

Albert Ali Salah · Alexey Karpov ·  
Rodmonga Potapova (Eds.)

LNAI 11658

# Speech and Computer

21st International Conference, SPECOM 2019  
Istanbul, Turkey, August 20–25, 2019  
Proceedings



 Springer

Albert Ali Salah · Alexey Karpov ·  
Rodmonga Potapova (Eds.)


# Speech and Computer


21st International Conference, SPECOM 2019  
Istanbul, Turkey, August 20–25, 2019  
Proceedings

*Editors*

Albert Ali Salah  
Utrecht University  
Utrecht, The Netherlands

Boğaziçi University  
Istanbul, Turkey

Rodmonga Potapova   
Moscow State Linguistic University  
Moscow, Russia

Alexey Karpov   
St. Petersburg Institute for Informatics  
and Automation of the Russian Academy  
of Sciences  
St. Petersburg, Russia

ISSN 0302-9743                      ISSN 1611-3349 (electronic)  
Lecture Notes in Artificial Intelligence  
ISBN 978-3-030-26060-6              ISBN 978-3-030-26061-3 (eBook)  
<https://doi.org/10.1007/978-3-030-26061-3>

LNCS Sublibrary: SL7 – Artificial Intelligence

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG  
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Svarabhakti Vowel Occurrence and Duration in Rhotic Clusters in French Lyric Singing . . . . .	227
<i>Uliana Kochetkova</i>	
The Evaluation Process Automation of Phrase and Word Intelligibility Using Speech Recognition Systems . . . . .	237
<i>Evgeny Kostuchenko, Dariya Novokhrestova, Marina Tirskaia, Alexander Shelupanov, Mikhail Nemirovich-Danchenko, Evgeny Choyzonov, and Lidiya Balatskaya</i>	
Detection of Overlapping Speech for the Purposes of Speaker Diarization . . .	247
<i>Marie Kunešová, Marek Hruží, Zbyněk Zajíc, and Vlasta Radová</i>	
Exploring Hybrid CTC/Attention End-to-End Speech Recognition with Gaussian Processes. . . . .	258
<i>Ludwig Kürzinger, Tobias Watzel, Lujun Li, Robert Baumgartner, and Gerhard Rigoll</i>	
Estimating Aggressiveness of Russian Texts by Means of Machine Learning . . . . .	270
<i>Dmitriy Levonevskiy, Dmitrii Malov, and Irina Vatamaniuk</i>	
Software Subsystem Analysis of Prosodic Signs of Emotional Intonation . . . .	280
<i>Boris Lobanov and Vladimir Zhitko</i>	
Assessing Alzheimer’s Disease from Speech Using the i-vector Approach . . .	289
<i>José Vicente Egas López, László Tóth, Ildikó Hoffmann, János Kálmán, Magdolna Pákáski, and Gábor Gosztolya</i>	
AD-Child.Ru: Speech Corpus for Russian Children with Atypical Development . . . . .	299
<i>Elena Lyakso, Olga Frolova, Arman Kaliyev, Viktor Gorodnyi, Aleksy Grigorev, and Yuri Matveev</i>	
Building a Pronunciation Dictionary for the Kabyle Language . . . . .	309
<i>Demri Lyes, Falek Leila, and Teffahi Hocine</i>	
Speech-Based Automatic Assessment of Question Making Skill in L2 Language . . . . .	317
<i>Eman Mansour, Rand Sandouka, Dima Jaber, and Abualsoud Hanani</i>	
Automatic Recognition of Speaker Age and Gender Based on Deep Neural Networks. . . . .	327
<i>Maxim Markitantov and Oxana Verkholyak</i>	
Investigating Joint CTC-Attention Models for End-to-End Russian Speech Recognition. . . . .	337
<i>Nikita Markovnikov and Irina Kipyatkova</i>	



# Software Subsystem Analysis of Prosodic Signs of Emotional Intonation

Boris Lobanov<sup>(✉)</sup> 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.** The main results of the update of the “IntonTrainer” system are the purposes of analyzing and studying the prosodic signs of emotional intonation are described. A distinctive functional feature of the updated system is the creation of an expanded set of prosodic signs of emotional intonation. The paper presents preliminary assessments of their effectiveness using the RAVDESS database of emotional phrases of English speech.

**Keywords:** Speech intonation · Basic emotions · Emotional intonation · Melodic portrait · Intonation analysis · Software model

## 1 Introduction

Well known that human speech conveys not only semantic but also emotional information. There are many different discrete sets of emotions. However, most studies are limited to analyzing the prosodic characteristics of the following 6 emotional states: “Neutrality”, “Joy”, “Sadness”, “Anger”, “Fear”, “Surprise”. There are also a number of emotions attributed quite often to the main ones, such as “Suffering”, “Aversion”, “Contempt”, “Shame”, and in addition, numerous shades of the above emotions.

