# On a Way to the Computer Aided Speech Intonation Training

Boris Lobanov, Yelena Karnevskaya and Vladimir Zhitko

The United Institute of Informatics Problems of National Academy of Sciences of Belarus, Minsk, Belarus

lobanov@newman.bas-net.by, zhitko.vladimir@gmail.com

**Abstract.** Presented in the paper is a software system designed to train learners in producing a variety of recurring intonation patterns of speech. The system is based on comparing the melodic (tonal) portraits of a reference phrase and a phrase spoken by the learner and involves active learner-system interaction. Since parametric representation of intonation features of the speech signal faces fundamental difficulties, the paper intends to show how these difficulties can be overcome. The main algorithms used in the training system proposed for analyzing and comparing intonation features are considered. A set of reference sentences is given which represents the basic intonation patterns of English speech and their main varieties. The system's interface is presented and the results of the system operation are illustrated.

**Keywords:** Speech intonation, melodic/tonal portrait, intonation analysis, computer system for teaching, intonation training

## 1 Introduction

Intonation plays a significant role in speech communication. It shows the general aim of an utterance and points out its information centre (nucleus) as well as giving prominence to the nonnuclear semantically relevant elements and deaccenting those lacking in novelty or semantic weight; it splits an utterance into phrases (clauses) and intonation-units (groups), each presenting a syntactically organized parcel of information, and integrates these parts into an utterance, distinguishing thereby between more and less closely connected 'chunks' of the speech flow. Intonation is widely recognized as an important aspect of speech that provides both linguistic and sociocultural information. Therefore, prosodic aspects of speech should be explicitly introduced to language learners to help them communicate effectively in a foreign language.

A current linguistic idea is that a foreign accent is more evident and stable in intonation than in segmental sounds. A foreign accent in intonation emerges mainly as a result of prosodic interference, an inevitable 'by-product' of bilingualism and, particularly, under the influence of the prosodic patterns of the learner's native language on those of the target language. Considering the variety of functions of intonation in

adfa, p. 1, 2011.

speech and its potential socio-cultural effects, deviations in this area can lead to serious semantic losses in communication. It is a well-known fact that it is incorrect intonation that is often the cause of the wrong impression a non-native language speaker might produce [1]. Native speakers of American English, e.g., made the following observation concerning the Russian accent in English: "Ask an average American what they think about the Russian accent, and the answer will be as follows: "*Russians don't sound very friendly. I feel like they don't like me at all. I am not sure whether it comes from their language or from their culture?*" (See also: https://www.youtube.com/watch?v=e0MZW3AbzxI). One of the reasons many Russian speakers of English sound unfriendly is the so called "flat" tone associated in American English with the above mentioned negative connotations. Obviously, many Russian speakers fail to capture the language-specific phonetic-phonological features of American/British English intonation and, moreover, are unaware of the drastic socio-cultural effects of the deviations from the prosodic form of an utterance. Helping nonnative learners eliminate such errors presupposes ensuring their familiarity and acquisition of the prosodic patterns of the foreign language being studied.

Accuracy of reproducing the foreign intonation patterns in the process of speaking as well as adequacy of identifying the patterns on the level of perception present considerable difficulty for the learners, particularly related to their ability to control their performance and perception (especially for those who have no ear for music). The lingaphone courses and equipment available at present provide only "a hearing" feedback for intonation accuracy control, which is obviously insufficient.

The present paper is concerned with the progress achieved in developing a computer trainer providing an additional *visual* feedback as well as a *quantitative assessment* of the learners' intonation accuracy in the foreign language teaching process.

In the course of creating the speech intonation training systems we faced a number of difficulties connected with the necessity of solving a number of technical problems, namely:

1. ***An adequate comparison of the pattern signal and a spoken one which is usually characterized by a non-linear time deformation and its beginning and end are not known beforehand.***

The solution of this problem has become possible thanks to the application of the modified method of a ***continuous dynamic time warping*** (CDTW) of two signals, developed by the author earlier [2]. The use of this method ensures automatic recognition of the end and beginning of a phrase being uttered simultaneously with its comparison with the pattern phrase.

2. ***Automatic segmentation of the signal being analyzed into areas for which the notion of F0 is relevant as far as the formation of the tonal contour of the phrase is concerned (the segments of vowels and most of the sonorants).***

This problem is being solved by means of a non-linear transfer of segment markers from the preliminarily marked pattern-phrase onto the phrase being uttered with the help of the author's earlier suggested technology of cloning the prosodic characteristics of speech [3].

