Auditory Estimation of Effectiveness of the AUP-Stylization Model of the Melodic Contour

*Boris Lobanov, **Helena Karnevskaya

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

Abstract

The present paper is concerned with the evaluation of perceptual discrimination of synthesized melodic (pitch) contours within the framework of the AUP stylization model, proposed earlier by the authors. It is argued that variation of the prosodic characteristics of synthesized speech is needed for conveying modal-pragmatic varieties of statements, imperatives, special and general questions and other utterance types. As a result of a series of listening tests it has been proved that AUP stylization model can be successfully used for subtle modifications in the pitch contour, which are perceptibly significant and can be evaluated by the listeners in terms of degrees of similarity with the pitch contours derived from natural speech.

1. Introduction

Enriching the intonation repertoire may be an effective source of ensuring better quality of synthesized speech. This goal can be reached, among other means, by implementing predictable variation within prosodic contours assigned to different communicative-syntactic utterance types: statements, imperatives, general and special questions, etc. The idea of "prosodic contour type"/"utterance type" co-occurrence has dominated the approach to creating the inventory of prosodic contours to be utilized in TTS-synthesis as well as the rules of their distribution and selection. This approach is functional in that it presupposes prosodic differentiation of the overall aim of communication as it is reflected in individual utterances. It is also convenient, particularly for TTS-synthesis purposes in that the choice of the contour relies on explicit lexical-syntactic clues.

The main principle of synthesizing prosodic parameters that we have utilized here is based on a model represented by a sequence of Accentual Unit Portraits (AUP-stylization model) [1, 2]. It was proposed over ten years ago and has been used successfully since then in several TTS synthesis models.

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 (accented) syllable. This syllable is the nucleus of an AU, whereas all the syllables preceding it form the pre-nuclear part and all the syllables following it form the post-nuclear part.

Despite the fact that AUP stylization model has been successfully applied in many versions of TTS-synthesis, no attempt has been undertaken so far to evaluate the subjective quality of intonation, synthesized according to the given principle. The present paper is intended to fill in the gap by setting an aim of auditory estimation of the AUP-stylization model effectiveness using the method of "quality subjective opinion testing" [3].

2. Method

2.1. Material

The database (DB) of intonation of interrogative phrases of the Russian dialog [4] was used for the present investigation. This DB contains about 400 samples for 8 types of General and 4 types of special questions, each type of both question-groups being represented by up to more than 10 various subtypes. The given intonation DB includes over 20 types of pitch accents, the acoustic-perceptual discrimination of which is associated in their semantic analysis with pragmatic varieties (subtypes) of questions. Importantly, not all of the question subtypes in the DB are distinguished by prosodic modifications alone. There are interrogative structures containing lexical-syntactic markers of a pragmatic kind. Four of such subtypes of General questions were selected to serve as the experimental material for the investigation reported in this paper. Their semantic-pragmatic labels are the following:

1. Verifying question with the particle *'ли'* ('if');
2. 'Asking oneself' question (rhetorical);
3. Genuine interrogation with the particle *'неужели'* ('really');
4. Certifying question with the particle *'правда'* ('isn't it/he, she etc.');

These varieties of General question were chosen, firstly, due to the significance of their intonational differences and, secondly, due to the presence of special lexical markers of each of the varieties, which simplifies their identification for TTS-synthesis.

At the next stage tokens of each subtype were selected and their phonograms were subjected to instrumental analysis with the help of the system "IntoClonator" [5]. The following tokens representing the aforesaid 4 varieties of Russian General questions were analyzed:

1. **Удал1О(/-)сь ли ему снять *квартИ(\)ру?*
2. **Не подар1И(//)ть ли мне ему *щенк2А(-\)?*
3. **Неуж1е(//-)ли он об этом *не зн2а(\\)л?*
4. **Пр1А(/)вда *симпат2И(\)чный?*

All these interrogative phrases consist of 2 accentual units (AU). The accented words are indicated with an asterisk (*) placed before the word to mark the beginning of the AU; **1**

and **2** indicate the order of the AU, the first and the second Aus, respectively. The vowel of the nucleus of the AU is printed in capital letters; the brackets following the nucleus contain the symbol of the type of the tonal accent as it is marked in the above-mentioned intonation DB from which it is taken.

