Inventory of Intonation Contours for Text-to-Speech Synthesis

Tetyana Lyudovyk, Valentyna Robeiko

International Research/Training Center for Information Technologies and Systems,
Kyiv, Ukraine

{tetyana_lyudovyk, robeiko}@uasoiro.org.ua

Abstract

This paper presents an intonation model which determines intonation contours over intonation phrases. The model is described by four elements: communicative type of an intonation phrase; number of accent groups in it; position of the nuclear accent group in it; and set of target intonation points. Individualization of the model is based on semi-automatic analysis of speaker database. The model was implemented in unit selection TTS system for Ukrainian.

1. Introduction

Modeling of intonation contributes to speech science by introducing clarity and defining more precisely the intonation system of a language. Adequacy of an intonation model can be tested by imposing it on synthetic speech.

The goal of modeling consists of analysis and generalization of intonation phenomena and their representation in a parametric form that is compact yet preserves the naturalness of intonation in the synthesized speech.

There is no generally accepted intonation model. Intonation models in TTS systems have varied from rule-based models derived from expert knowledge to data driven statistical models.

General disadvantage of rule-based methods of intonation modeling consists in their inflexibility and insufficient account of individual speaker peculiarities.

Data-driven methods of intonation modeling often do not promote the understanding of linguistic phenomena. Using empirically composed large sets of features with wide range of values sometimes makes modeling "blind", because the relative importance of different features is unknown [1].

Our approach to intonation modeling combines rule-based and data-driven methods by defining general set of intonation contours and a procedure of individualizing these contours based on the automated analysis of speech data.

The intonation model described in this paper has been realized in the TTS system for Ukrainian [2, 3].

2. Stages of general intonation model development

Speech communication is based on general models common for all people speaking a particular language.

In this work intonation modeling is based on the assumption that the intonation serves primarily the communicative function. Intonation can be understood as the systematic use of pitch for communication [4]. P. Taylor notes one of the reasons why good models of prosody have proved hard to develop is that researchers have often tried to study prosody without reference to its communicative function.

Three stages of intonation analysis have been carried out. First, we began with the acoustic-phonetic study of ten non-annotated speech corpora and six prosodically annotated speech databases of different speakers to examine Lobanov's intonation model.

2.1. Lobanov's model of intonation

Lobanov's intonation model [5] has been successfully used for a long time in TTS systems for Russian. According to this model, the minimal intonation 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 part (all the phonemes following the fully stressed syllable). Phonemic content and number of syllables in the pre- and post-nucleus do not influence significantly the intonation contour of a certain type of phrase intonation.

Phrase intonation is characterized by:

- phrase type (finality, non-finality, interrogation, exclamation etc.);
- number of AUs.

For example, a declarative phrase composed of 4 AUs is marked as F-4. The last AU in a phrase is considered the prominent one, because usually the distinct intonation movement is associated with a phrase end.

Each AU is described by a set of target intonation points, which determine F0 values. F0 values between these target points are calculated by means of linear interpolation.

The Lobanov's model has two significant advantages. The first one concerns the distinction between informationally important (nucleus) and non-important (pre-nucleus and post-nucleus) portions of an AU. The second advantage concerns the detailed description of the nucleus by six target intonation points, which allows to model slight but categorical details and thus contributes to the natural quality of generated intonation.

2.2. Preliminary study of Ukrainian intonation

The acoustic-phonetic study of 10 speech corpora revealed general regularities of Ukrainian intonation. Table 1 presents analyzed speech material. Special attention has been paid to the comparative analysis of material obtained from different speakers reading the same text.

