**УДК 378.16**

# LINGUISTIC AND ACOUSTIC RESOURCES OF THE COMPUTER-BASED SYSTEM FOR ANALYSIS AND INTERPRETATION OF SPEECH INTONATION

Zdaranok Y.A. (*yuliyazdaranok@gmail.com*)
Belarussian National Technical University, Minsk, Belarus

This paper considers suprasegmental parameters such as intonation, stress and speech rhythm. Correct articulation of sounds and correct pronunciation in the target language depend on prosodic structures presented by standard intonation patterns. Linguistic and acoustic resources provide the basis for computer-aided intonation training in and outside the classroom. There is a great need for new and innovative computer-aided programs. But there is no necessary software available for such teaching system and, therefore, such system appears to be of great relevance. Presented work describes as well the algorithm of designing of the computer-based system for analysis and interpretation of speech intonation.

**Keywords:** suprasegmental parameters, intonation, pronunciation, linguistic and acoustic resources, computer pronunciation training

## 1 Introduction

Intonation plays a significant role during the speech comprehension. Speech intonation shows a communicative intention of an utterance, its logical meaning, a prominence of the most significant theme in relation to general themes (actual division of a sentence), a distinction between semantically associated segments of speech, and an integration of speech elements within these segments. [Lobanov B., 2006].

Proper pronunciation of the target language is associated with the correct articulation of sounds and also with suprasegmental parameters. Suprasegmental parameters are aspects of speech, referred to prosody. By teaching prosody is important to understand and describe the suprasegmental parameters that are detected in the target language. It is also very important to describe the prosodic pattern of speech.

By learning prosody is necessary to know an intonation contours palette to convey the diversity of thoughts in speech. Therefore, the intonation should be taught in the context of a well-structured dialogue or discourse.

Speech is a universal means of communication. It includes the processes of generation and perception (reception and analysis) messages for communication purposes in all languages of the world where leading thought or mental image is implemented in the speech by acoustic instruments.

The sentence is a combination of grammatically and phonetically structured performance of human thought during the speech. It is known that the sentence possesses definite phonetic features: speech melody, sentence-stress, tempo, rhythm, pauses and timbre. Each feature performs a definite task, and all of them work simultaneously [Lobanov B., 20116]. An utterance consists of one or more phrases. The phrase has a semantic completeness and syntactic structure. The phrase is the largest unit with a complete phonetic intonation. The main distinguished unit in the phrase is the core, accomplished by pre-nucleus and post-nucleus elements.

## 2 Intonation pattern of melodic portrait

The present work is a follow up study to the previously introduced model of universal melodic portraits (UMP) of accentual units (AU) for representation of phrase intonations in text -to-speech synthesis [Lobanov B., Okut T., 2014]. According to this model, a phrase is represented by one or more of AUs. Each unit, in turn, can be composed of one or more phonetic word. If there is more than one word in an AU, than only one word bears the main stress while other words carry a 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 stressed vowel).

The UMP model assumes that topological features of melodic AU for particular type of intonation do not depend on a number or quality of phonemic content of a pre-nucleus, nucleus or post-nucleus, nor on the fundamental frequency range specific for a given speaker.

The UMP model allows to represent intonation constructs 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 F0 min and F0 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 speaker’s voice register and diapason.

The normalization is calculated by the formula

$$F_0^N = \left(F_0 - F_{0_{min}}\right) / \left(F_{0_{max}} - F_{0_{min}}\right) \qquad (1)$$

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

$$F_0^N = (F_0 - M)/\varsigma \qquad (2)$$

where M is mathematical expectation, ζ is standard deviation. Note that M can be interpreted as a register and ζ – as a diapason of speaker's voice. Therefore, the normalized space for UMP may be presented as a rectangle with axes (*TN*, *F0*$^N$) as schematically shown in Figure 1, while the interval [0 – 1/3] on the abscess *TN* Structure of linguistic resource is a pre-nucleus, [1/3 – 2/3] is a nucleus, and [2/3 – 1] is a post- nucleus. The intervals on the ordinate *F0*$^N$: [0 -1/3] – low level, [1/3 – 2/3] – mid-level, [2/3 – 1] – high level.

