An Intelligent System of Speech Intonation Analysis and Training

Boris Lobanov
The United Institute of Informatics Problems of National Academy of Sciences of Belarus
Minsk, Belarus
lobanov@newman.bas-net.by

Vladimir Zhitko
The United Institute of Informatics Problems of National Academy of Sciences of Belarus
Minsk, Belarus
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. 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 Russian, British English, American English, German and Chinese speech and their main varieties.**

***Keywords*—Speech intonation, melodic/tonal portrait, intonation analysis, intonation training, computer system for language learning (CALL), semantic of intonation.**

I. Introduction

Intonation plays a significant role in speech communication. It shows the general aim of an utterance and points out its information center (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 socio-cultural 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 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]. 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 system of speech intonation analysis and training providing an additional visual feedback as well as a quantitative assessment of the learners' intonation accuracy in the foreign language teaching process.

II. Basic problems to be solve

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].

Multi-lingual intelligent system of speech intonation analysis and training is presented here as a software package called “IntonTrainer”. The software package “IntonTrainer” (hereinafter, “Application”) is intended for analysis and representation on the screen of the pattern and spoken phrases intonation (F0 - basic tone trajectory), as well as for their comparison and estimation of intonation similarity. Estimation of intonation similarity is carried out on the basis of representation of intonation in the form of universal melodic portraits (UMP) [10].

III. Graphical User Interface (on example of British English learning)

The initial **Application** window that opens after the program is started is shown in “Fig. 1”.

Before you start, you can preview the **Application**: settings (top right corner of the initial window) and correct them. In this window, the user can select the recording type of the signal from the microphone: (1) recording for N seconds, (2) manual control, (3) automatic, or (4) recording for N seconds + template length. In the 4th mode the choice of N = 1s is recommended. In addition, it is possible to specify the number of recorded phrases (files) stored in the **Records** folder. The “About” button opens a window with information about the developers.

After clicking the “Start” button, the main window opens, containing a structured list of reference phrases (“Fig. 2”)

Figure 1. The initial window of the Application.

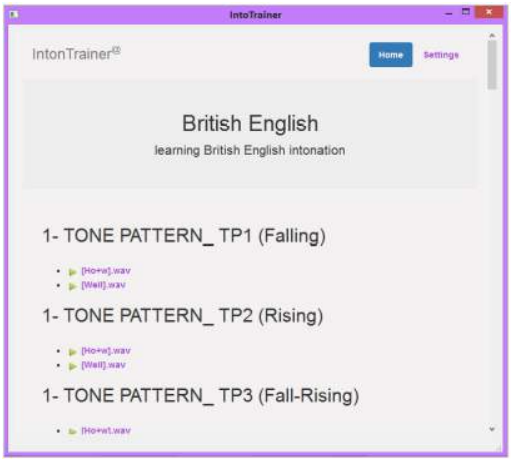

Figure 2. Main window.

IV. Study of intonation constructions of speech

By scrolling the page of the window from top to bottom, the user is given the opportunity to see the examples for the main tone patterns (TP1-TP3) of English speech. Each example provides audio and visual representation of UMP, pairwise comparison of different TPs, explains peculiarities of TP usage, as well as TP implementation in dialogues, prose and verse.

For example, clicking the mouse cursor at directory:

{**TONE PATTERN_ (TP1) (Falling) [We+ll].wav**},

will be open the “Graph” window in which the results of the intonation analysis of this phrase are displayed graphically (“Fig. 3”).

In “Fig. 3” the red column on the left shows the range of the melody change, i.e. frequency of the pitch (F0), expressed in octaves. On the right, a linear graph of the UMP is displayed in red, the core of which is marked with frequent vertical lines. Below the graphs, the minimum and maximum values of F0 for the selected phrase are listed, as well as the text of the phrase in which the nuclear vowel is indicated by the “+” sign. Listening to the selected phrase is carried out by pressing

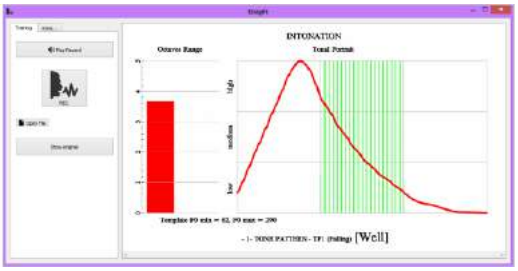

Figure 3. Analysis results window.

the **"Play Record"** button.