Six speech databases have been created under the framework of unit selection speech synthesis on the basis of the six speech corpora: two from male voices (isolated

Table 1. Speech material analyzed at the preliminary stage

Speaker	Text type	Number of intonation phrases with neutral intonation				Number of intonation phrases with logical or emphatic accent
		Finality	Non-finality	Question	Exclamation	
Svyatoslav	isolated sentences	267	268	3	14	114
Olexandr	isolated sentences	5	6	—	17	—
Yuriy	isolated sentences	4	7	—	9	—
Larysa	isolated sentences	11	13	—	13	—
Dmytro	radio news	15	31	—	—	9
Anzhelika	radio news	11	4	—	—	—
Viola	radio news	14	12	—	—	—
Valentyna	instructions	18	8	15	—	6
Mykola	radio interview	—	—	23	—	—
Maryna	isolated sentences	74	78	7	6	3

sentences and radio news) and four from female voices (isolated sentences, radio news, and instructions).

Experiments with the TTS system for Ukrainian have shown that Lobanov's intonation model can be successfully used under the unit selection framework (earlier we used this model with a formant synthesizer).

2.3. Correction of intonation model

2.3.1. Units of intonation

Intonation phrase (syntagm) is considered the basic intonation unit. Intonation phrase (IP) is divided into accent groups (minor phrases). An accent group (AG) consists of one or more words united by one accent.

Pre-nuclear, nuclear and post-nuclear parts can be distinguished in an IP intonation contour, each of them carrying a different functional load. The nuclear part (nuclear, main, prominent AG) is an intonation center of an IP. It has a distinct F0 contour which allows to differentiate communicative types (discourse situations). Pre-nuclear and post-nuclear parts of an IP are optional.

The intonation model determines the intonation contour over the whole IP, rather than over syllables [6], or over words or phonemes. The intonation contour is represented by a sequence of target F0 points.

2.3.2. Model elements

Investigation of the speech material and experiments with synthesized speech shown poor intonation modeling results for IPs with logical or emphatic accent placed on any AG other than last. This led to the inclusion of one more element into the intonation model, namely position of a prominent AG in an IP.

The general intonation model determines the intonation contour over the whole intonation phrase and is described by four elements [3]:

- communicative type of an IP;
- number of AGs in an IP;
- position of the prominent (main, nuclear) AG in an IP;
- set of target intonation points.

The number of AGs is determined by the number of accented vowels in the IP. The position of the nuclear AG corresponds to the position of the last AG if there is no logical or emphatic accent. Otherwise the position of the nuclear AG corresponds to the position of the AG carrying the logical or emphatic accent.

Target intonation points determine F0 values. The accent center (nucleus) of an AG is its accented vowel. It is modeled by 6 target points. The part of an AG preceding the accent center is modeled by 2 target points. The same applies to the succeeding part of an AG. Thus, an IP with 3 AGs is described by 30 intonation points; an IP with 4 AGs is described by 40 intonation points, etc. F0 values between target points are calculated by means of linear interpolation.

2.4. Deeper study of intonation and testing of the intonation model

Aiming at reflecting the full range of communicative functions in synthesized speech, we selected for our work a Ukrainian fiction text with dialogues (80 minutes) read by the professional male speaker Valeriy.

2.4.1. Recordings and database development

The recording sessions were not monitored. In fact, a speaker received an orthographic text, made recordings in a quiet room within one day (two sessions), and supplied these recordings to the researchers. A speech database containing 18785 units (phones-in-context) was developed with manual correcting of automatically obtained transcription and segmentation into phones.

Stressed and unstressed vowels are treated as different phonemes.

The manual correction of segmentation assured appropriate pitch synchronous boundaries between phones. Segmentation of units into pitch periods was carried out automatically. Unvoiced phones were not segmented.

2.4.2. Intonation annotation in the speech database

The database annotation contains no high-level linguistic information nor symbolic prosodic labels like ToBI [1].

To annotate intonation, we rely only on objective low-level numerical feature that is the pitch period length. We claim that the sequence of pitch period lengths during an intonation phrase is the best intonation description independent of any intonation theory.

Borders between intonation phrases and between words are unmarked. We found that it was incorrect to mark

borders relying on corresponding orthographic text (as in [7]), because the speaker often violates syntactic structure and ignores punctuation marks like question marks or points. To mark borders automatically without mistakes relying on acoustic cues (e. g. pauses) is also incorrect, because sometimes intonation phrases are not divided by a pause (3 % of borders between intonation phrases), while often there is an inner pause within an intonation phrase (25 %). Sometimes the speaker feels the need to place phrase breaks at equal intervals somewhat independently of the top down linguistic structure [4].

