pause - (0), before labial - (1), pre-dental, dorsal and velar - (2) hard consonants, and before soft consonants - (3). In total for 6 vowels we avail of Nv = 2*5*4*6 = 240 allophones. Allophones of *consonants* are created only with regard to the right context: before a pause - (0), before unvoiced - (1) and voiced - (2) consonants, before unstressed- (3) and stressed - (4) vowels. In total, for all the 36 consonants we have Nc = 5*36 = 180 allophones. Overall, for the Russian phonemes in a mini-set we have 240+180 = 420 allophones. Similar observations for the other two languages – Belarussian and Polish – can be made.

At the beginning the mini-sets of ANW for TTS synthesis of each language are created manually. At the next step the mini-sets of ANW for the automatic creation of the maxi-sets of ANW and the sets of Multi-ANW (sequences of two and more ANWs – MANW) are utilized. Automatic creation of maxi-sets of ANW and MANW DB is realized by data driven voice “cloning” technology [9].

The received estimations of allophone quantity, calculated theoretically, are strongly overestimated in that, firstly, many positional and combinatory situations do not occur in speech altogether and, secondly, for many allophones acoustic distinctions are so insignificant, that they can be neglected. As a result, as experience shows, the quantity of allophones used in a maxi-set appears to be more than 2 times, and in a mini-set – 1,5 times smaller. The results of the calculation of the theoretical set and the one, used practically, the quantity of allophones for each of the three languages result as is shown in table 2. For the designation of allophones the symbols of corresponding phonemes (in Latin letters) with 3 digital indexes are used, where the first index designates the positional type of a phoneme, the second index – the type of the left context, and the third index – the right context. In table 3 uniform designations of allophones, used for speech synthesis in three Slavonic languages are presented.

Language	**Belarussian**				**Polish**				**Russian**			
Quantity of allophones	*Theoretical*		*Used in practice*		*Theoretical*		*Used in practice*		*Theoretical*		*Used in practice*	
Type of the set	*Maxi*	*Mini*	*Maxi*	*Mini*	*Maxi*	*Mini*	*Maxi*	*Mini*	*Maxi*	*Mini*	*Maxi*	*Mini*
Vowels	2700	240	1480	170	3600	320	2050	224	2700	240	1550	175
Consonants	1400	180	720	76	2040	215	920	113	1440	180	840	81
Total	4100	420	2200	246	5640	535	2970	337	4140	420	2390	256

Table 2. Allophones number of different types in Belarussian, Polish, and Russian languages

	Labial consonants					*Pre-dental consonants*					*Dorsal consonants*					*Velar consonants and vowels*			
№	**Bel**	**Pol**	**Rus**	**Name**	**№**	**Bel**	**Pol**	**Rus**	**Name**	**№**	**Bel**	**Pol**	**Rus**	**Name**	**№**	**Bel**	**Pol**	**Rus**	**Name**
1	*п*	*p*	*п*	P_{ijk}	16	*м*	*t*	*м*	T_{ijk}	31	*ч*	*cz*	-	Ch_{ijk}	46	*к*	*k*	*к*	K_{ijk}
2	*ф*	*f*	*ф*	F_{ijk}	17	*ц*	*c*	*ц*	C_{ijk}	32	*ш*	*sz*	*ш*	Sh_{ijk}	47	*х*	*h*	*х*	H_{ijk}
3	*б*	*b*	*б*	B_{ijk}	18	*с*	*s*	*с*	S_{ijk}	33	*дж*	*dż*	-	Dh_{ijk}	48	*гх*	*g*	*г*	G_{ijk}
4	*в*	*w*	*в*	V_{ijk}	19	*д*	*d*	*д*	D_{ijk}	34	*ж*	*ż*	*ж*	Zh_{ijk}	49	*к'*	*k'*	*к'*	K'_{ijk}
5	*м*	*m*	*м*	M_{ijk}	20	-	*dz*	-	Dz_{ijk}	35	*р*	*r*	*р*	R_{ijk}	50	*х'*	*h'*	*х'*	H'_{ijk}
6	*ў*	*ł*	-	W_{ijk}	21	*з*	*z*	*з*	Z_{ijk}	36	-	*ć*	*ч'*	Ch'_{ijk}	51	*гх'*	*g'*	*г'*	G'_{ijk}
7	*п'*	*p'*	*п'*	P'_{ijk}	22	*н*	*n*	*н*	N_{ijk}	37	-	*ś*	*ш'*	Sh'_{ijk}	52	*й*	*j*	*й*	J'_{ijk}
8	*ф'*	*f'*	*ф'*	F'_{ijk}	23	*л*	*l*	*л*	L_{ijk}	38	-	*dź*	-	Dh'_{ijk}	53	*у*	*u*	*у*	U_{ijk}
9	*б'*	*b'*	*б'*	B'_{ijk}	24	-	*t'*	*м'*	T'_{ijk}	39	-	*ź*	-	Zh'_{ijk}	54	*о*	*o*	*о*	O_{ijk}
10	*в'*	*w'*	*в'*	V'_{ijk}	25	*ц'*	*c'*	-	C'_{ijk}	40	-	*r'*	*р'*	R'_{ijk}	55	*а*	*a*	*а*	A_{ijk}
11	*м'*	*m'*	*м'*	M'_{ijk}	26	*с'*	*s'*	*с'*	S'_{ijk}	41	-	-	-	-	56	*э*	*e*	*э*	E_{ijk}
12	-	-	-	-	27	*дз'*	*d'*	*д'*	D'_{ijk}	42	-	-	-	-	57	*ы*	*y*	*ы*	Y_{ijk}
13	-	-	-	-	28	*з'*	*z'*	*з'*	Z'_{ijk}	43	-	-	-	-	58	*і*	*i*	*и*	I_{ijk}
14	-	-	-	-	29	*н'*	*n'*	*н'*	N'_{ijk}	44	-	-	-	-	59	-	*ą*	-	O'_{ijk}
15	-	-	-	-	30	*л'*	*l'*	*л'*	L'_{ijk}	45	-	-	-	-	60	-	*ę*	-	E'_{ijk}