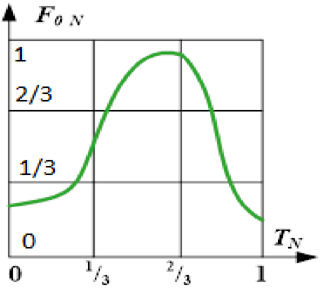

Figure 1 – Main view UMP IK

## 3 Structure of linguistic and acoustic resources

In present research we use the resources of English texts and audio-files [Ockenden, 2005], which include:

1. 44 everyday situations, each containing four dialogues in authentic conversational English;
2. All dialogues consist of 1051 sentences, including 704 affirmative sentences, 325 interrogative sentences and 22 exclamatory sentences, spoken by certain number of male and female speakers;
3. Situations relevant to those, who are studying or travelling in England, including eating out, entertainment and travel, as well as more general functions such as greetings, complaining and apologizing.

Each dialogue is structured by:

the speaker (man and/or woman),

the number of participants in the dialogue,

the type of sentence: questions, statements and exclamations,

the number of phrase units,

number of AU in the phrase,

the specification pre-nucleus, nucleus, post-nucleus.

In his “Advice to Foreign Learners” A.C.Gimson emphasizes necessity of learning “the English usage of falls and rises to signify the mood of the speaker, so that an over-use of rises will not give an unintentional impression of, for example, diffidence or complaint, and too many falls create an unwitting effect of impolite assertiveness” [Grimson, 1996].

For that reason, the processing of acoustic materials was conducted according to the following intonation criteria:

the falling tune;

the rising tune;

the falling-rising tune.

(a) The falling tune

The voice falls from a high to a low note on one stressed syllable. It is used in short complete statements, for questions beginning with a question word, for question tags when the speaker is sure that what he says is right or for orders and exclamations.

(b) The rising tune

The voice rises on the last stressed word or on the unstressed syllables following the last stress. It is used for statements intended to encourage, for questions which are answered by, for questions beginning with question words when the speaker wishes to show some special interest, for question tags when the speaker is not sure that what he says is correct, for sentences ending with “please”; for “goodbye”; for “thank you” when it is used to show gratitude for a simple matter (passing the salt etc.)

(c) The falling-rising tune

The voice falls on the most important part of the sentence and rises again. It is used for apologies, for expressing tentative opinions.

According to the grammar rules of English there are two types of commonly used interrogative sentences: general and special questions. The statistics below show how often are interrogative sentences used depending on the situational moment during the interaction.

Table 1 – Statistics: General question

| *Starts with* | *Number of questions* | *Starts with* | *Number of questions* |
|---|---|---|---|
| Is | 16 | Need | 1 |
| Are | 12 | Has | 4 |
| Am | 1 | Can | 23 |
| Was | 2 | Could | 17 |
| Were | 1 | Shall | 2 |
| Will | 6 | May | 2 |
| Do | 26 | Must | 1 |
| Does | 4 | Would | 16 |

| | | | |
|---|---|---|---|
| Did | 2 | TOTAL | 140 |
| Have | 21 | | |

Table 2 – Statistics: Special question

| ***Question word*** | ***Number of questions*** | ***Question word*** | ***Number of questions*** |
|---|---|---|---|
| What | 66 | Whose | 0 |
| When | 13 | Wherefore | 0 |
| Why | 2 | Whatever | 0 |
| Where | 14 | Wherewith | 0 |
| Who | 2 | Whither | 0 |
| How | 52 | Whence | 0 |
| Which | 8 | However | 0 |
| Whom | 0 | TOTAL | 157 |

Table 3 (English) and Table 4 (Russian) gives us a clear idea that the minimum (F0 min) and maximum (F0 max) of fundamental frequency - F0 differs in the entire ensemble of intonation patterns IKi [Bryzgunova, 1982] in the utterance spoken by native English speaker and native Russian speaker. That makes obvious that the voice of pitch in English is higher than the voice of pitch in Russian.

