# Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis

*Boris Lobanov, Liliya Tsirulnik, Dmitry Zhadinets &**Helena Karnevskaya

*United Institute of Informatics Problems N.A.S., Belarus
**Minsk Linguistic State University, Belarus
lobanov@newman.bas-net.by

## Abstract

The paper is concerned with the study of final/non-final phrase intonation and its language- and speaker-specific peculiarities. A phrase, according to the model used, is represented by a sequence of accentual units consisting of pre-nucleus, nucleus and post-nucleus. Experimental data obtained from the prosodic analysis of a text, recorded by Russian and Polish native speakers, has made it possible to create accentual units' "portraits" for different types of final/non-final phrase intonation. The implementation of these "portraits" in the unified text-to-speech synthesis system for Slavonic languages with the ability of personal speaking manner cloning is discussed.

## 1. Introduction

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 [1], acoustic-phonetic superpositional models that interpret F0 contours as complex patterns resulting from the superposition of several components [2], IPO model that represent intonation as an inventory of pitch movements [3], and Tilt model that utilizes the continuous parameterization of F0 contours [4]. 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 Russian. A very useful perceptual description of Russian intonation according to the IPO model was developed by C. Ode [5]. Another model, the superpositional one, was successively utilized in the Bell Laboratories Russian TTS system [6].

The main principle of synthesizing prosodic parameters that we have utilized here is based on a 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 [7] and has been used successfully since then in several TTS synthesis models. The present research is carried out in the context of developing a unified text-to-speech synthesis system for Slavonic languages [8] within the framework of the on-going research into the "cloning" of individual speaking manner [9] that involves intensive prosodic studies. The paper is concerned, in particularly, with the study of AUPs finality/non-finality (or completeness/incompleteness in compliance with other terminology) phrase intonation types, namely - its language- and speaker-specific peculiarities. The implementation of these "portraits" in the unified text-to-speech synthesis system for Slavonic languages with the ability of personal speaking manner cloning is discussed.

## 2. Fundamentals of AUP stylization model

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 1, 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 2, where $F_0$ contours of a three-words

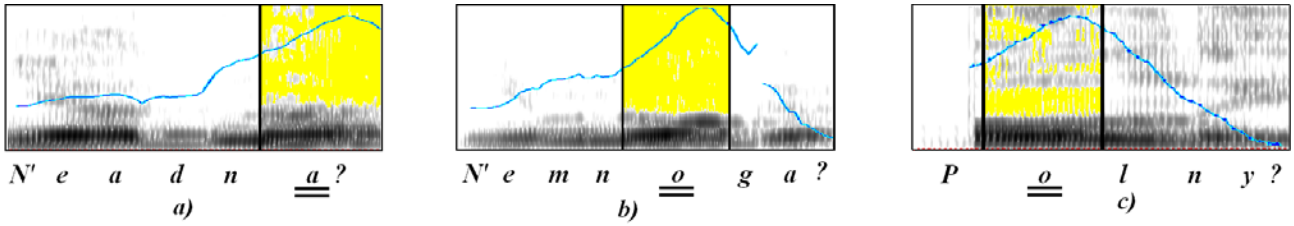



*Figure 1: $F_0$ contours of interrogation for the Russian word-phrases: a) "Не одна?"/N'eadn`a/-"Not one?", b) "Не много?" /N'emn`oga/-"Not much?", c) "Полный?"/P`olny/-"Full up?" (the accented vowels are underlined with a double line)*

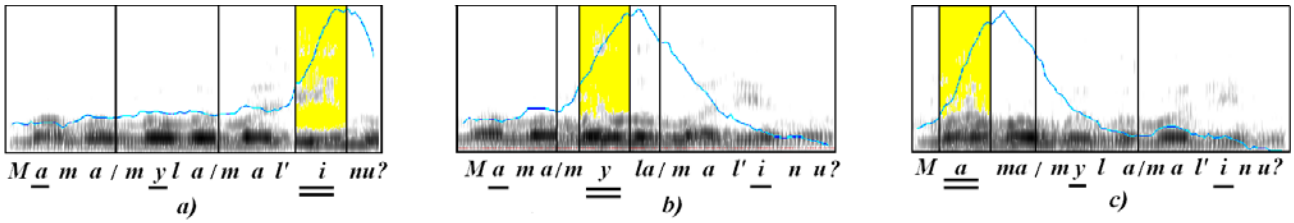
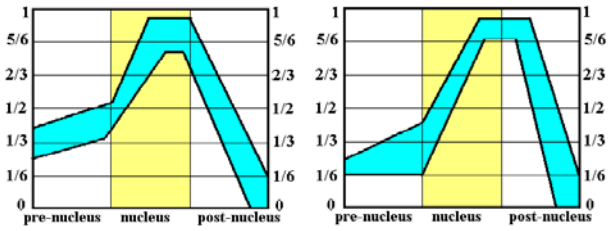



