Software Model of Analysis and Visualization of Emotional Intonation of the Spoken Phrases

Boris Lobanov, 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*—The purpose of this work is to develop a software model that provides a variety of ways to analyze and visualize the intonation of the main types of human emotions. A set of basic emotions and their acoustical correlations in human speech are described. A database of emotional speech and a software model of analysis and visualization of emotional intonation are presented. The software model created is based on suggested before a computer training system named "IntonTrainer".**

***Keywords*—speech intonation, melodic portrait, intonation analysis, basic emotions, emotional intonation, software model**

I. Introduction

It is well known that human speech conveys not only linguistic messages, but also emotional information. In the theories of emotion, the emotional states are often mapped into a two or three-dimensional space. The two major dimensions consist of a valence dimension (pleasant–unpleasant) and an activity dimension (active–passive) [1]

Generally accepted that there exist a set of basic or fundamental emotions such as: **Calm (neutrality) – Joy – Sadness – Anger – Fear – Surprise**. Below for each of the listed emotions a description of the characteristic psychological state of the person is given.

Calmness (neutrality) – a serene, balanced state of mind, no anxiety, doubts, excitement, worries ...

Joy is a positive emotional state, connected with the possibility of sufficiently fully satisfying an actual need.

Sadness – a negative emotional state associated with the received information about the impossibility of meeting the most important vital needs

Anger is an emotional state, negative in sign, usually occurring in the form of affect and caused by the sudden appearance of a serious obstacle

Fear is a negative emotional state that appears when a subject receives information about a real or imaginary danger.

Surprise – an emotional reaction to sudden circumstances

In the previous works of the authors [2], [3] a software system, called "IntonTrainer", designed to train learners in pronouncing a variety of intonation patterns of speech was described. This work is devoted to the further development of the system in the direction of using it not only as a means of Computer Assisted Language Learning (CALL), but also as a means of experimental research of various functional aspects of intonation.

Intonation is widely recognized as an important aspect of speech that provides both linguistic and sociocultural information. Many people define the purpose of intonation to express the emotional side of speech as the most specific for intonation, although it is little studied in linguistics. Understanding how emotions are expressed in speech is important not only for its own sake but it's also important for understanding how can we know how much of the F0 variability reflecting an emotional content? For this reason, the development of a specialized system that allows for detailed analysis and visualization of emotional intonation is relevant.

II. Acoustic correlates of emotions in human speech

Emotions in human speech may vary according to different physical characteristics [4]. Several researchers have studied the acoustic correlates of emotions in the acoustic features of speech signals [5], [6]. According to [7], there is considerable evidence for specific models of voice expression for different emotions. Emotions can cause changes in respiration, phonation and articulation, which, in turn, affect the acoustic characteristics of the signal [8]. There is also much evidence in the acoustic patterns of vocal affect expression [9].

At present there is little systematic knowledge about the details of acoustic patterns that describe specific emotions in human voice expressions. Typical acoustic features that are considered to be strongly involved in this process include the following:

- the level, range and shape of the contour of the fundamental frequency (F0), which reflect the frequency of vibration of the speech signal and are perceived as the pitch;
- the level of vocal energy, which is perceived as voice intensity, and distribution energy in the frequency spectrum that affects voice quality;

- formants that affect articulation;
- speech speed.

For example, some emotional states, such as anger, fear, and happiness (or joy), are considered to have a high level of arousal [5]. They are characterized by a tense voice with a higher speech speed, high F0 and a wide pitch range. However, sadness (or quiet sadness) and boredom are similar with slower speech, lower energy, lower tone, reduced pitch range and variability of both emotions [10]. There is an influence of voice emotions on the excitation of t nervous systems, primarily in frequency and time, and secondly, in loudness and pronunciation [4].

The common characters of acoustic features of the emotions based on the dimensional analysis with F0 variance, intensity, and speech duration (between silence periods) were also discussed in [11]. Happiness, fear, shyness and sadness are quite even with F0 variance, and surprise, anger, and dominance have strongly varying F0; for the intensity, anger, surprise, disgust, and dominance have the highest value, and sadness and shyness are weakest; the longest duration occurs with happiness, disgust, and surprise, and shyness and sadness have relatively longer pauses between utterances.