Table 3 – Shows the minimum (F0 min) and maximum (F0 max) of fundamental frequency in English

| Intonation type | Affirmative | | Special | | General | |
|---|---|---|---|---|---|---|
| F0 [Hz] | min | max | min | max | min | max |
| Sample 1 | 92 | 184 | 100 | 330 | 109 | 280 |
| Sample 2 | 90 | 180 | 98 | 280 | 98 | 286 |
| Sample 3 | 100 | 230 | 60 | 235 | 101 | 252 |
| Sample 4 | 105 | 230 | 65 | 232 | 99 | 211 |
| Mean value | 96.75 | 206 | 80.75 | 268.25 | 101.75 | 257.25 |
| Diapason | 2.13 | | 3.34 | | 2.53 | |

Table 4 – Shows the minimum (F0 min) and maximum (F0 max) of fundamental frequency in Russian

| Intonation type | Affirmative | | Special | | General | |
|---|---|---|---|---|---|---|
| F0 [Hz] | min | max | min | max | min | max |
| Sample 1 | 80 | 147 | 85 | 154 | 85 | 170 |
| Sample 2 | 78 | 150 | 85 | 155 | 91 | 196 |
| Sample 3 | 81 | 144 | 80 | 155 | 84 | 185 |
| Sample 4 | 82 | 146 | 83 | 157 | 84 | 185 |

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| Mean value | 80.25 | 146.75 | 83.25 | 157.5 | 86.25 | 182.25 |
| Diapason | 1.83 |  | 1.89 |  | 2.11 |  |

## 4 An algorithm of computer system for speech intonation training

Computers were used for language learning since 1960 in the last century. Nowadays, however, the importance of emphasizing intonational aspects of speech while teaching foreign language and while creating computer-based systems for speech analysis and synthesis has the same significance as 50 years ago.

In the United Institute of Informatics Problems NAS Belarus, Minsk under the guidance of B.M. Lobanov were created a Prototype of the Computer System for Speech Intonation Training [Lobanov B., Zhitko V, 2017]. The development such of an intelligent automatic system for the processing of speech signals needs a strict algorithm of intonation analysis in a composition of the computer teaching system of foreign speech [Lobanov B., Zhitko V, Zdaranok Y., 2016].

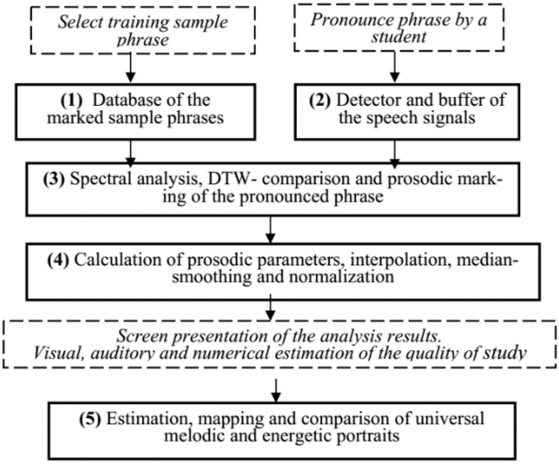
Select training sample phrase
Pronounce phrase by a student
(1) Database of the marked sample phrases
(2) Detector and buffer of the speech signals
(3) Spectral analysis, DTW- comparison and prosodic marking of the pronounced phrase
(4) Calculation of prosodic parameters, interpolation, median-smoothing and normalization
Screen presentation of the analysis results.
Visual, auditory and numerical estimation of the quality of study
(5) Estimation, mapping and comparison of universal melodic and energetic portraits

Figure 2 – The block diagram of the computer training system of foreign speech intonation

Figure 2 contains a block diagram illustrating the sequence of algorithms for the analysis and interpretation of speech intonation within the developed computer teaching system. The main goal of the system is to provide a student with a compact and easily interpretable image for the results of analysis of melodic and energy contours of phrases with different into-nation. The system would also provide a visual, auditory and numerical evaluation of the quality of learning of a foreign speech intonation by a student.