The analyzed speech material contains many cases of prominence, which are difficult to mark automatically.

3. Communication types of intonation phrases

Our goal is to analyze prosodic annotations of a speech database and to create an inventory of intonation contours related to this database and thus relative to the speaker intonation.

Two of the intonation model elements (number of AGs in an IP and set of target intonation points) can be found easily given the database annotation supplemented with breaks between IPs. But the other two elements, the position of the prominent AG in an IP and, most importantly, the communicative type of an IP, turned out to be difficult to identify not only automatically but even by phoneticians.

We began with a list of 10 communicative types: finality, non-finality, wh-question, yes/no question, exclamation, contrast, explication, parenthesis words, expressive finality, and not-identified type. In many cases it was difficult to distinguish between types, for example between non-finality and contrast, or between exclamation and expressive finality.

Thus we carried out a perception experiment aimed at analyzing what communicative types are assigned by listeners to different intonation phrases from real speech. 20 listeners (students and professors of linguistic university, all native speakers of Ukrainian) were asked to listen to 49 IPs selected from the investigated speaker's recordings. Listeners were supplied with a list of 9 communicative types and the orthographic text corresponding to the recordings, where punctuation marks were absent and all the words were in lower case. The task was to indicate the communicative type of each IP. Each IP was played three times, and the experiment lasted 30 minutes.

This experiment helped correct the list of communication types present in the speaker's recordings. Thus, communicative types of explication and contrast have been excluded. On the contrary, 3 communicative type have been added: enumeration; attributive relative clause with a relative pronoun; and first part of complex wh-question.

Table 2 represents the resulting distribution of IPs contained in the investigated read fiction text.

3.1.1. Stylization

In order to compare IPs with different segmental structure and to derive invariant intonation contours the stylization of F0 tracks was performed.

Stylization consists in determining of F0 values at the target intonation points. Intonation contour of one-accent IP is described by 10 F0 values:

$$F0_1^{\,1}, F0_1^{\,2}, ..., F0_1^{\,10},$$

where $F0_1^{\,1}$ is the F0 value at the first not unvoiced phoneme (vowel or voiced consonant) of the AG;

$F0_1^{\,2}$ is the F0 value at the last not unvoiced phoneme of the AG among phonemes preceding the accent center;

$F0_1^{\,3}$ ... $F0_1^{\,8}$ are F0 values at the accent center;

$F0_1^{\,9}$ is the F0 value at the first not unvoiced phoneme among phonemes succeeding the accent center;

$F0_1^{\,10}$ is the F0 value at the last not unvoiced phoneme of the AG.

Intonation contour of an IP divided into n AGs is described by $10n$ F0 values:

$$F0_1^{\,1},\ F0_1^{\,2}, \dots,\ F0_1^{\,10},\ F0_2^{\,1},\ F0_2^{\,2}, \dots,\ F0_2^{\,10}, \dots,$$
$$F0_n^{\,1},\ F0_n^{\,2}, \dots,\ F0_n^{\,10}.$$

Table 2. Distribution of intonation phrases with different communicative types

Communicative type of intonational phrase	Total number of intonation phrases	Number of accent groups in an intonational phrase						
		1	2	3	4	5	6	7
neutral finality	350	49	94	113	67	18	4	5
expressive finality	207	11	57	51	37	39	7	5
non-finality	327	58	112	95	34	23	2	3
yes/no question	21	12	2	2	1	2	2	—
wh-question	20	—	6	10	3	—	1	—
exclamation	59	19	18	9	10	2	1	—
enumeration	17	5	8	—	4	—	—	—
parenthesis words	10	2	7	—	1	—	—	—
first part of complex wh-question	11	—	2	5	2	2	—	—
attributive relative clause with a relative pronoun	7	—	2	5	—	—	—	—
unidentified	26	8	8	7	3	—	—	—
total	1055	164	316	297	162	86	17	13

Figures 1 and 2 represent non-stylized and stylized intonation contours of the IP „А ви прислухайтесь," ("Lend your ear"), consisting of two AGs.