Table 3.Enumeration of allophone names, used for Belarussian, Polish, and Russian speech synthesis

2.3. Creation of language and speaker specific DB of ANWs and MANWs

The process of creation of the language and speaker specific DB of ANWs and MANWs includes the following operations:

- Formation of the representative text corpuses and speech recordings corresponding to these texts (speech base) from different speakers;
- Processing of the created speech base including phonemic segmentation of the speech signal, allophonic marking of segments and preservation of the obtained set in a ANWs DB.

At the first stage text corpuses are created on the basis of a specially selected mini-set of words in the quantity equal to the minimal number of allophones in each of the languages being used. Speech recordings, corresponding to text corpuses, are produced in studio conditions by specially instructed professional announcers. Below, in table 4 fragments of the lists of words for the creation of the mini-sets of consonants ANWs, and in table 5 - the mini-sets of vowels ANWs for the 3 languages are shown.

It is obvious, that though the use of a maxi-set of allophones for synthesis will provide a higher quality of speech, its "manual" creation is almost impossible (the order of 4000 allophones for each of the languages and speakers!) if not possible at all. Creating a mini-set (the order of 300 allophones) "manually" is quite real. A mini-set, in the same way as a maxi-set, provides speech synthesis from any text though the quality of the synthesized speech will be not as high.

Allophone's index (right context) / Language	Phrase pause (0)	Unvoiced consonants (1)	Voiced consonants (2)	Unstressed vowels (3)	Stressed vowels (4)
Belarussian	***Цяжар***	***Дзірка***	***Скарба***	***Сябраваць***	***Урад***
Polish	***Akr***	***Krtań***	***Grdyka***	***Środowisko***	***Program***
Russian	***Спор***	***Марка***	***Кордон***	***Караван***	***Парад***

Table 4. Fragments of the lists of words for the creation of the mini-sets of consonants ANWs (for /R/ consonant)

Second index (left context0) **Languages:** *Belarussian, Polish, Russian* / Third index (right context) **Languages:** *Belarussian, Polish, Russian*		**0**		**1**		**2**		**3**	
		phrase pause		*п, ф, б, в, м, ў*		*т, ц, с, д, з, н, л, ч, ш, дж, ж, р, к, х, г, у, о, а, э, ы*		*к', х', г', й, ц', с', дз', з', н', л', п', ф', б', в', м', і*	
		phrase pause		*p, f, b, w, m,ł*		*t, c, s, d, dz, z, n, l, cz, sz, dż, ż, r, k, h, g, u, o, ą, a, e, ę, y*		*k', h', g', j, ć, ś, dź, ź, r', t', c', s', d', z', n', l', p', f', b', w', m', i*	
		phrase pause		*п, ф, б, в, м*		*т, ц, с, д, з, н, л, ш, ж, р, к, х, г, у, о, а, э, ы*		*к', х', г', й, ч', ш', р', т', с', д', з', н', л', п', ф', б', в', м', и*	
0	*phrase pause*	A000	*А*	A001	Аўра	A002	Анджей	A003	Альфа
	phrase pause		А		Amper		Adres		Ani
	phrase pause		А		Автор		Атом		Ася
1	*п, ф, б, в, м, ў*	A010	Барацьба	A011	Вабны	A012	Фарба	A013	Майстар
	p, f, b, w, m, ł		Ва		Zaspawać		Najbardziej		Pawie
	п, ф, б, в, м		Судьба		Баба		Вата		Батя
2	*т, ц, с, д, з, н, л, ч, ш, дж, ж, р*	A020	Кабала	A021	Зграбны	A022	Цацка	A023	Талент
	t, c, s, d, dz, z, n, l, cz, sz, dż, ż, r		Та		Samym		Znacznie		Zaletami
	ш, ж, р, т, ц, с, д, з, н, л		Еда		Запад		Дата		Тася
3	*к, х, г, у, о, а, э, ы*	A030	Дачка	A031	Кава	A032	Казка	A033	Камень
	k, h, g, u, o, ą, a, e, ę, y		На		Gapa		Bogaty		Zagapić
	к, х, г, у, о, а, э, ы		Нога		Гавкать		Сказка		Галя
4	*ц', с', дз', з', н', л', п', ф', б', в', м', к', х', г', й, і*	A040	Мітусня	A041	Сябар	A042	Немаўляты	A043	Сядзеш
	t', c', s', d', z', n', l', ć, ś, dź, ź, r', p', f', b', w', m', k', h', g', j, i		Pnia		Rozdziawa		Posiada		Kopiami
	т', с', д', з', н', л', ч', ш', р', п', ф', б', в', м', к', х', г', й, и		Шутя		Тяпка		Тяга		Тянет