The diagram of the computer teaching system of foreign speech intonation Block 1 contains the database of sample phrases with different intonation patterns which is compiled from multimedia textbooks. Every sample phrase has the preliminary placed prosodic marks that include phrase boundaries and placement of its nucleus.  Based on a given goal of intonation learning, a student chooses the needed sample phrase, defines it and pronounces it. The pronounced phrase is recorded on the buffer (block 2). In block 3, the signals from the both a sample and pronounced phrases are spectrum analyzed and compared using the method of dynamic time warping (DTW). This is accompanied by a transfer of prosodic marks and labeling of a pronounced phrase. In block 4, prosodic phrase parameters, such as frequency of the basic tone F0 [Lobanov B., Levkovskaya T., 1997] and energy of the signal A0 are calculated. These parameters are further interpolated on the non-vocal areas, median-smoothed and normalized. In block 5, an estimation and comparison of universal melodic and energetic portraits are produced. Figure 3 presents some illustration of system's output for the interrogative phrase: «Did Sasha eat the porridge»? The image shows successive processing F0 (t) and A0 (t) and a comparison of the sample phrase and the student-spoken phrase speech signals.

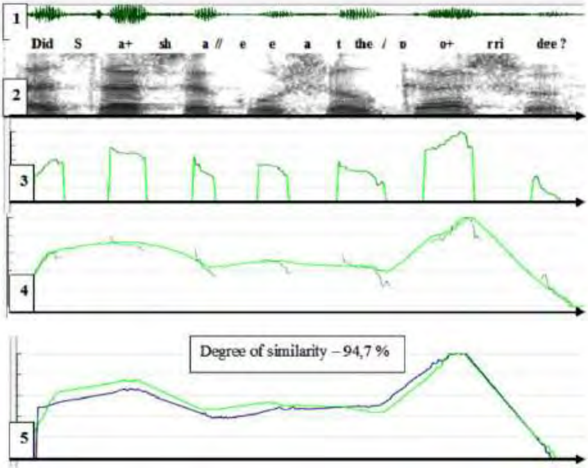

Figure 3 – The illustration of speech signals processing: 1- oscillogram, 2 – spectrum, 3 – F0(t) (original), 4 – F0(t) (after interpolation and median smoothing), 5 – comparison of two melodic curves F0 (t) - sample and spoken phrases

Application of an intonation mapping analyzer as a part of a speech recognition system is expected to increase reliability of recognition through the prominence of accented words and intonational segmentation of a speech flow. Intonation analysis will also be helpful for subsystems of identification of individual and emotional factors of speaker’s speech. The use of intonation system in speech synthesis systems will give an opportunity to improve the intonational prominence of the synthesized speech so that it will positively affect listener’s comprehension.

## 5 Conclusion

Practically, to know a foreign language means to generate the skills and develop the ability to think as a native speaker and to understand other people’s thoughts. In order to sound right and intelligible to the listener the utterance should be conveyed into the correct intonational pattern. This means that the internal or external performance of speech should be presented with an appropriate dynamic acoustic connotation in accordance to the rules of the target language.

Computer-aided intonation training is specifically designed to evaluate and improve the pronunciation in foreign languages.

Due to computer-aided intonation training system the specific pronunciation mistakes will be identified at the word or subword level, providing an opportunity to improve pronunciation in and outside the classroom according to visual feedback.