3. ***Precise calculation of F0 of the pattern speech signal and of that produced by the learner within a very wide voice range {30 – 1000 Hz}, for male and female voices pooled.***

The task is solved by using the traditional methods of singling F0 out of a speech signal. Seeking a solution to the given problem has been the subject matter of a large number of publications (see e.g. [4]).

4. ***Automatic interpolation of current values F0 on the segments for which measuring F0 is invalid, i.e. on most of the consonants.***

This task is solved by using well-known interpolation mathematical formulas determining the way of finding intermediate values on the basis of an available discrete set of given values.

5. ***An adequate calculation of a similarity measure between the pattern signal and the uttered one under the condition of their differences in duration and F0 voice-ranges.***

This task is solved by using a representation of an intonation curve in the form of a unified melodic portrait (UMP) described below in the next section of the paper. Calculation of the similarity measure of two UMPs is carried out with the help of traditional formulas either by means of calculating a samples correlation coefficient or through determining the vector distance between the curves.

In dealing with these problems, we relied on the results of earlier research in the field of developing automatic intonation assessment systems for computer aided language learning [5-8] as well as the results of our earlier research in the area of speech intonation analysis and synthesis [9-11].

## 2 Intonation stylization model and acoustic database

The present work is a follow up study to the previously introduced model of universal melodic portraits (UMP) of accentual units (AU) for the representation of phrase intonations in text-to-speech synthesis [9]. According to this model, a phrase is represented by one or more AUs (Accent Unit is often referred to as Accent Group). Each unit, in turn, can be composed of one or more words. In the latter case, only one word bears full stress while the other words carry partial stress. Each AU consists of ***pre-nucleus*** (all phonemes preceding the main stressed vowel), ***nucleus*** (the main stressed vowel) and ***post-nucleus*** (all phonemes following the main stressed vowel).

The UMP model assumes that typological features of an AU pitch movement for a particular type of intonation do not depend either on the number or quality of segments in the phonemic content of the pre-nucleus, nucleus or post-nucleus, or on the fundamental frequency range specific for a given speaker. The model allows of representing the intonation constructions of a given language as a set of melodic patterns in normalized space {***Time – Frequency***}. Time normalization is performed by bringing pre-nucleus, nucleus and post-nucleus elements of AU to standard time lengths. This sort of normalization levels out the differences in melodic contours caused by the number of words and phonemes in an AU.

For fundamental frequency normalization $F_{0\,min}$ and $F_{0\,max}$ are determined within the ensemble of melodic contours produced by a certain speaker. This sort of normalization cancels out the differences of melodic contours caused by the speaker's voice register and range.

The normalization is calculated by the formula

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

We note that the value

$$R = [(F0\ max / F0\ min) - 1 \quad (2)$$

expressed in an octave scale, can be used to estimate the range of the F0 change (*wide-medium-narrow*).

In certain cases it may be beneficial to use statistical normalization instead of (1)

$$F_0^N = (F_0 - M) / \zeta \quad (3)$$

where *M* is mathematical expectation, ζ is standard deviation. Note that *M* can be interpreted as a register and ζ – as a range of the speaker's voice.

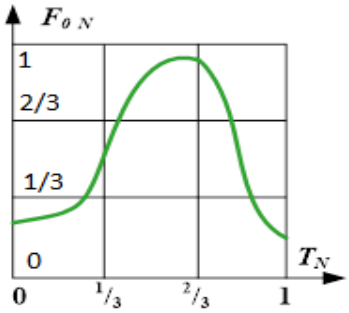


**Fig. 1.** The UMP model

Therefore, the normalized space for UMP may be presented as a rectangle with axes ($T_N$, $F_0^N$) as schematically shown in **Figure 1**, while the interval [0 – 1/3] on the abscese $T_N$ is a pre-nucleus, [1/3 – 2/3] is a nucleus, and [2/3 – 1] is a post- nucleus. The intervals on the ordinate $F_0^N$: [0 -1/3] – low level, [1/3 - 2/3] – mid-level, [2/3 - 1] – high level.

UMP representation focuses on the peculiarities of the shape of the F0 curve on the nucleus with less attention to the quantitative and qualitative composition of the pre- and post-nucleus. Within the framework of the UMP it is possible to describe the melodic curve minutely, using well-known terms, such as:

- *"low-medium-high*" - for the pitch level,
- "*falling-level-rising*" - for the direction of the pitch change,
- "*wide-medium-narrow*" - for the range of the pitch change.