Table 5. *Fragments of the lists of words for the creation of the mini-sets of vowels ANWs (for stressed vowel /A/)*

However, due to the creation of a mini-set of allophones it becomes possible to start an automatic process of "cutting" a maxi-set of ANWs, as well as that of larger units – multi-allophones (a set of MANWs), realized in the form of an allophone sequence – diallophones, allosyllables and other speech units. For automating the process of creating a DB of ANWs and MANWs the technology of a personal voice cloning is used.

The general scheme of the procedure of creating a mini- and maxi-DB of ANWs and DB of MANWs is presented in fig.1.

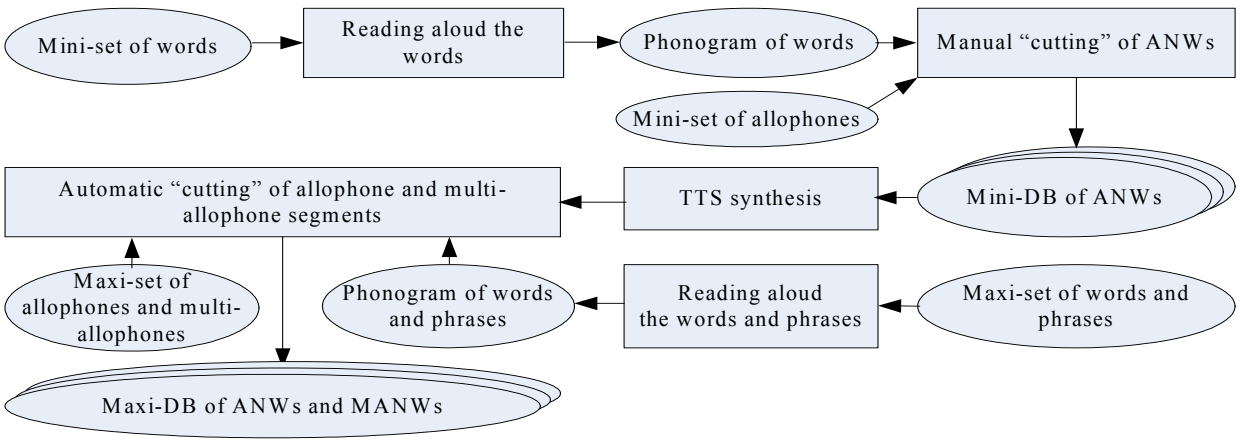


Figure 1. The general scheme of the procedure of creating a mini- and maxi-DB of ANWs and MANWs

3. Study and modeling of language and speaker specific *prosodic* peculiarities

3.1. Fundamentals of AUP intonation stylization model utilized in TTS

A large variety of models have been applied in speech synthesis systems to specify prosodic parameters, including phonological models that represent the prosody of an utterance as a tone-sequence [10], acoustic-phonetic superpositional models that interpret F0 contours as complex patterns resulting from the superposition of several components [11], IPO model that represent intonation as an inventory of pitch movements [12], and Tilt model that utilizes the continuous parameterization of F0 contours [13]. A very useful perceptual description of Russian intonation according to the IPO model was developed by C. Ode [14]. All of the approaches rely on a combination of data-driven and rule-based methods. They explore natural speech databases, and vary in terms of what is derived from the analysis to drive intonation synthesis.

Most of the intonation models, mentioned above, were developed and tested for English, French, German, Dutch, and some others languages. But there are only a few examples of the development and utilization of these models for Slavonic languages. The main principle of synthesizing prosodic parameters that we have utilized here is based on an original model which actually resembles the above mentioned ones yet differs from them in the underlying method of phrase intonation representation, namely, by a sequence of Accentual Unit Portraits (AUP-stylization model). It was proposed over ten years ago [1] and has been used successfully since then in several TTS synthesis models. This section is concerned with the study of AUPs finality/non-finality (or completeness/incompleteness in compliance with other terminology) phrase intonation types, namely - its language-specific peculiarities, and with the implementation of these "portraits" in the unified text-to-speech synthesis system for Slavonic languages.