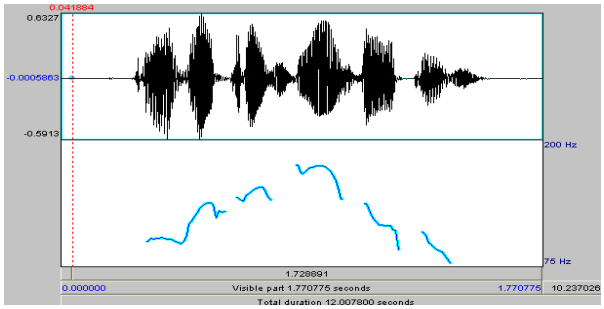

Figure 1: Oscillogram (top) and non-stylized intonation contour (down) of the intonation phrase „А ви прислухайтесь," ("Lend your ear") uttered by the speaker Valeriy.

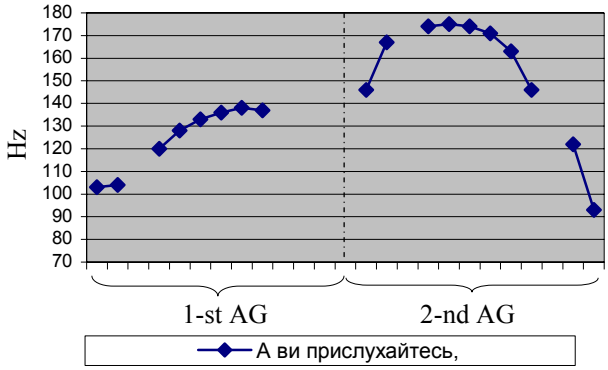

Figure 2: Stylized intonation contour of the intonation phrase „А ви прислухайтесь," ("Lend your ear") uttered by the speaker Valeriy.

Stylization allows comparing IPs with different segmental content abstracting from intonationally insignificant segments and microprosody influence, e.g. F0 change at consonants [8]. For example, there is an F0 lowering by ≈10 Hz (speaker Valeriy) and by ≈15-20 Hz (speaker Svyatoslav) at voiced fricatives. It is considered traditionally that F0 variation caused by segmental structure of speech segments are not perceived as intonationally significant. Experiments testify that F0 contour may be considerably simplified without a loss of intonation perception.

Now the stylization of intonation contours is performed in automated mode using speech database annotations containing the information about pitch periods lengths and, therefore, about F0 movement.

The most difficult non-automated stage of the stylization is the detection of IPs borders, because not all IPs are separated by pauses and not all pauses indicate such borders. We plan to analyze the dependence of the presence of IP borders on pause duration and on range and form of F0 contour at stressed vowels. First results in this direction allow us to automatically find potential IP borders.

3.1.2. Classification

The next step in the inventory of intonation contours deriving is classification of stylized intonation contours of all the IPs according to communicative types listed in Table 2.

Each communicative class is divided into sub-classes according to the number of AGs, and each sub-class is divided into sub-sub-classes according to the position of the prominent AG in the IP.

Each sub-sub-class is given a name consisting of three parts corresponding respectively to communicative type, number of AGs, and position of prominent AG: X_Y_Z.

The number of IP contours which intonation model can determine is equal to $l\sum_{n=1}^{m} n$, where l is the number of distinguished communicative types, m is the maximum number of AGs in an IP, and n is the position of prominent AG in the IP. Now we distinguish 10 communicative types. The maximum number of AGs in an IP is equal to 7. Then the proposed model can generate 280 different intonation contours.

We continued to investigate stylized F0 contours within each sub-sub-class, namely the direction of F0 movement (falling, rising or narrow). We discovered further subdivision into sub-types for finality (3 sub-types), non-finality (4 sub-types), and exclamation (3 sub-types). Figures 3, 4, and 5 represent averaged variants (regarded as models) of finality, non-finality and exclamation for IPs consisting of two AGs, the second one being the prominent one.