In [10] the positive experience of creating melodic portraits of complex narrative sentences of Russian speech with the use of the PAE model and UMP is described, and in [11] it was shown that the representation of intonation in the form of UMP allows to reveal the characteristic differences when comparing melodic portraits of English and Russian phrases of dialogue speech.

## 3 Acoustic Database

The developed prototype of the system is realized in 2 variants for implementation in multimedia course-books for advanced learners of English [12] and Russian [13] intonation. Application of this system makes it possible for the students not only to listen to phrases pronounced with standard intonation but also observe the model

F0(t) и A0(t) curves on display, reproduce these phrases, compare their F0(t) и A0(t) curves with the original ones and obtain a numerical evaluation of their similarity. Used as models are male-and-female-spoken sample phrases from the above-mentioned multimedia course-books.

In practice of teaching English intonation 10 tonal patterns are used which represent the pitch varieties of the four basic types of pitch change in English (see: Table 1). The principle of selecting the varieties is both structural and functional: on the one hand - perceptible discrimination and identification, and, on the other hand - a tendency towards association with a particular modal-pragmatic type of utterance (statements, general questions, requests, implications, apologies, etc.). In Table 1 the [+] sign indicates the position of the nuclear vowel of the phrase.

**Table 1.** The basic types of English tonal patterns

| Type of tone pattern | N° | Pitch varieties | Types of utterances. Common usage. | Typical examples |
|---|---|---|---|---|
| ***Rising*** | 1 | Mid Wide | General, Elliptical questions, Tags | Is it di+fficult? |
| | 2 | Low Wide | General questions, Tags, Non finality | Can I speak to Ma+ry? |
| | 3 | High Narrow | Interrogative repetitions | Na+tive? |
| | 4 | Low Narrow | Statements, Tags | Ye+sterday. |
| ***Falling*** | 5 | High Wide | Statements, Imperatives, Special questions | Li+sten to me, please! |
| | 6 | Mid Wide | Statements, Imperatives, Tags, Special questions | Whe+re is she? |
| | 7 | Low Narrow | Statements, Imperatives, Tags | It's in the So+uth. |
| ***Falling-Rising*** | 8 | Undivided | Imperatives, Questions, Statements, Non-finality, Conversational formulas | They are re+ady. |
| | 9 | Divided | | No+t no+w. |
| ***Rising-Falling*** | 10 | Undivided | Statements, Special questions | It's wo+nderful. |

The acoustic signals realizing each of the given phrases are marked for the boundaries of each of the vowels contained as well as for indicating the functional status of the vowel: pre-nucleus, nucleus, post-nucleus of an accentual unit. In the database used, there are 4 to 5 commonly used samples for each of the ***10 tonal patterns*** of the phrases, as well as several samples of conversational speech and a piece of narrative prose. In addition to the most commonly used samples, the database includes examples of different ***types of utterances*** (see Table 1) for each of the 10 tonal patterns, read by a professional British English speaker.

As far as computer training is concerned, we proceed from the model of ***intonation patterns*** (IP) by Elena Bryzgunova [14], which is widely used in teaching Russian

speech intonation. This model includes seven patterns: IP1 (the falling tone), IP2 (the falling tone with some prosodic emphasis), IP3 (the rising tone with a subsequent fall), IP4 (the falling-rising tone). IP5 (combination of the rising, level and falling tones), IP6 (combination of the rising and level tones), IP7 (combination of the rising tone with a glottal stop).

In the database used, there are 5 samples of common ***types of utterances*** for each ***intonation pattern*** as well as several samples of conversational speech and a piece of narrative prose.

## 4 Block diagram of the intonation training system

Figure 2 contains a block diagram illustrating a sequence of algorithms for the analysis and evaluation of speech intonation within the computer system developed. The main goal of the system is to provide a student with a compact and easily interpretable pitch image of the results obtained in the course of analyzing the pitch and energy contours of the phrases carrying different intonation patterns. The system will also provide an auditory, visual and numerical evaluation of a student's performance in the intonation of a foreign language.