In accordance with the AUP stylization model, the minimal prosodic unit is the Accentual Unit (AU), consisting of one or more words, having only one fully stressed syllable. An AU, in its turn, consists of the nucleus (the fully stressed syllable), the pre-nuclear part (all the phonemes preceding the fully stressed syllable) and the post-nuclear parts (all the phonemes following the fully stressed syllable).

The main assumption of AUP stylization is, that the topological properties of prosodic parameters do not change (or change insignificantly) with the changes of the phonemic context and number of syllables in the pre- and post-nucleus for a certain type of phrase intonation. This fact can be clearly seen from figure 2, where F_0 contours for various one word-phrases with a different accent position are shown. These phrases were recorded by the speaker who pronounced the words with the interrogative type of intonation.

An AU may consist of more than one word but only in a case when the phrase has only one accented (prominent) word. This is illustrated in figure 3, where F_0 contours of a three-words phrases with a different position of the focused word in a phrase are shown. The phrase "Мама мыла малину?" (the English translation "Did mother wash raspberry?") was recorded three times by the speaker who pronounced it with the interrogative type of intonation, and with three different positions of the focus.

As is clear from fig.3, each of these phrases consists of only one AU, and the behaviour of F_0 contour is rather similar to that of the nucleus, pre- and post-nucleus of a single word shown in fig.2.

All mentioned above gives us good reasons to represent the AUP of F_0 contour in a time-frequency space with a the relative equal duration of the three AU's parts - nucleus, pre- and post-nucleus.

In fig. 4a the common AUP of F_0 contours for the interrogative type of intonation that corresponds to one-word phrases from fig. 2 is shown, and in fig.4b – the contour corresponding to three-word phrases from fig. 3. As seen from figures 4a and 4b, the difference between AUPs is not very significant.

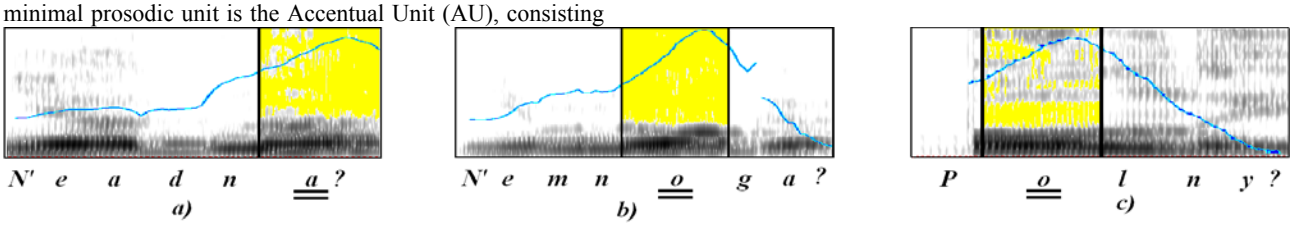


Figure 2: F_0 contours of interrogation for the Russian word-phrases: a) "Не одна?"/N'eadn`a/-"Not one?", b) "Не много?" /N'emn`oga/-"Not much?", c) "Полный?"/P`olny/-"Full up?" (the accented vowels are underlined with a double line)

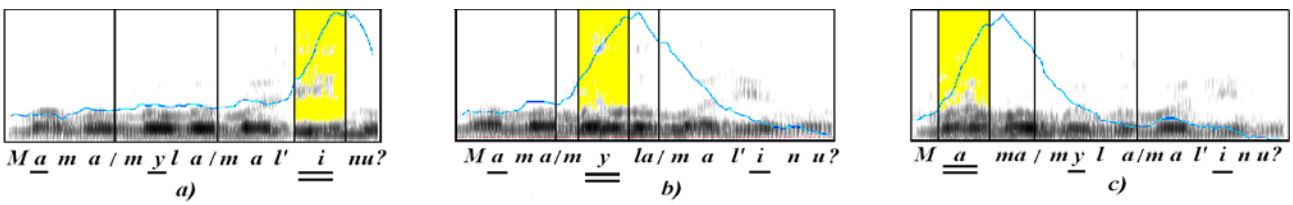
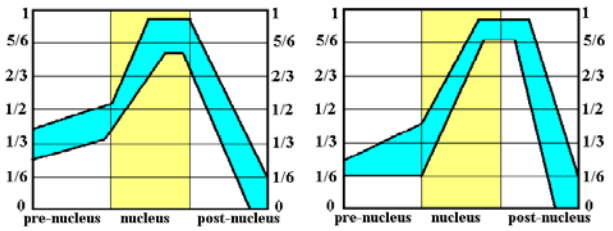


Figure 3: F0 contours of interrogation for the Russian phrase "Мама мыла малину?" /Mama myla mal'inu/ with the focused words: a)"mal'inu", b) "myla", c) "mama"(the strong accented vowels are underlined with a double line)



Figire 4: AUP for interrogative type of intonation for one-word phrases (on the left) and three-word phrases (on the right)

The main principles of AUs pitch contour "portraits" creation are illustrated in Fig.5 by an example of a Russian phrase: *"которые могут быть представлены"*, in transcription*:* "*kat**o**rye m**o**gut byt' pr'etst**a**vl'eny*" (the fully stressed vowels are underlined); the English translation is "*that can be represented*". It is part of an utterance, spoken by a male speaker and carrying a non-final intonation type contour consisting of 3 AUs.