III. Emotional Speech Database

In present study we use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [12]. The "RAVDESS" is a validated multimodal database of emotional speech and song. The database is gender balanced consisting of 24 professional actors, vocalizing lexically-matched statements in a neutral North American accent. Speech includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. All conditions are available in face-and-voice, face-only, and voice-only formats. The set of 7356 recordings were each rated 10 times on emotional validity, intensity, and genuineness. Ratings were provided by 247 individuals who were characteristic of untrained research participants from North America. A further set of 72 participants provided test-retest data. High levels of emotional validity and test-retest intrarater reliability were reported. Corrected accuracy and composite "goodness" measures are presented to assist researchers in the selection of stimuli. All recordings are made freely available under a Creative Commons license and can be downloaded at https://doi.org/10.5281/zenodo.1188976.

IV. Analysis and Visualization of Emotional Intonation

Software model that makes available of analysis and visualization of emotional intonation is based on computer trainer system (see: https://intontrainer.by "Fig. 1") provides additional visual feedback, as well as a quantitative estimation of the correctness of speech intonation in the process of teaching various foreign languages [2], [3]. To create a system that allows for detailed analysis and visualization of emotional intonation we add to the system some new functions described below.

Figure 1. The initial window of the Application.

A. Extended parametric display of melodic intonation portraits

The standard graphic of the UMP obtained when selecting in the **Main Settings Window** (see: "Fig. 2") of the **"IntonTrainer"** system **Show UMP** and **Show F0**. In case if the option **Show derivative** F0 is selected, then we will get a joint image of the UMP and its derivative for the phrase "Dogs are sitting by the door", pronounced with neutral emotion (see "Fig. 3").

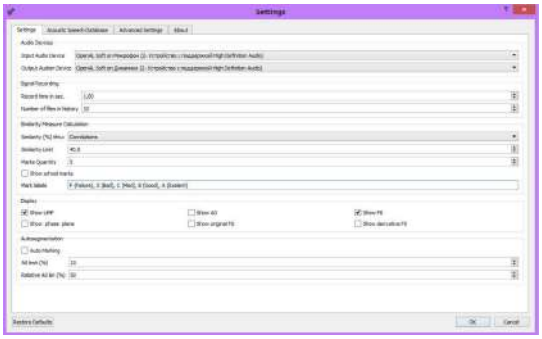

Figure 2. The settings window of the Application.

In the **Main settings window** there is also one new display option for UMP and its derivative. With an additional selection of the **Show phase plane mode**, we obtain the map in coordinates "F0 - dF0", shown in "Fig. 4".

In "Fig. 5" and "Fig. 6" shows the results of display, analysis and comparison of the two the same spoken phrases with **neutral** and **sadness** emotions in the two modes described above.

The extended parametric display of melodic portraits of intonation by the joint analysis of the UMP and its

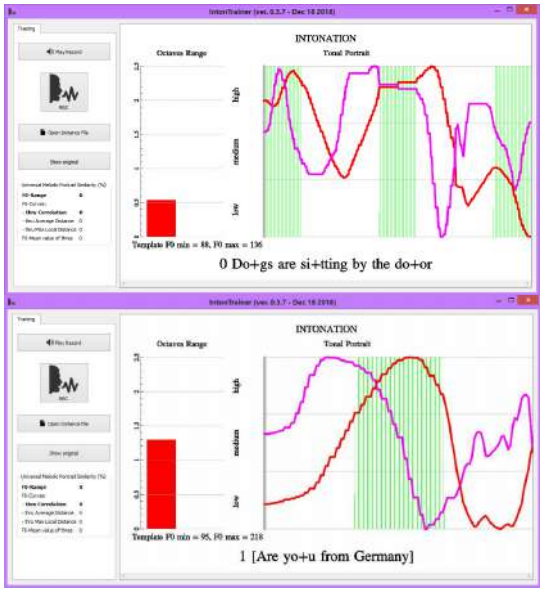

Figure 3. Joint display of the UMP and its derivative.