*Figure 2: $F_0$ contours of interrogation for the Russian phrase "Мама мыла малину?" /Mama myla mal'inu/ with the focused words: a)"mal'inu", b) "myla", c) "mama"( the strong accented vowels are underlined with a double line)*

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.2, 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.1.

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. 3a the common AUP of $F_0$ contours for the interrogative type of intonation that corresponds to one-word phrases from fig. 1 is shown, and in fig.3b – the contour corresponding to three-word phrases from fig. 2. As seen from figures 3a and 3b, the difference between AUPs is not very significant.




*Figire 3: AUP for interrogative type of intonation for one-word phrases (on the left) and three-word phrases (on the right)*

As it will be shown from the subsequent experiments discussed below, the regularities of AUPs construction shown above for the interrogation are also typical of the final/non-final types of intonation. Similar considerations can be made concerning the representation of amplitude and timing phrase contours by AUPs.

The main principles of AUs pitch contour "portraits" creation are illustrated in Fig.4 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. 4 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. 4 b).Finally, the AU's pitch and duration are normalized (Fig. 4 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:

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

For the given speaker the $F_{0min}$ value was equal to 70 Hz and $F_{0max}$ – 180 Hz (see Fig.3 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.3 c).

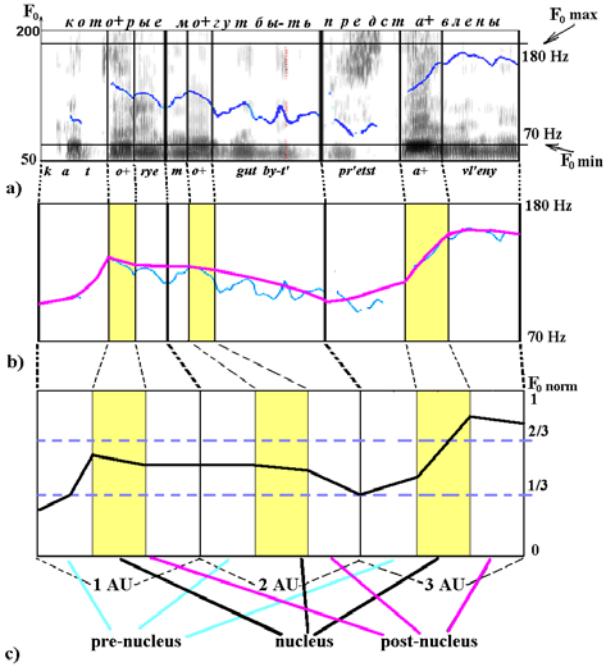



*Figure 4: Scheme of pitch contours "portrait" creation: a) $F_0$ values computation, b) $F_0$ curve interpolation, c) $F_0$ curve normalization*

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. Experimental material and annotation

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. 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 [10]. 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.

## 4. Experimental results and discussion

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 2. 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 section 3 are shown in Fig.5. 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.

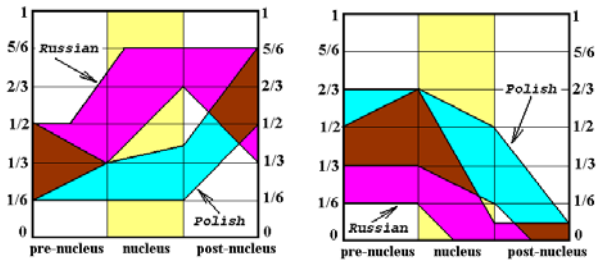



*Figure 5: Intonation "portraits" of final AU in Russian and Polish for non-final intonation (on the left) and final intonation (on the right)*

As is evident from fig. 5, 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 6 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.