First, the F_0 values are computed for every vocalized segment (Fig. 5 a). Then, the AUs boundaries as well as pre-nucleus, nucleus, and post-nucleus areas for each AU are marked and F_0 values for voiceless segments are interpolated (Fig. 5 b). Finally, the AU's pitch and duration are normalized (Fig. 5 c).

For F_0 normalization the minimum F_0 value (F_{0min}) and the maximum F_0 value (F_{0max}) are determined from the full phonogram being analyzed. Generally, F_{0max} is located on the AU nucleus of an exclamatory phrase, while F_{0min} is associated with the AUs nucleus of a final phrase in an utterance located at the end of a paragraph. For F_0 value normalization (F_{0norm}) the following formula is used:

$$F_{0\ norm} = (F_0 - F_{0\ min}) / (F_{0\ max} - F_{0\ min}) \qquad (1)$$

For the given speaker the F_{0min} value was equal to 70 Hz and F_{0max} – 180 Hz (see Fig.5 a). F_0 values can also be represented in Log or ERB-scales. The AUs duration normalization is carried out through equalization of pre-nuclear, nuclear, and post-nuclear parts (see Fig.5 c).

Thus, we obtain a set of normalized "portraits" of pitch contours for different types of phrase intonation. These normalized sequences of AUPs are utilized then by TTS synthesis system independently of particular AUs' phonemic contents. Speech re-synthesis by using AUPs thus obtained does not noticeably diminish the perceived intonation quality.

3.2. Comparative study of language and speaker specific peculiarities in intonation contours

The aim of the study is the description of language-and speaker-specific peculiarities of phrase intonation according to AUPs stylization model, namely of final/non-final intonation types. The experimental material for the study of language- and speaker-specific intonation cues was provided by a specially selected representative text spoken by several speakers. The text was sorted out so as to represent each of the intonation types considered above.

In the first part of the experiment aimed at studying language-specific distinctions in phrase intonation, Russian and Polish native female speakers were asked to read out corresponding texts of a similar scientific content in both languages.

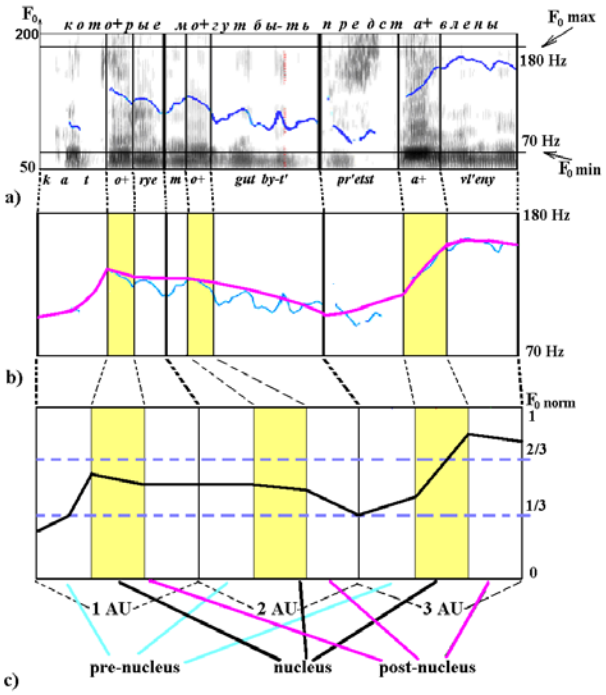


Figure 5: Scheme of pitch contours "portrait" creation: a) F_0 values computation, b) F_0 curve interpolation, c) F_0 curve normalization

The texts in both languages comprised more than one thousand words and approximately 300 intonation phrases. Both texts were spoken two times by the speakers at normal speed. The two recordings were aurally tested and the better one was used for further analysis.

In the second part of the experiment devoted to the study of speaker-specific distinctions in Russian phrase intonation, we used a phonetically balanced Russian text corpus designed at the experimental phonetics department of St.-Petersburg University [15]. The text includes about one thousand words and 250 intonation phrases. The text was spoken two times by two professional Russian male speakers. The two recordings were aurally tested and the better one was used for further analysis.

The recorded speech corpus was then processed by experienced phoneticians with the help of the Praat speech processing software.