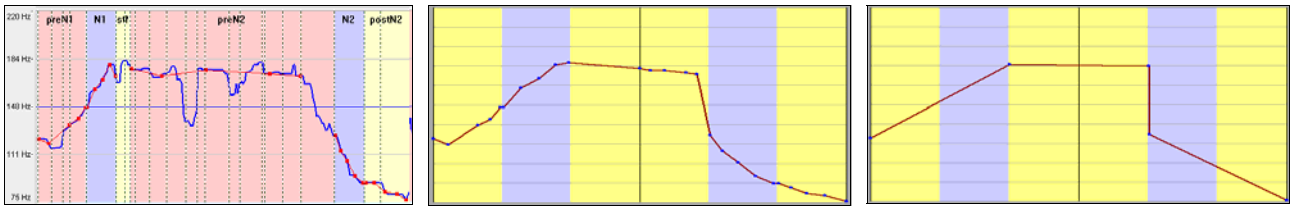
The overall intonation pattern of a phrase is formed by the combination of the tonal accents of the successive tonal units. Thus the above tokens are represented by four intonation patterns:

Intonation Type I: /- Rising-Level + \ Falling Neutral
Intonation Type II: // Wide Rise +/-\ Rising-Level-Falling
Intonation Type III: // Wide Rise + \\ Wide Fall
Intonation Type IV: / Rising Neutral + \ Falling Neutral

Figures 1–4 below demonstrate the F0 contours of each of the tokens obtained with the help of the "IntoClonator":

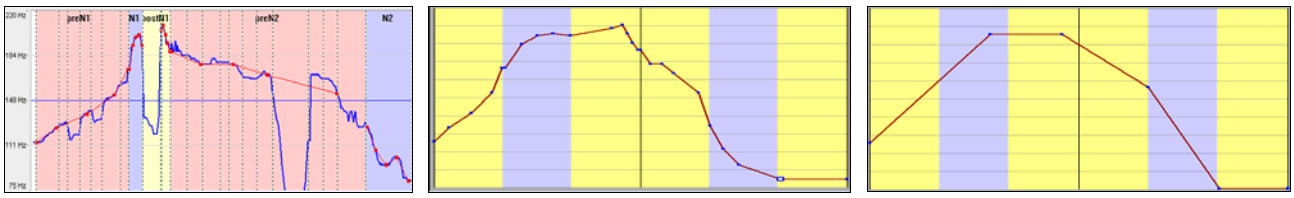
a) the original F0 contour;
b) the AUP stylization of the first and second accentual units of the original F0 contour;
c) a simplified AUP stylization of the original F0 contour.

The F0 contour (c) in each figure was simplified by means of approximating the representation of the pitch movements within the accentual units by straight lines. This simplified version of an F0 contour of the AUPs resembles the labelling of the pitch accents in the intonation DB [4], where their identification relies entirely on the direction of the pitch change, pitch level and width of the pitch interval. The AUP's stylization model, by comparison, takes account, besides, of the overall configuration of the F0 movement and shape of the pitch change over the sequence of syllables embraced by the AU.



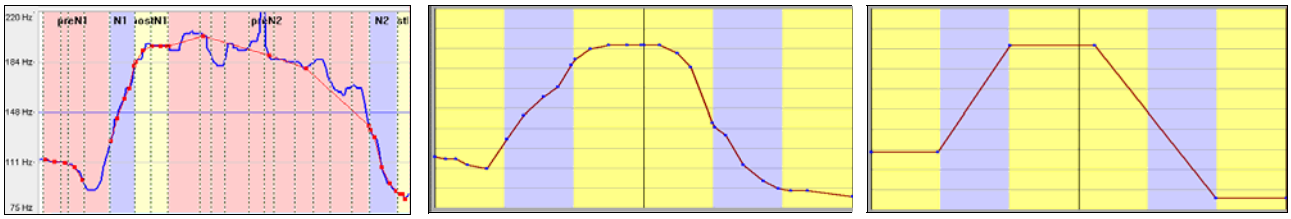
a) b) c)

Fig. 1. F0 contour for the phrase "*Удалось ли ему снять квартиру?" – Has he managed to hire a flat?*
(Verifying question with the particle *'ли'*): a) Original F0 contour, b) F0-AUP, c) simplified F0-AUP