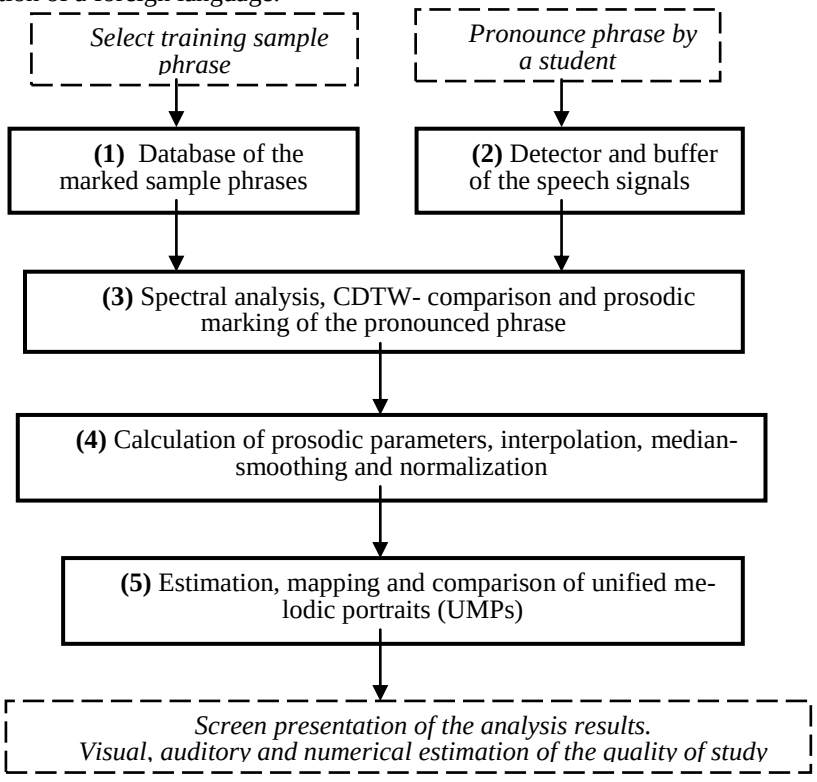


**Fig. 2.** Block diagram of the computer intonation training system

Block 1 contains the database of sample phrases (teacher's phrases) with different intonation patterns, compiled from multimedia course-books (see, e.g. [12] for the English language, or [13] - for Russian). Every sample phrase is preliminarily marked for the perceptible prosodic phrase boundaries and the location of its nucleus (*see figure 3*).

Depending on the concrete goal of intonation training, the student selects the sample phrase needed, listens to it and pronounces it. The student's phrase is recorded on the buffer in block 2.

In block 3, the signals from both the sample and the student-spoken phrase are spectrum analyzed and compared using the algorithm of continuous dynamic time warping (CDTW). This is accompanied by the transfer of prosodic marks and labeling of a pronounced phrase (*see figure 3*).

In block 4, prosodic phrase parameters, such as the fundamental frequency of the tone F0 and energy of the signal A0 are calculated. These parameters are further interpolated on the non-vocal areas, median-smoothed and normalized (*see figure 4*).

In block 5, an estimation and comparison of F0 trajectories first in real time space (see figure 5 and 7) and then in normalized UMP-space (see figure 6 and 8) are produced.

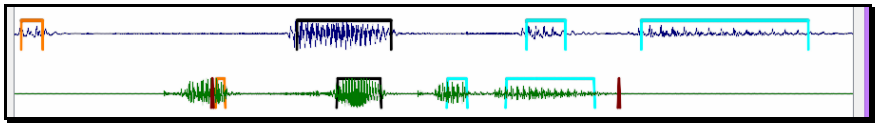

**Fig. 3.** Illustration of speech signals marking: the phrase "*It's Saturday*" pronounced by the teacher (above) and by a learner (below)

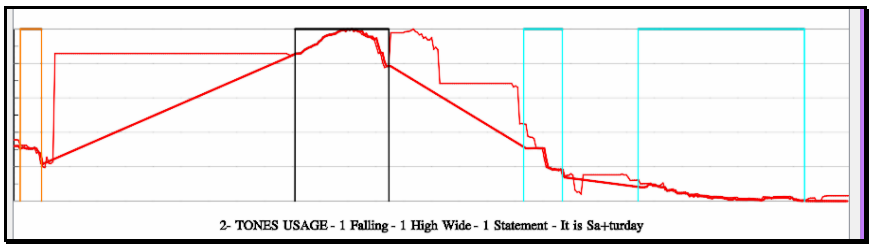


**Fig. 4.** Illustration of F0 trajectory processing for the teacher's phrase *"It's Saturday*": original (*light curve line*) and interpolated, median-smoothed and normalized (*dark curve line)* tracks