The audio files obtained during the recording and their transcript served as the database for the research. Initially the speech material was analyzed aurally and irrelevant segments, such as noises, sighs and eh-'fillers', were removed. Then an expert analyzed the audio recording into phrases. The decision about the end of a phrase was drawn from various features, such as a breath-pause, a pitch change of a phrase (F_0 contour), a specific dynamic structure (amplitude envelope) and a particular rhythmic pattern (sound duration pattern). Punctuation marks in the script as well as other formal textual signs were taken into account when analyzing the audio recording. Phrase boundaries and the type of the

phrase intonation were marked in the audio wav-file and in the transcript.

After that each phrase was divided into AUs. The AU boundaries are marked in the audio wav-file and in the script. Besides, strong and weak accents for each AU were marked. Each AU of the phrases was analyzed into the nucleus, pre-nucleus and post-nucleus. The next stage of processing was the computation of pitch contours for the phrases, i.e. F_0 values were computed for the vocalized speech segments.

The procedure of speech and text materials processing described above was performed then to analyze individual intonation properties according to AUPs stylization model.

The research was focused on the finality/non-finality intonation types as they are most commonly observed in reading aloud both in Russian and Polish. No consideration has been given to other intonation types, such as interrogation or exclamation. AUPs for various subtypes of final/non-final intonation in Russian and Polish were created with the help of the procedure described in section 3.1. The main attention in this study of language-specific and individual features of pitch contour realization is focused on the final AU of the phrase as the most informative part as far as revealing the peculiarities of a particular intonation's type is concerned.

The generalized results of the language-specific analysis of intonation contours obtained from the Polish and Russian text corpuses described in this section are shown in Fig.6. It displays the AUPs areas in normalized "time-frequency" space for most frequently occurring pitch contours. The AUPs areas include pitch contours of more than 60% of phrases with a final/non-final intonation type in the texts studied. The values of F_{0min} and F_{0max}, used for the normalization of the observed F_0 values (see formula 1), were found at 170Hz and 350Hz for the Polish speaker and 160Hz and 380Hz for the Russian speaker.

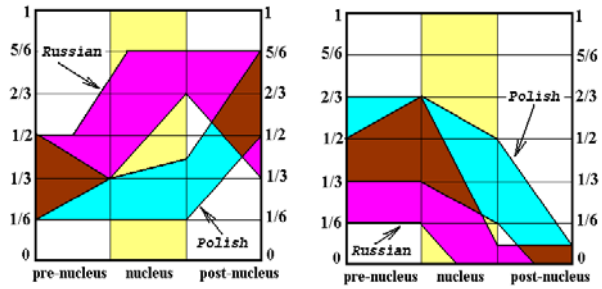


Figure 6: Intonation "portraits" of final AU in Russian and Polish for non-final intonation (on the left) and final intonation (on the right)

As is evident from fig. 6, both final and non-final pitch contours in Russian and in Polish diverge considerably. The most significant differences are on the post-nuclear parts of AU both for non-final and final intonation types. The non-final intonation contour typically characterized by a rising pitch movement is realized in Russian on the nucleus of an AU whereas in Polish it is characterized by the falling pitch change on the nucleus and by the rising pitch change on the post-nucleus. Similar observations hold true for the final intonation contours. The final phrase contour generally characterized by the falling tone is carried in Russian by the pre-nucleus and nucleus of an AU whereas in Polish it is on the nucleus and post-nucleus. This phenomenon can be interpreted by the fact that post-nucleus is almost universally present in a Polish word due to the penultimate-syllable word-stress while in Russian the post-nucleus may be lacking altogether owing to the non-fixed word-stress position.

Pitch contour regularities for the non-final AUs in a non-final and final types of phrase intonation were observed too. It was found that Russian and Polish pitch contours differ not only in the final AU but also in the initial and intermediate AUs of the phrase, although not so significantly.

Fig 7 displays in the normalized "time-frequency" space of AUPs the most frequent pitch contours (about 70% of the overall number) obtained from two Russian speakers for the phrases with final/non-final intonation. The values of F_{0min} and F_{0max}, used for the normalization of the observed F_0 values were found to be equal to 70Hz and 150Hz for the first speaker and 80Hz and 180Hz for the second speaker.

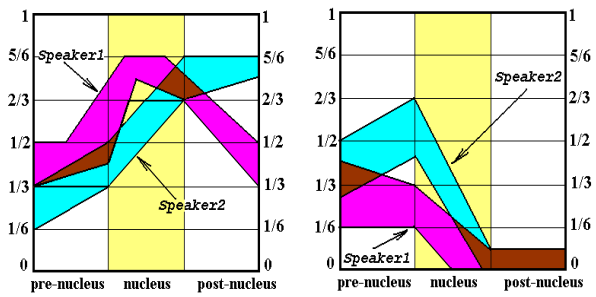


Figure 7: Intonation "portraits" of final AU for two Russian speakers for non-final intonation (on the left) and final intonation (on the right)

3.3. Implementation of intonation contours in TTS system

The implementation of intonation contours in TTS system is provided by the prosodic module the interface of which is shown in figure 8.