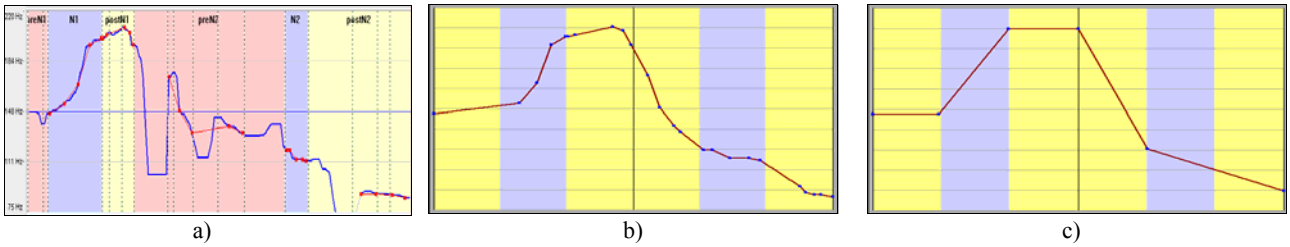
a) b) c)

Fig. 2. F0 contour for the phrase "*Не подарить ли мне ему щенка?" – Shouldn't I give him a puppy as a present?*
('Asking oneself' question with the particles *'не'*, *'ли'*): a) Original F0 contour, b) F0-AUP, c) simplified F0-AUP



a) b) c)

Fig. 3. F0 contour for the phrase: *Неужели он об этом не знал? – Didn't he know about it?*
(Genuine interrogation with the particle *'неужели':* a) Original F0 contour, b) F0-AUP, c) simplified F0-AUP



a) b) c)

Fig. 4. F0 contour for the phrase: *Правда симпатичный? – Isn't he handsome?*
(Certifying question with the particle *'правда'):* a) Original F0 contour, b) F0-AUP, c) simplified F0-AUP

2.2. Testing procedure

The perceptual Testing procedure included three categories of stimuli.

1. The original speech signal selected from DB [3].
2. Synthesized speech signal obtained by replacing the natural F0 contour with its AUP stylization (AUP).
3. Synthesized speech signal obtained by replacing the natural F0 contour with its Simplified AUP-stylization (SAUP).

Evaluation of AUP stylization model effectiveness was obtained as a result of a series of listening comparison tests. Subjects were presented with pairs of sound stimuli (the first and the second stimulus). The first stimulus in each pair was the original speech signal, whereas the second might be either a synthesized signal or the same original one. The listeners were instructed to give 2 marks (from 2 to 4) as a response to 2 questions. The first question was: "What's the degree of intonation similarity of the second stimulus to the first one?" The second question was "What's the 'voicing' quality (the way it sounds) of the second stimulus compared to the first one? Is it very good – (4), fairly good – (3), or poor – (2)?" For each of the subtypes of General Qustions the following pairs of stimuli were prepared (see table 1).

Table 1: Description of the pairs of stimuli

№	Intonation Sybtype of the 1st stimulus	Stimulus Code	Description of the 2nd stimulus in the pair
1.	1	10	Original, Int. type I
2.	1	11	Synth. with AUP, Int. type I
3.	1	12	Synth. with AUP, Int. type II
4.	1	13	Synth. with AUP, Int. type III
5.	1	14	Synth. with AUP, Int.type IV
6.	1	15	Synth. with SAUP, Int.type I
7.	2	20	Original, Int. type II
8.	2	21	Synth. with AUP, Int. type I
9.	2	22	Synth. with AUP, Int. type II
10.	2	23	Synth. with AUP, Int. type III
11.	2	24	Synth. with AUP, Int. type IV
12.	2	25	Synth. with SAUP, Int. type II
13.	3	30	Original, Int.type III
14.	3	31	Synth. with AUP, Int.type I
15.	3	32	Synth. with AUP, Int. type II
16.	3	33	Synth. with AUP, Int. type III
17.	3	34	Synth. with AUP, Int. type IV
18.	3	35	Synth. with SAUP, Int. type III
19.	4	40	Original, Int. type IV
20.	4	41	Synth. with AUP, Int. type I
21.	4	42	Synth. with AUP, Int. type II
22.	4	43	Synth. with AUP, Int. type III
23.	4	44	Synth. with AUP, Int. type IV
24.	4	45	Synth. with SAUP, Int. type IV