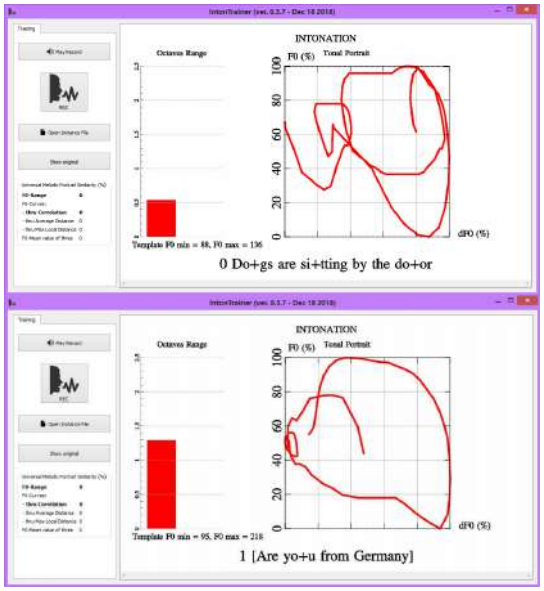

Figure 4. Joint display in the coordinates “F0 - dF0”.

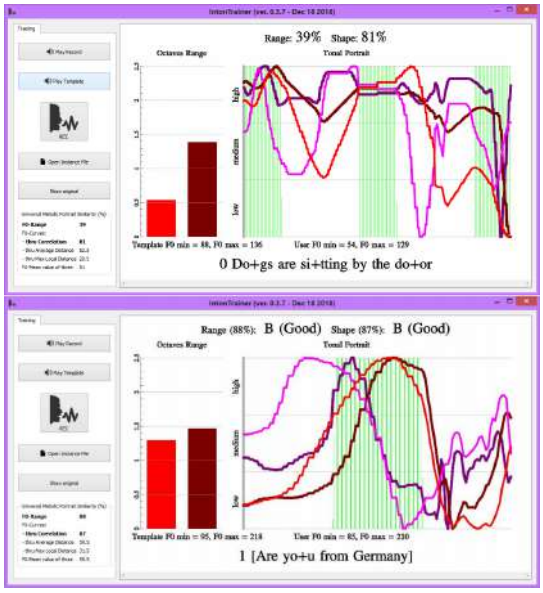

Figure 5. Results of displaying and comparing spoken phrases: UMPs and their derivatives.

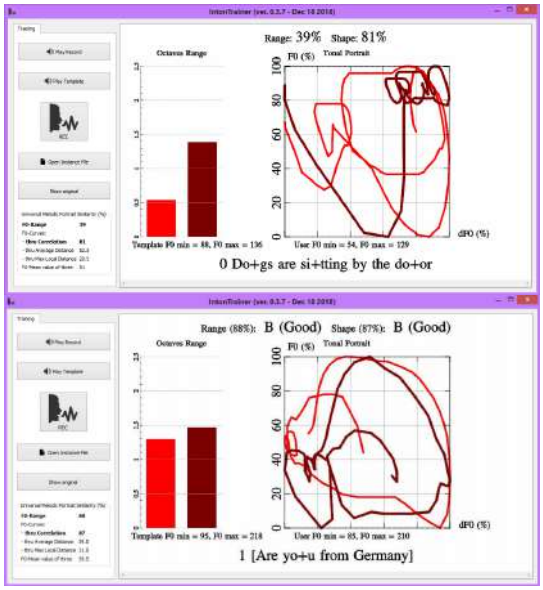

Figure 6. Results of displaying and comparing spoken phrases in the coordinates “F0 - dF0”.

derivative allows one to take into account more subtle differences in the intonation of the reference and spoken phrases. Adding the ability to display the UMP and its derivative in the coordinates “F0 - dF0” made it possible to make a visual comparison of the intonation of the reference and the spoken phrases more vivid.

B. Automating the procedure for marking the analyzed signals to voice regions

At the previous version there was the only possibility of preliminary automatic marking of reference sound files into voice regions, namely: by selecting the operation **Mark Out File** in the advanced settings section - **“Acoustic Speech Database”**.

In new version of the **“IntonTrainer”**, there is an additional possibility of direct automatic marking of reference signals. To do this it is necessary to select in the **Main Settings Window** the **Auto Marking** mode. Signal segmentation into voice regions is carried out on the basis of information about the presence of periodicity in the signal while the signal amplitude is present at sufficiently high amplitude - A0 (t). The user in the

section Auto segmentation is also given the opportunity to independently specify the absolute threshold values - A0 limit (%) and relative - Relative A0 limit (%).

“Fig. 7” shows the UMP of the phrase ”Are yo+u from Germany?” built in the presence of markup on P, N, T - regions, and in “Fig. 8” shows the trajectory of F0 in the case when each of the voice regions of the phrase is assigned the same index N.