When the **"Show original"** button is pressed, an additional window opens ("Fig. 4"), at the top of which a waveform of the phrase signal with pre-nucleus marks (red line), nucleus (black line) and post-nucleus (blue line) is displayed. In the middle part of the window the real curve of the change F0 is depicted showing the parts of the pre-nucleus, the core and the behind-nucleus, from which the UMP is formed, shown in "Fig. 3".

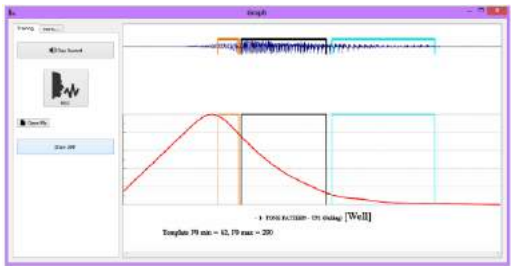

Figure 4. Additional window displaying F0 in real time.

Looking through the structured list of reference phrases in the manner described above (see "Fig. 2"), the user is introduced to and studies main intonation patterns (TP1 – TP3) of English speech, their pairing, proper of use and implementation in dialogue, prose and verse.

V. Individual intonation training

When using the Application for individual intonation training in the study of Russian as a foreign language, as well as for improving oral-speech intonation skills in such professions as call center operators, radio speakers, etc., the user must use an external or built-in microphone. In this case, the user should press the **"Rec"** button, wait for a short "beep-signal" and pronounce the phrase in the microphone, the text of which is indicated in the lower part of the window in "Fig. 3". After recording to the **"RECORDS"** folder and processing of the entered speech signal, the user will hear the 2nd "beep-signal", and the image in the graphics window ("Fig. 3") will be replaced with the image shown in "Fig. 5". In the upper part of the window the results of comparison of the reference and pronounced phrases are shown: Pr - proximity in % on the variation range F0 and Ps - proximity in % in the form of the trajectory F0.

In "Fig. 5" the red column on the left shows the range of change F0 of the reference phrase, and the brown one - the spoken phrase. On the right, the linear graph of the UMP of the reference phrase is displayed in red, and the brown one of the spoken phrase. Below the graphs, the minimum and maximum values of F0 of the reference and spoken phrases are given.

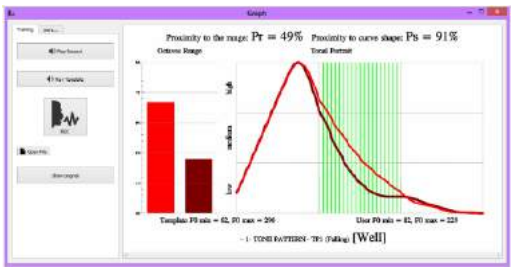

Figure 5. The window displaying the results of analysis and comparison.

Listening to the selected reference phrase is carried out by pressing the **"Play Template"** button and the pronounced phrase **"Play Record"**.

Pressing the "Show original" button opens an additional window ("Fig. 6"), at the top of which waveforms of signals of both phrases are displayed with labels of pre-nucleus (red lines), nucleus (black lines) and post-nucleus (blue lines). In the middle part of the window, the trajectories of the change F0 of the reference (red) and the spoken (brown) phrases in real time are depicted showing the sections of the pre-nucleus, nucleus and post-nucleus, from which the UMPs shown in "Fig. 5". The information contained in this window can be useful for controlling the correctness of the transfer of pre-nucleus, nucleus, and post-nucleus labels from the reference phrase to the spoken phrase. Errors in the transfer of labels can greatly distort the actual form of the UMP phrase.