Apparently, both stimuli in each pair are identical in their phonemic structure. However, their intonation types were either the same, when the second stimulus in the pair was synthesized with the Intonation type of the original token ('own' stimuli: 10, 11, 15, 20, 22, 30, 33, 35, 40, 44, 45), or not the same, when the second stimulus in the pair was synthesized with one of the other intonation types ('alien' stimuli: 12, 13, 14, 21, 23, 24, 31, 32, 41, 42, 43)

The synthesized signal in each pair was recorded at 1 sec. interval after the original one. Token pairs were then presented to the listeners through headphones and were separated from each other by 10 seconds of silence. The auditory tests were carried out by five subjects – all experts in phonetics from Minsk State Linguistic University, familiar with testing requirements and principles of intonation analysis and evaluation. Their marks were registered in a special response-sheet.

3. Results and Discussion

The data obtained were statistically verified and proved statistically significant. The results are shown in tables 2 –3 and figures 5–6.

The results of the evaluation of the degree of stimuli similarity in intonation as well as their similarity (sameness) in the voicing quality are shown in tables 2 and 3, respectively.

Table 2: Intonation similarity of the stimuli pairs

Intonation type of stimuli	Type of a pair			
	Both original	Original **vs** AUP (own)	Original **vs** SAUP(own)	Original **vs** AUP(alien)
	1	2	3	4
1. Verifying question with the particle *'ли'*	4,0	3,9	3,3	2,1
2. 'Asking oneself' question (rhetorical)	4,0	3,6	2,4	2,5
3. Genuine interrogation with the particle *'неужели'*	4,0	3,6	3,6	2,5
4. Certifying question with the particle *'правда'*	4,0	3,4	3,0	2,0
Absolute Mean values	4,0	3,6	3,1	2,3
Normalized Mean values	1	0,6	0,3	-0,7

The marks in column 1 refer to the pairs with two original tokens, i.e. identical not only in the phonemic structure but in the intonation type as well. The marks in column 2 and 3 refer to the pairs in which the second stimulus was a synthesized AUP-stylization (column 3) or a synthesized SAUP-stylization (column 4) of the original F0 contour. Both AUP and SAUP stylization contours in this case represent the intonation type of the original token. The marks in column 4, on the other hand, show the results relating to the degree of perceptible similarity between the original token and a phonemically identical stimulus synthesized according to AUP's stylization but with another intonation type, belonging to a different original token. For example, the mark 3,1 in column 4 of table 2 was obtained as a result of comparing the original token with Intonation Type I "*Не подарить ли мне ему щенка?" – "Shouldn't I give him a puppy as a present?"* with a phonemically identical stimulus synthesized with an 'alien' intonation contour, borrowed from Type 2, 3 or 4.

As a way of summing up the data presented in table 2, it seems important to point out the following.

1. The marks in column 2 as compared to the frame of reference – the marks in column 1 – appear to approach the 'very good mark' (on average, only 0,4 scores lower), which testifies to a fairy high quality of the AUP's stylization of an F0 contour of a given type.
2. In cases when subjected to comparison were two phonemically and lexically identical tokens (one natural and the other – synthesized), but the AUP's stylization is based on a different intonation type ('alien' stimuli), the marks are nearer to 'poor', which proves the effectiveness of AUP stylization for perceptual discrimination of F0 contours representing different intonation types.
3. The application of a simplified SAUP stylization leads to a noticeable drop in the marking values, which is an argument in favour of using a complete model of AUP's stylization.

The data presented in table 3 make it possible to draw conclusions similar to those referring to the data in table 2. Yet, the values of marking here are lower even for the pairs of two original identical tokens.