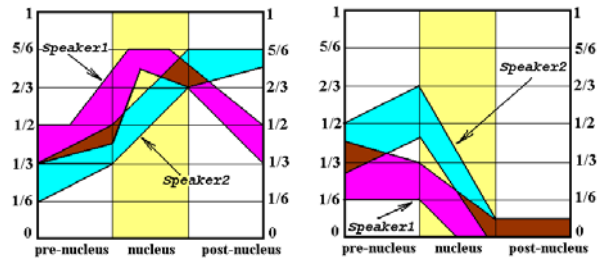



*Figure 6: Intonation "portraits" of final AU for two Russian speakers for non-final intonation (on the left) and final intonation (on the right)*

As is evident from figure 6, intonation "portraits" for these two particular speakers diverge considerably for both non-final and final intonation. Significant differences are observed on the nucleus and post-nucleus for the non-final intonation type and on the pre-nucleus and the nucleus - for the final intonation type.

The experimental results thus obtained have confirmed the usefulness of the suggested AUP stylization model of phrase intonation for language- and speaker-specific peculiarities representation in TTS synthesis. The implementation of these "portraits" in the unified text-to-speech synthesis system for Slavonic languages with the ability of personal speaking manner cloning is described below.

## 5. Implementation 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 7. 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.

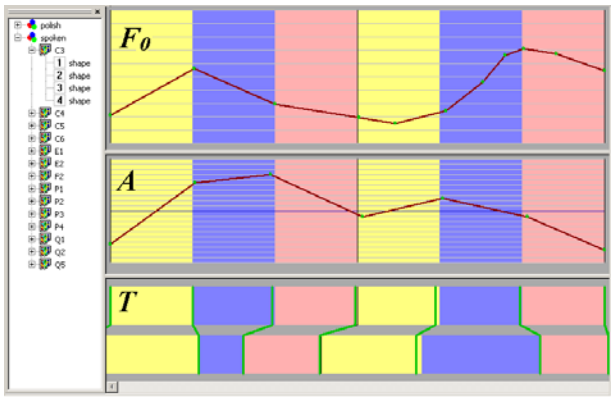



*Figure 7: Interface of the TTS prosodic module (an example of the $F_0$, A and T contours for 2 AUs non-final phrase intonation is shown)*

Any phrase of a text is considered to belong to one of the following intonation types: finality, non-finality, interrogation, exclamation. Besides, every type of intonation has several subtypes. At the moment we utilize 5 variants of finality, 18 variants of non-finality, 6 variants of interrogation and 6 variants of exclamation. For each of the variants of intonation the 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.7) 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 $F_{0\ max}$.

## 6. Conclusion

The paper has presents the first results of the quantitative analysis of the pitch contours for two Slavonic languages and, besides, the peculiarities of two Russian speakers individual intonation have been revealed. This made it possible to create a basic set of normalized "portraits" of pitch contours (AUPs) to assign some of the language-and speaker-specific peculiarities to synthetic speech with different prosodic types. The report will be illustrated by phonograms of speech synthesized by the intonation rules developed.

## 7. Acknowledgment

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.

## 8. References

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

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

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

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

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

[6] Pavlova E.; Pavlov Y.; Sproat R.; Shih Ch.; van Santen J., 1997. Bell laboratories Russian text-to-speech system. Proc. *Eurospeech'97*. Rhodes-Greece, 2451-2454.

[7] Lobanov B., 1987. The Phonemophon Text-to-Speech System. Proc. *ICPhS*. Tallin, 61-64

[8] Hoffmann R., Shpilewsky E., Lobanov B., Ronzhin A., 2004. Development of multi-voice and multi-language Text-to-Speech (TTS) and Speech-to-Text (STT) conversion system (languages: Belorussian, Polish, Russian). Proc. *Int. Conf. SPECOM'2004*. St.-Petersburg, 657-661.

[9] Lobanov B., Tsirulnik L., 2004. Phonetic-Acoustical Problems of Personal Voice Cloning by TTS. Proc. *Int. Conf. SPECOM'2004*. St.-Petersburg, 17-21.

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