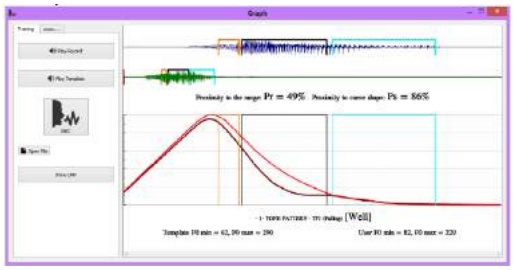

Figure 6. Additional window displaying F0 in real time.

VI. Additional options for using the Application

The software package "IntonTrainer" is an open system and allows its various modifications. First of all, this refers to the used set of reference data, which can optionally be supplemented or formed anew in accordance with the task at hand. For example, if you configure the Application DB for the task of learning the intonation of American English.

An important factor in the formation of the acoustic database of the studied phrases is their prosodic marking on the areas (regions) of pre-nuclear, nuclei and post-nuclei. Currently, this operation is performed manually using the standard application **“Sound Forge”**, but in the future it will be automated. The speech signal of the phrase is recorded in a “wav” format with a sampling of 8 kHz, 16 bits and is labeled into regions P1 (pre-nuclear), N1 (nuclear) T1 (per-nuclear) as shown in “Fig. 7” for a single-nuclear (one-accented) phrase: **“We+ll”**, pronounced by a male voice.

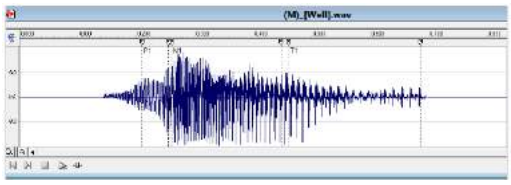

Figure 7. Example of markup of one-accented reference phrase.

Note that if the phrase is one accent, then the index is assigned to all regions. If the phrase contains 2 or more accent units (accent groups), then the P, N, T regions are assigned indices corresponding to the numbers of the accent units in the phrase. In “Fig. 8” shows an example of marking 2 accent phrases: **“Befo+r you open the do+or, ...”**.

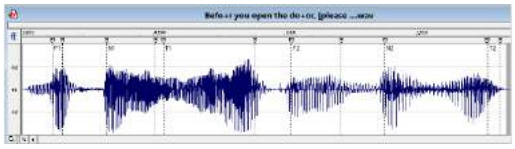

Figure 8. Example of marking a two-accent phrase.

As already mentioned, the software package **“Inton-Trainer”** can be used not only for personal purposes of linguistic education and learning intonation of oral speech, but also in a number of general scientific and practical studies. For example, the **Application** can be successfully used in experimental-phonetic or forensic studies, during which it becomes necessary to compare the standard intonation with the intonation of the phrases studied from various sources. In this case, instead of using an external or built-in microphone, acoustic implementations of these phrases are triggered by pressing the “Open File” button (see “Fig. 3”-“Fig. 6”) from the specially created “TEST” folder (see “Fig. **??**”).

VII. Conclusions

To date, there are demo versions of the “IntonTrainer” system, focused on learning the intonation of Russian, British English, American English, German and Chinese (see site https://intontrainer.by). The software package is recommended for use in the following current fields:

- In linguistic education (Used as a means of visualizing intonation). Primary introduction and study of the basic tone patterns (TP) of oral speech, their pairwise comparisons, application features, as well as their implementation in dialogue, prose and verse.
- In self-learning of intonation of oral speech (Used as a means of intonational training). Individual training for correct pronunciation of TP when studying a foreign language or improving intonation skills of native language in some professions: call center operators, radio and TV announcers, etc.
- In scientific and practical research (Used as a means of comparing intonation from different sources). Experimental phonetics, medical diagnostics, psychological testing, criminalistics, etc..

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 Audio-Processing, 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.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗАТОР И ТРЕНАЖЁР РЕЧЕВОЙ ИНТОНАЦИИ

Лобанов Б. М., Житко В. А.

Объединённый институт проблем информатики НАН Беларуси, Минск, Беларусь

Данная работа посвящена описанию разработанной компьютерной системы, ориентированной на начальное обучение РКИ в рамках освоения учащимися интонационных конструкций русской речи. Понятие интонационных конструкций (ИК1 — ИК7), предложено в 1960-х гг. и эффективно используется во многих современных методических пособиях по обучению РКИ.