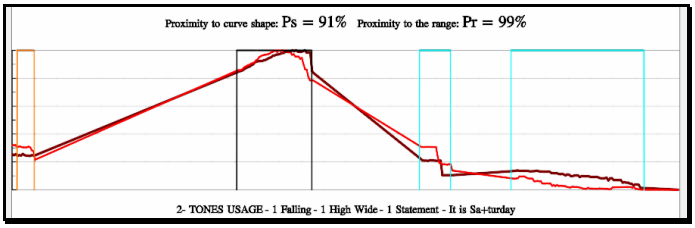


**Fig. 5.** Illustration of F0 curve comparison in real time space between the teacher's (*light curve line*) and a student's (*dark curve line*) phrase *"It's Saturday*" ***(correct pronunciation)***

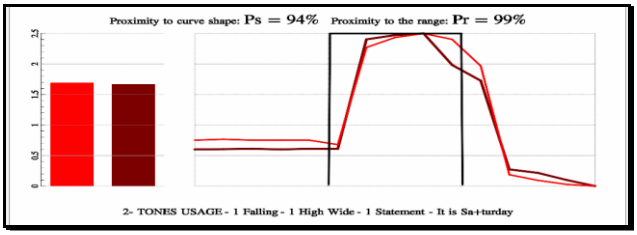


**Fig. 6.** Illustration of F0 range and curve comparison in UMP-space between the teacher's (*left column, light curve line*) and a student's (*right column, dark curve line*) phrase *"It's Saturday*" ***(correct pronunciation)***

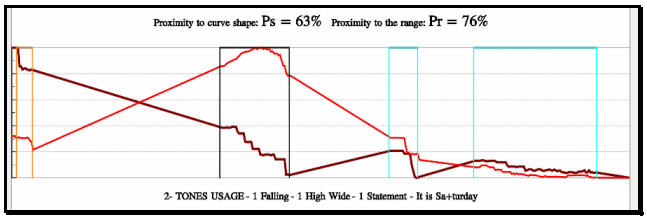


**Fig. 7.** Illustration of F0 curve comparison in real time space between the teacher's (*light curve line*) and a student's (*dark curve line*) phrase *"It's Saturday*" ***(wrong pronunciation)***

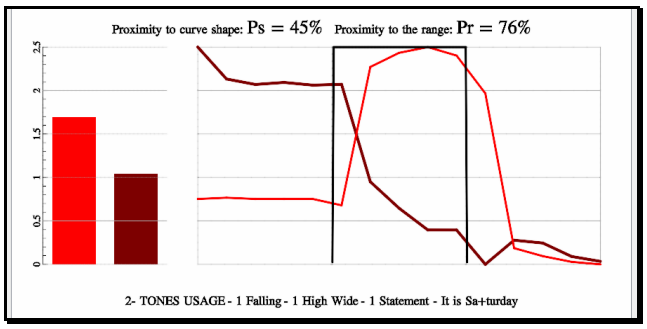


**Fig. 8.** Illustration of F0 range and curve comparison in UMP-space between the teacher's (*left column, light curve line*) and a student's (*right column, dark curve line*) phrase *"It's Saturday*" ***(wrong pronunciation)***

At the top of Figures 5-8, numerical estimates are presented as a percentage of the proximity of the teacher's and student's phrases: for the F0 shape of curves – (Ps) curves and for their ranges – (Pr). The measure of proximity is defined as the vector distance between them.

On the left side of Figures 6 and 8, light and dark columns are shown. They expresses the values R in an octave scale calculated by using formula (2). Value R is used to show the difference range of F0 change between the teacher's and student's phrases.

## 5 Software realization of the system

Software realization of the system named ***"IntonTrainer"*** is written on C++ programming code by using Qt framework. It can be compiled under Windows platform (from XP to 10 versions), as well as under Linux platform.

The application core is divided into several modules that implement standalone functions. Such modules can implement audio signal recording, voice detection, CDTW processing, etc. As these modules are independent from each other, we can easily build different applications by substituting these modules for other ones or integrating them in external systems.

For building the main user interface a built-in web engine is used. The user interface is built on html5, css3 and js (ReactJs js frame-work). The "Developer mode" user interface is built on standard Qt forms.