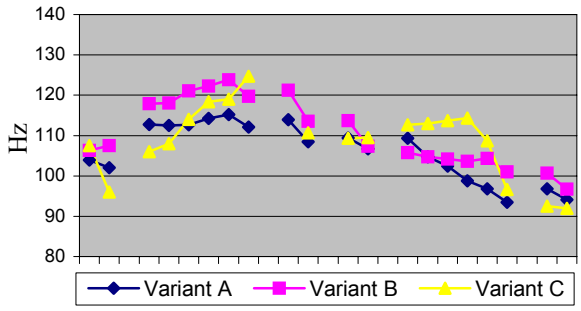

Figure 3: *Models for finality (speaker Valeriy).*

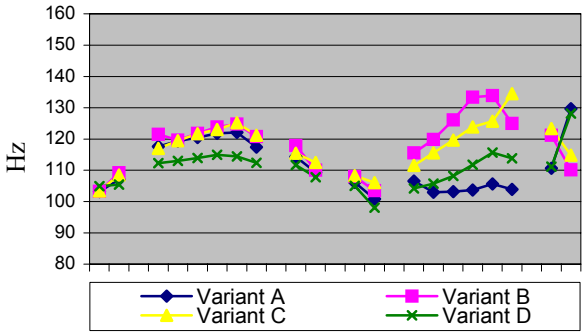

Figure 4: *Models for non-finality (speaker Valeriy).*

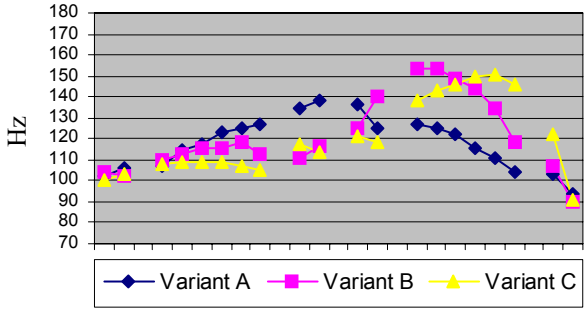

Figure 5: *Models for exclamation (speaker Valeriy)*

It should be added, that the proposed intonation model allows increase in the number of communicative types at the expense of a more detailed specification of communicative sense, e.g. differentiation between a proper question and a specifying question.

4. Individualization of Intonation Model

Synthesis of individualized speech implies the training of intonation model on speaker's data that is elaborating individual inventory of intonation contours differing in F0 range and shape. Training of the model is performed in semi-automatic way based on the annotation of the speaker's database. Intonation peculiarities other than F0 contours should be accounted for as well. In our case it concerns for example the insertion of an extra pause before the last AG of an IP. (This is characteristic of some actor's readings).

First, breaks between IPs are indicated, and then communicative sub-sub-classes are assigned. There is not enough knowledge at present to automate this step because neither the syntax structure nor even punctuation marks are helpful.

Second, the stylization of IPs according to the intonation model is carried out automatically on the basis of speech database annotation which contains the detailed description of F0 movement along each vowel and voiced consonant in the form of a sequence of pitch period lengths. Stylization consists in determining of F0 values at the target intonation points of an AG: two F0 values for pre-nucleus, six for nucleus and two for post-nucleus. Tables (where the rows correspond to IPs in sub-sub-classes, and the columns correspond to target points) and diagrams of intonation contours are obtained and models of intonation contours are derived either by averaging or by selecting one representative contour.

Weak points of the described procedure are: the difficulty of automatic identification of an IP communicative type and prominent AG position; errors of automatic F0 values calculation; and lack of data for several communicative types.

The set of obtained intonation contours forms the individualized intonation model and describes the intonation of a particular speaker.

The challenge and the goal of the future work is to automate the process of breaks between IPs identifying, communicative type of an IP with indication of prominent AG marking, and variants of model contours within a set of IPs with equal communicative type, equal number of AGs and equal position of prominent AG discovering. For example, while speaker Valeriy has four variants of N_2_2 intonation contours (non-final IP with two AGs, the second being the prominent one), the speaker Svyatoslav has only three variants of N_2_2 contour. In Figure 6, averaged N_2_2 intonation contours of two speakers are presented.