There is a great potential on both domestic and international markets for a new and innovative product such as the proposed computer system for intonation training integrated into a foreign language educational courseware. There is no necessary software available for such teaching system and, therefore, such system appears to be of great relevance. In presented work, the linguistic and acoustic resources and the design and the output of the computer-based system for analysis and interpretation of speech intonation were described.

## References

**[Lobanov B., 2006]** Language and speaker specific implementation of intonation contours in multilingual TTS synthesis / B. Lobanov [et al.] // Speech Prosody: proceedings of the 3-rd International conference, Dresden, Germany, May 2–5, 2006. Dresden, 2006. Vol. 2. P. 553–556.

**[Lobanov B., 2016]** Lobanov B. Comparison of Melodic Portraits of English and Russian Dialogic Phrases // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». Iss. 15 (22). M.: RSHU, 2016. P. 382–392.

**[Lobanov B., Okut T., 2014]** Lobanov B., Okut T. Universal Melodic Portraits of Intonation Patterns of Russian Speech // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». Iss. 13 (20). M.: RSHU, 2014. P. 330–339.

**[Ockenden, 2005]** Ockenden M. Situational Dialogues // The English Centre, Eastbourne; Revised Edition. Longman, 2005.

**[Grimson, 1996]** Gimson A.C. Inoduction to the Pronunciation of Eglish, London, 1966, p. 261.

**[Bryzgunova, 1982]** Bryzgunova E. Intonation // Russian grammar // Science. M., 1982. P. 96–122.

**[Lobanov B., Zhitko V, 2017]** Lobanov B. A Prototype of the Computer System for Speech Intonation Training / B.M. Lobanov, V.A. Zhitko // Open Semantic Technologies for Intelligent Systems : Papers from the Annual International Conference. Vol.1 (Minsk, 16-18 February 2017). / edited by. : V. V. Golenkov. — Minsk : BSUIR, 2017. — P. 163-166.

**[Lobanov B., Zhitko V, Zdaranok Y., 2016]** Lobanov, B.M. Computer-based System of Analysis and Interpretation of Speech Intonation / B.M. Lobanov, V.A. Zhitko, Y.A. Zdaranok // International Congress on Computer Science: Information Systems and Technologies / BSU ; edited by. S.V. Ablameiko. — Minsk : BSU, 2016. — P. 589-594.

**[Lobanov B., Levkovskaya T., 1997]** Boris Lobanov, Tatiana Levkovskaya (1997). “Continuous Speech Recognizer for Aircraft Application “Proc. of International Conference - Speech and Computer - SPECOM’97, Napoca, Romania, pp. 817-820.

# ЛИНГВИСТИЧЕСКИЕ И АКУСТИЧЕСКИЕ РЕСУРСЫ КОМПЬЮТЕРНОЙ СИСТЕМЫ ДЛЯ АНАЛИЗА И ИНТЕРПРЕТАЦИИ РЕЧЕВОЙ ИНТОНАЦИИ

Здаранок Ю.А. (*yuliyazdaranok@gmail.com*)
Белорусский национальный технический университет,
Минск, Республика Беларусь

В данной статье рассматриваются супрасегментарные параметры, такие как интонация, стресс и речевой ритм. Правильная артикуляция звуков и правильное произношение на целевом языке зависят от просодических структур, представленных стандартными образцами интонации. Лингвистические и акустические ресурсы служат основой для обучения интонации с помощью компьютера в классе и вне его. Существует огромная потребность в новых и инновационных программах для автоматизации. Но для такой системы обучения нет необходимого программного обеспечения, и поэтому такая система, по-видимому, имеет большое значение. Представленная работа описывает также алгоритм проектирования компьютерной системы для анализа и интерпретации речевой интонации.

**Ключевые слова:** супрасегментные параметры, интонация, произношение, лингвистические и акустические ресурсы, обучение компьютерному произношению.