The main user interface is independent from the application core and can be modified or even replaced by another one. The use of html/css/js standard allows an easy change of application front-end for different purposes. For the interaction with application core there exist a number of special links formats processed by application core. Such links can open different applications dialogues (like settings, developer mode and so on), process input audio signals and play audio files.

Thus we can easily build in different training systems by replacing the front-end and training data files.

The starting page of the User Interface is shown in figure 9.

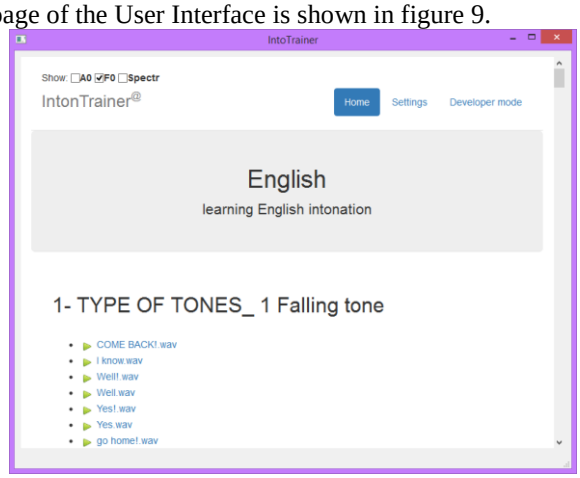


**Fig. 9.** Starting page of the User Interface

## 6 Conclusions

At present, using the IntonTrainer system, experiments are conducted to learn by students the intonation of Russian and English. Preliminary results indicate a significant effectiveness of its use.

To date, there are demo versions of the "IntonTrainer" system, focused on learning the intonation of Russian and English. For those who want to test the system, a site https://intontrainer.by/ is open.

A working version of the prototype system will be demonstrated to the conference participants.

## References


1. Chun, D. M.: The neglected role of intonation in communicative competence and proficiency. Modern Language Journal, 72, (1988), pp. 295-303.
2. Lobanov B.M., Levkovskaya T.V.: Continuous Speech Recognizer for Aircraft Application // Proceedings of the 2nd International Workshop "Speech and Computer" – SPECOM'97 - Cluj-Napoca, (1997), pp. 97-102.
3. Lobanov, B.M., Tsirulnik L.I., Sizonov O.N.: «IntoClonator» – Computer system of cloning prosodic characteristics of speech (in Russian) // Proceedings of the International Conference "Dialogue 2008", Moscow, (2008), pp. 330-338.
4. Shimamura T. and Kobayashi H.: Weighted Autocorrelation for Pitch Extraction of Noisy Speech // IEEE Transactions on Speech and AudioProcessing, Vol. 9, (2001), pp. 727–730
5. Anne Bonneau, Vincent Colotte.: Automatic Feedback for L2 Prosody Learning. Speech and Language Technologies, (2011), pp.55-70.
6. Yi Xu: ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In Proc. of the TRASP'2013, Aix-en-Provence, France (2013), pp. 7-10.
7. Juan Arias, Nestor Yoma, Hiram Vivanco.: Automatic intonation assessment for computer aided language learning. Speech Communication 52 (2010), pp. 254–267.
8. Dávid Sztahó, Gábor Kiss, László Czap, Klára Vicsi.: A Computer-Assisted Prosody Pronunciation Teaching System. In Proc. of the WOCCI 2014, Singapore, (2014), pp. 121-124.
9. Lobanov B.M. et al: Language- and speaker specific implementation of intonation contours in multilingual TTS synthesis // Speech Prosody: Proceedings of the 3-rd International conference, Dresden, Germany, (2006), pp. 553-556.
10. Lobanov B., Okrut T.: Universal Melodic Portraits of Intonation Patterns of Russian Speech (in Russian) // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 13 (20). — Moscow, (2014). pp. 330-339.
11. Lobanov B.M.: Comparison of Melodic Portraits of English and Russian Dialogic Phrases // Computational Linguistics and Intellectual Technologies: Papers from the Annual Inernational Conference "Dialogue". Issue 15 (22). – Moscow, (2016), pp. 382-392.
12. Karnevskaya, E.B. (editor): Practical English Phonetics. Advanced course // Aversev, Minsk, (2016), – 409 p.
13. *Odintsova I.V.: Sounds. Rhythm. Intonation (in Russian)* — Flinta-Nauka, Moscow, (2011), - 378 p.
14. Bryzgunova E.A.: Sounds and Intonation of Russian Speech (in Russian) — Nauka, Moscow, (1968), - 267 p.