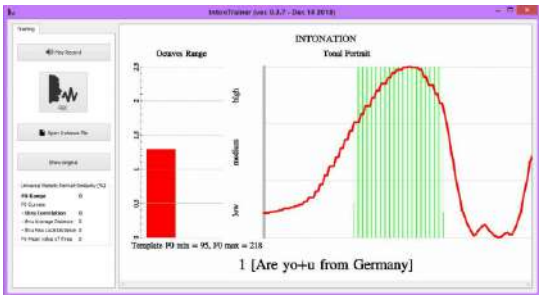

Figure 7. An example of displaying the trajectory F0 (UMP) by manual marking on the P, N, T - regions.

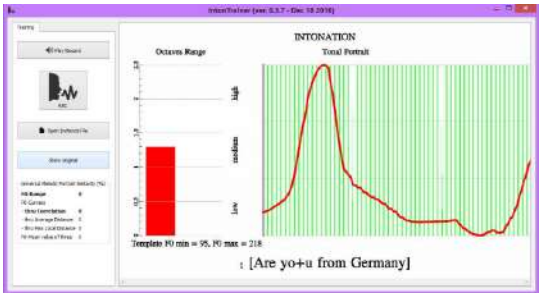

Figure 8. An example of displaying the trajectory F0 by automatically given the same index N for each of voiced regions.

V. Conclusion

The term Emotional Intelligence (EI) became widely known in 1995 with the publication of Goleman’s book: “Emotional Intelligence – Why it can matter more than IQ?”. EI is the capability of individuals (or AI systems) to recognize their own emotions and those of others discern between different feelings and label them appropriately, use emotional information to guide thinking and behavior, and manage and/or adjust emotions to achieve one’s goal(s). This work is the first step towards the creation of software tools for an objective analysis of the physical components necessary for assessing and simulating so-called emotional intelligence.

VI. Acknowledgement

This paper was supported by a BRFFR grant (**Ф17МС-039**)

References

[1] Scherer, K.R., Schorr, A., Johnstone, T., Appraisal Processes in Emotion: Theory, Methods, Research.Oxford University Press, New York and Oxford, 2001

[2] Lobanov B. On a Way to the Computer Aided Speech Intonation Training / B. Lobanov, H. Karnevskaya and V. Zhitko // Proceedings of 19th International Conference on Speech and Computer. Hatfield, Hertfordshire, UK , Springer, 2017, pp. 582-592

[3] Lobanov, B. A Prototype of the Software System for Study, Training and Analysis of Speech Intonation / Lobanov, V. Zhitko, V. Zahariev // Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings – Springer, 2018. – P. 337-346.

[4] Picard, R., Affective Computing, MIT Press, 1997

[5] Banse, R., Sherer, K.R., Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 1996, 614–636

[6] Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., A Database of German Emotional Speech, Proceedings Interspeech 2005, Lissabon, Portugal

[7] Banse, R., Sherer, K.R., Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 1996, 614–636

[8] Scherer, K. R., Vocal correlates of emotion, in A. Manstead & H. Wagner (Eds.), Handbook of psychophysiology: emotion and social behaviors (pp.165-197). London: Wiley, 1989

[9] Scherer, K.R., Kappas, A., 1988: Primate vocal expression of affective state, in D. Todt, P. Goedeking, & D. Symmes (Eds.), Primate vocal communication (pp. 171-194). Berlin: Springer

[10] Breazeal, C., 2001. Designing Social Robots, MIT Press, Cambridge, MA

[11] Abelin, A., Allwood, J., Cross-linguistic interpretation of emotional prosody. In: Proceedings of the ISCA Workshop on Speech and Emotion, 2000

[12] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391

ПРОГРАММНАЯ МОДЕЛЬ АНАЛИЗА И ВИЗУАЛИЗАЦИИ ЭМОЦИОНАЛЬНОЙ ИНТОНАЦИИ В УСТНОЙ РЕЧИ

Лобанов Б. М., Житко В. А.

Данная работа посвящена описанию программной модели позволяющей проводить анализ и визуализацию интонации различных эмоций в устной речи. Описаны базовые эмоции и их акустические проявления в человеческой речи. Представлен набор эмоциональных фраз и их визуализация и анализ. Программная модель реализована на ранее предложенной обучающей системе “IntonTrainer”.

Received 28.12.18