We hope to advance in automation through detecting the correspondence between intonation contours on one hand and communicative nuances and lexical-syntactic structure of IPs on the other. For example, it is clear now that a change of a definite intonation contour occurs when the speaker is "uncertain" about what he is reading.

The most important yet difficult goal is to develop procedures for input text analysis with modeling of text interpretation by a particular speaker.

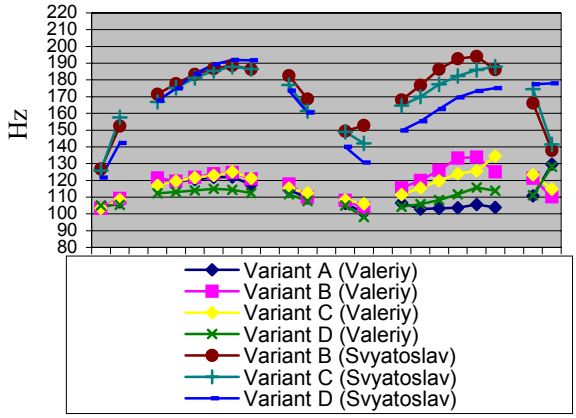

Figure 6: *Models of intonation contours for non-finality of two speakers Svyatoslav and Valeriy.*

5. Implementation in TTS system

Obtained individualized intonation models were used in the TTS system for Ukrainian.

Communicative type is assigned based on punctuation mark and some lexical cues. Sequences of F0 values corresponding to target intonation points, are calculated by the prosody generation module, and are characteristic of the speaker whose voice is used for synthesis. In unit selection module, the fundamental frequency is one of the main criteria of selection. Concatenation of selected phones with definite intonation is performed by the acoustic processor. Phone waves may be either modified according to calculated values of durations and F0 or concatenated as they are, without modification (pure unit selection).

6. Testing the intonation model

To test the intonation model incorporated in the TTS system, a formal listening test was carried out. 22 listeners (students and professors of linguistic university, specialists in Ukrainian language) were asked to listen to 60 synthesized passages containing IPs of 10 communicative types (Table 2). All the passages were taken from the Ukrainian translation of the Lewis Carroll's "Alice in Wonderland", because this text has a natural variety of prosody [1]. Test material was synthesized with Valeriy's voice.

Listeners were supplied with a list of 10 communicative types and the orthographic text corresponding to the synthesized passages, where punctuation marks were absent and all the words were in lower case. The task was to indicate the communicative type of each IP. Each passage of synthesized speech was played three times, and the experiment lasted 30 minutes.

The results are presented in Figure 7. The communicative type recognized the best was enumeration (89 %). This corresponds to the results of our experiment with real Valeriy's speech. Then the listeners noted that this speaker had a distinct intonation contour for enumeration. So, it was not difficult to implement this contour in our TTS system. On the contrary, there is no big difference between contours of finality, expressive finality and exclamation, which is reflected in corresponding recognition results. The poor recognition of "first part of complex wh-question" is due to the fact that some listeners judged the whole questions and recognized them as wh-questions.

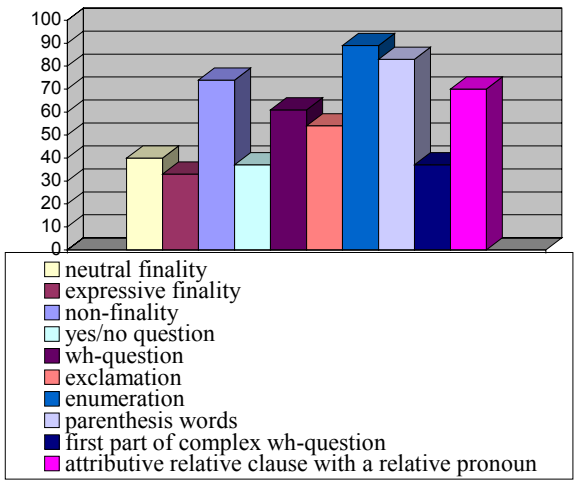

Figure 7: *Results of IPs communicative type recognition (%).*

Table 3 represents recognition results of some IPs not containing lexical cues (in Ukrainian). Communicative type "parenthesis words" was recognized by all listeners in all corresponding IPs.