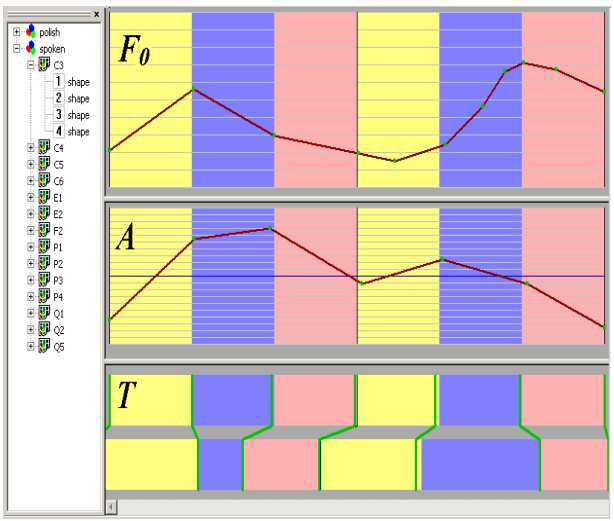


Figure 8: Interface of the TTS prosodic module (an example of the F_0, A and T contours for 2 AUs non-final phrase intonation is shown)

The tonal – (F_0), dynamic – (A) and rhythmical – (T) contours of the phrase are presented by a sequence of prosodic portraits of AUs constituting the phrase. The limitation of the prosodic module used is that a phrase may contain from one to four AUs.

The intonation module provides a basic inventory of prosodic "portraits" of AUs in the various positions within the phrase, and namely: initial, intermediate and final. To determine the intonation type and subtype of the phrase of a text the following indicators are used: the punctuation marks as explicit indicators; coordinative and subordinative conjunctions as well as some other resulting cues of utterance parsing as implicit markers. Using the interface of the TTS prosodic module (fig.8) it is possible to assign the language- and speaker-specific peculiarities by choosing an appropriate set of prosodic AUPs. The module also allows to carry out effective prosodic portrait adjustment as well as changing the values of $F_{0\,min}$ and $F_{0\,max}$.

4. General description of the Slavonic multi-language and multi-voice TTS-synthesizer

The general structure of the multi-lingual and multi-voice TTS-synthesizer looks in the following way (see Fig.9). The incoming orthographic text undergoes a number of successive analytical operations carried out with the help of specialized processors.

The ***textual*** processor is devised to transform the incoming orthographic text into a prosodically marked one. The processor performs the following tasks:

- dividing an orthographic text into utterances;
- transforming numbers, abbreviations, shortenings;
- dividing an utterance into phrases;
- placing word's stress;
- dividing phrases into accentual units (AU);
- marking the intonation type of the phrases;

The prosodically marked text is then sent to ***phonemic*** processor, that performs the following tasks:

- phonemic transcription of the orthographic text.
- determination of positional and combinatory allophones from the in-coming phonemic text;
- generation of the allophone and multi-allophone sequences that are necessary to synthesize.

The ***prosodic*** processor performs the following tasks:

- splitting AU into the elements of accentual units (EAU): pre-nuclear, nuclear and post-nuclear parts;
- generating the fundamental frequency (F_0) contour as well as the amplitude (A) and phoneme duration (T) values according to AUPs for each EAU.

The ***acoustical*** processor uses the information coming from the phonemic and prosodic processors to provide:

- the prosodic parameters modification of ANWs and MANWs;
- concatenation of ANWs and MANWs to the appropriate sequence.

Finally, by concatenating of ANWs and MANWs and their modifications in accordance with the current values of F_0, A, T it generates the ***speech signal***.