Table 3: Voicing quality similarity of the stimuli pairs

Intonation type of stimuli	Type of a pair			
	Both original	Original **vs** AUP (own)	Original **vs** SAUP(own)	Original **vs** AUP(alien)
	1	2	3	4
1. Verifying question with the particle *'ли'*	3,7	3,4	2,8	2,2
2. 'Asking oneself' question (rhetorical)	3,5	3,4	2,4	2,4
3. Genuine interrogation with the particle *'неужели'*	3,6	3,7	2,3	2,8
4. Certifying question with the particle *'правда'*	4,0	3,7	2,7	2,9
Absolute Mean values	3,7	3,6	2,8	2,6
Normalized Mean values	0,7	0,6	-0,2	-0,4

A possible explanation of this fact might be greater complexity of the task for listeners who had been unaware of the peculiarities of stimulus combinations and had subconsciously anticipated 'obligatory' difference.

Figures 5–6 demonstrate the mean absolute and normalized values of quality similarity estimation.

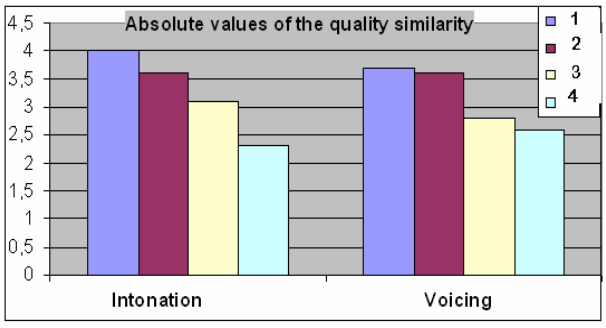


Fig. 5. Absolute values of the quality similarity of intonation and voicing for the different intonation types, namely: I – (1), II – (2), III – (3), IV – (4)

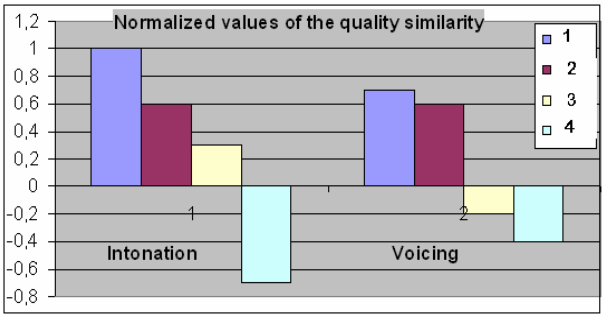


Fig. 6. Normalized values of the quality similarity of intonation and voicing (legends as in **Fig. 5**)

The graphical representation of the data in figures 5–6 illustrates the above conclusions by comparing both the absolute and normalized values of perceptual quality similarity for four different intonation types, namely I – (1), II – (2), III – (3), IV – (4).

4. Conclusions

The underlying hypothesis of the present research was the efficiency of the AUP's stylization model for implementing subtle modifications within F0 contours, thus producing modal-pragmatic varieties of basic types of utterance and improving the quality of synthesized speech. Perceptual discrimination of the pitch modifications proved by the different degrees of subjective similarity of the contours being compared has confirmed this assumption. The AUP's stylization model has demonstrated a potential flexibility towards the framework of its invariant structural principle. At the same time, the results of perceptual testing show that a simplified version of AUP's stylization causes the quality of synthesized speech to diminish.

5. References

[1]. Lobanov B., Tsirulnik L., Zhadinets D., Karnevskaya E. "Language- and Speaker Specific Implementation of Intonation Contours in Multilingual TTS Synthesis", *Proc. of the 3rd International conference "Speech Prosody",* Dresden, Germany, May 2–5: 553-556, 2006,

[2] Lobanov B., Tsirulnik L., Sizonov O. "AUP's Modeling of Speaker Specific Intonation Contour Peculiarities", *Proc. of the 12-th International conference "Speech and Computer: SPECOM'2007"*, Moscow, Russia: 312-317, 2007.

[3] Method for subjective performance assessment of the quality of speech voice output devices, *ITU-T Recommendation – Geneva: ITU-T Study Group 12,* p. 13, 2004.

[4] Kodzasov S. At al. "Baza dannyh 'intonatsija russkogo dialoga': voprositel'nyj repliki (in Russian)", *Proc. of Int. Conf. DIALOGUE'2005,* Moscow, Russia: 245-249, 2005.

[5] Lobanov B., Tsirulnik L., Sizonov O. "«IntoClonator» – kompjuternaja sistema klonirovanija prosodicheskih harakteristik rechi (in Russian)", *Proc. of Int. Conf. DIALOGUE'2008,* Moscow, Russia: 330-338, 2008.