Table 3. *Results of IPs communicative type recognition*

IPs (English equivalent for commodity)	Imposed communicative type	Recognized as
You'll see me there	expressive finality	expressive finality
Yes	expressive finality	neutral finality
Poor Alice	exclamation	exclamation
there's no use in crying like that	exclamation	expressive finality
Alice felt	non-finality	non-finality
I shall have to ask	non-finality	first part of wh-question
all dripping wet, cross	enumeration	enumeration
White (Rabbit) with pink eyes	enumeration	relative clause
pulling me out of the window	yes/no question	yes/no question
Not like cats	yes/no question	non-finality

7. Discussion

This work revealed the communicative polysemantics of information contained in texts to be read. Readers interpret texts according to situation, to audience and even to their own character. We studied several cases when, for example, some speakers break an IP in AGs and others do not, some add prominence, others do not while reading the same text. Thus, while synthesizing speech we have to model the reading of a text by a specific speaker.

Now the individualized analysis of input texts is not performed to the full extent. Decision about the communicative type of an IP is made based on punctuation marks and some lexical cues. Then a corresponding individual intonation model trained on speaker data is applied to create a speaker "tailored" target specification of an input text.

The attached audio file ("Alice.wav") presents the synthesized beginning of the Ukrainian version of "Alice in Wonderland".

8. Conclusions

The presented work concentrates on the study of intonation in communicative aspect. The full range of communicative types present in a large speech corpus was investigated. The presence of complex sentences allowed to discover specific types associated with parts of questions and attributive relative clauses with relative pronouns.

The derived intonation model based on communicative types and its individualization based on semi-automatic analysis of speaker data were implemented in unit selection TTS system for Ukrainian and tested during a formal listening test. The results testify that the listeners identify the communicative types of synthesized utterances.

The proposed model allows to synthesize speech in different styles (e.g. neutral and expressive) using the same speech database but different intonation contours (e.g. neutral and expressive finality, non-finality, questions, etc.).

It should be noted also that the synthesis technology under the framework of which the proposed intonation model is used, may be applied to languages other than Ukrainian. Similar approach to intonation modeling is used in [5] for Russian and Polish.

9. References

[1] Strom, V., Clark, R., King, S., "Expressive Prosody for Unit-selection Speech Synthesis", Proc. of INTERSPEECH 2006 – ICSLP, pp. 1296-1299.

[2] Lyudovyk, T., Sazhok, M., "Unit Selection Speech Synthesis Using Phonetic-Prosodic Description of Speech Databases", Proc. of International Conf. "Speech and Computer" (SPECOM'2004), St.-Petersburg, Russia, 2004, pp. 594-599.

[3] Lyudovyk, T., "Linguistic Processor Training on Speaker Data for Unit Selection Text-to-Speech", Proc. of International Conf. "Speech and Computer" (SPECOM'2006), St.-Petersburg, Russia, 2006, pp. 315–320.

[4] Taylor, P., "Text-to-Speech Synthesis", Manuscript, http://mi.eng.cam.ac.uk/~pat40/ttsbook_draft_2.pdf.

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

[6] Clark, R. A. J., King, S., "Joint Prosodic and Segmental Unit Selection Speech Synthesis", Proc. of INTERSPEECH 2006 – ICSLP, pp. 1312-1315.

[7] Colotte, V., Beaufort, R., "Linguistic features weighting for a Text-To-Speech system without prosody model", Proc. of INTERSPEECH 2005, pp. 2549-2552.

[8] Mertens, P., "The Prosogram: Semi-Automatic Transcription of Prosody Based on a Tonal Perception Model", Proc. of Speech Prosody 2004, Nara (Japan), 23-26 March, pp. 549-552.