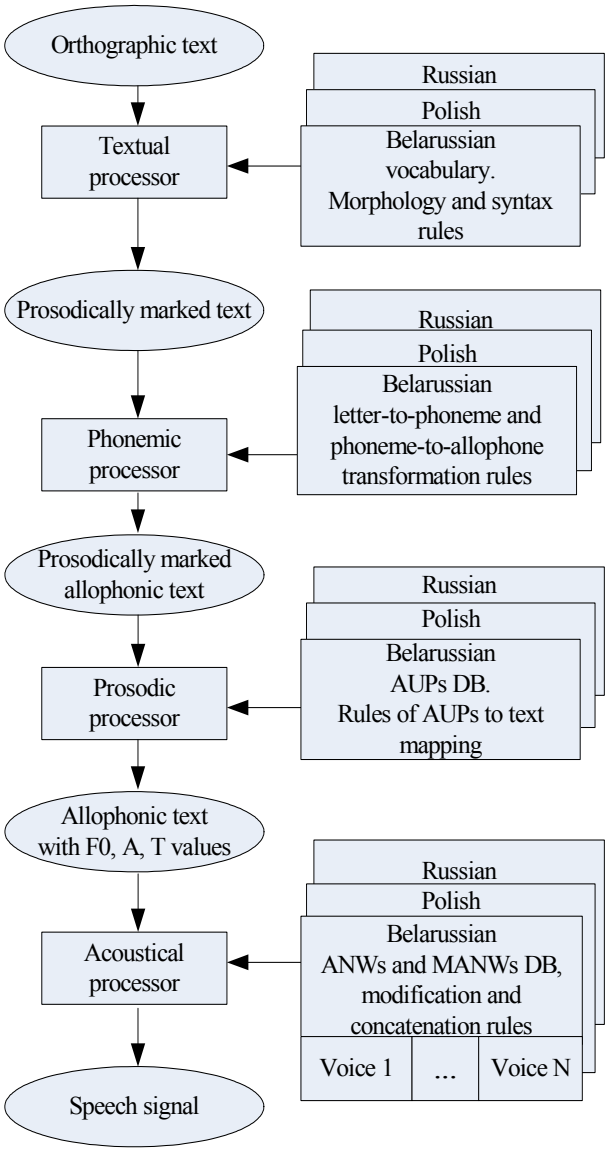


Figure 9: General structure of the multi-lingual TTS – synthesizer

5. Conclusion

In the paper only the basic, key questions of construction of the Slavonic multi-language and multi-voice TTS-synthesis system, concerning phonemic and prosodic peculiarities of speech, are considered. The detailed description of each of TTS system blocks (fig. 9) is beyond given paper.

By present time development a beta-version of Slavonic multi-language and multi-voice TTS-synthesis system is finished and its testing and debugging is carried out**.** Synthetic speech in three languages provided by the system will be demonstrated at the conference during the report.

5. Acknowledgements

This paper was supported by the European Commission under grant INTAS Ref. No 04-77-7404. The authors wish to express their thanks for the support. We should like to thanks our colleagues Dmitry Zhadinets, Andrey Davydov and Oleg Sizonov for their significant contribution to computer modelling and testing of the system. And last not least the authors wish to express the thanks to our Polish colleagues for many useful discussions during the preparation of the paper.

6. References

[1] Lobanov B. "The Phonemophon Text-to-Speech System", *Proc. of the XI International Congress of Phonetic Sciences*, Tallin: 61-64, 1987.

[2] Lobanov B., Karnevskaya H. "MW Speech Synthesis from Text", *Proc. of the XII International Congress of Phonetic Sciences*. Aix-en-Provense, France: 406-409, 1991.

[3] Lobanov B., Jokisch O. et al. "A Bilingual German/Russian Text-to-Speech System", *Proc. of the 3nd International Workshop "Speech and Computer" – SPECOM'98*, St.-Petersburg, Russia: 327-330, 1998.

[4] Lobanov B., Shpilewski E. et al. "Polish TTS in Multi-Voice Slavonic Languages Speech Synthesis System", *Proc. of the International Conference SPECOM'2004*, St. Petersburg, Russia: 565-570, 2004.

[5] Boguslavsky I., Lobanov B., Karnevskaya H. "Generation of Intonation and Accentuation of Synthetic Speech on the Base of Morpho-Syntactic Knowledge", *Proc. of the International Workshop "Integration of Language and Speech"*, Moscow, Russia: 11-28, 1996.

[6] Lobanov B., Karnevskaya H. "TTS-Synthesizer as a Computer Means for Personal Voice (On the example of Russian)", *Phonetics and its Applications*. Stuttgart: Steiner: 445-452, 2002.

[7] Lobanov B., Tsirulnik L. "Phonetic-Acoustical Problems of Personal Voice Cloning by TTS", *Proc. of the International Conference SPECOM'2004*, St. Petersburg, Russia: 17-21, 2004.

[8] Lobanov B. et al. "A personalized text-to-speech synthesized "LobanoPhone-2000", *Proc. of the International Conference "100 Years of Russian Experimental Phonetics"* , St.-Petersburg, Russia:101-104, 2001 (in Russian).

[9] Tsirulnik L. "Automated System for Individual Phonetic-Acoustical Speech Peculiarities Cloning", *Informatics*, 2(10):47-56, 2006 (in Russian).

[10] Silverman, K. et al. "TOBI: a standard for labelling English prosody", *Proc. ICSLP*: 867-870, 1992.

[11] Fujisaki, H. "Prosody, Models, and Spontaneous Speech", *Computing Prosody,* Springer-Verlag: 27–42, 1996.

[12] de Pijper, J. *Modelling British English Intonation.* Foris, Dordrecht, 1983.

[13] Taylor, P. "Analysis and synthesis of intonation using the Tilt model". *J. Acoust. Soc. of America,* 2000.

[14] Ode, C. *Russian intonation: a perceptual description*. Amsterdam, 1989.

[15] Bogdanov D., Krivnova O., Podrabinovitch A., Farsobina V. "The base of Russian language speech fragments "ISABASE", *Proc. Intellectual technologies of information input and output*, Moscow, Russia: 20-23, 1998 (in Russian).