Today, there is not enough knowledge about the details of acoustic models that describe certain emotions of the human voice. Typical acoustic characteristics that are believed to be involved in this process include the following [1, 2]:

- Level, range and shape of the pitch contour;
- Level of vocal energy and speech rate.

Recently, some important new speech characteristics have been investigated, such as formant frequencies, linear prediction coefficients (LPC), and the Mel-frequency cepstral coefficients (MFCC) [3–5].

In one of the recent works devoted to the analysis of prosodic characteristics of emotions [6], it is proposed to use the following description of the pitch contour:

- The number of maxima in the contour of the main tone in the voiced segment;
- Average value and peak variance;
- Medium tilt;
- Average gradient between two sample points on the pitch curve;
- Variance of pitch gradients.

Our previous work [7] was devoted to the analysis and comparison of the pitch contours of various intonation patterns with the help of software system “IntonTrainer”. It aimed to for study, training, and analysis of speech intonation. The software system “IntonTrainer” contains subsystems that include sets of reference phrases that represent the basic intonation models of Russian, English (British and American versions), German and Chinese speech.

The purpose of this work is to update the existing system by supplementing it with a subsystem for analyzing the prosodic signs of emotional intonation. Such a subsystem should provide analysis and visualization of an effective set of prosodic signs of emotional intonation using the available databases of emotional speech. In this work, for testing purposes, we use the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [8].

## 2 Visual Representation of Emotional Intonation Features

To create a subsystem that allows for detailed analysis and visualization of emotional intonation we add to the “IntonTrainer” system some new functions described below (see: folder name “English Emotions” at the site <https://intontrainer.by>).

The initial Application window is shown in Fig. 1.



Fig. 1. Initial window

After clicking the “Start” button, the main window opens, containing a structured list of reference phrases indicating the name of the announcer, the name of the emotion and the text of the phrase in which it is reflected (see Fig. 2). The numbers 0 or 1 indicated two levels of emotional intensity.

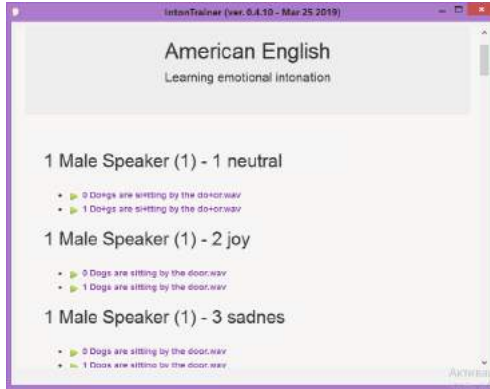


Fig. 2. Main window

By selecting the desired directory with the cursor, for example: “1 Male Speaker (1) - 1 neutral \_1 Dogs are sitting by the door” opens a window (Fig. 3) in which are displayed the results of the intonation analysis of this phrase in a graphic view. The continuous curve in Fig. 3 displays the trajectory of F0 change on the voice sections of the phrase and is presented in the form of the Normalized Melodic Portrait (NMP).



Fig. 3. Window displaying the NMP curve of the phrase “Dogs are sitting by the door” (Neutral emotion)

Segmentation of speech signal into voice regions is carried out automatically (by selecting the Auto Marking mode). Segmentation is based on the information about periodicity in the signal (voice presence), while the presence of a sufficiently high signal amplitude -  $A_0(t)$ . The construction of the NMP curve, in contrast to the Universal Melodic Portrait of the UMP [7], does not require “manual” marking of the phrase on the “pre-core”, “core” and “post-core” sections. The horizontal dashed line on the screen shows the average value of the

NMP curve. Two vertical lines show the information on the position of the center of the NMP curve and its width on the normalized time axis.

It is also possible to calculate and display the derivative of the NMP in a similar way (see Fig. 4). The height of the column (to the left of the NMP) shows the range of variation of the F0 in octaves.

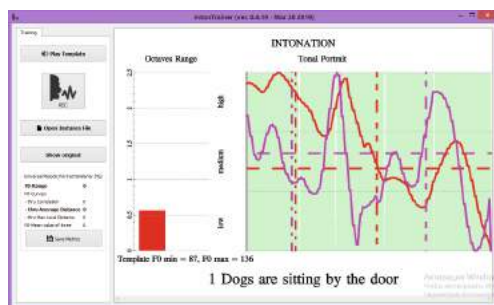


Fig. 4. Displaying of the NMP and its derivative

In the left part in Figs. 3, 4 control buttons are shown with which the following functions are available:

- “Play Template” - listening to reference phrases.
- “Rec” - quick recording of user phrases through a microphone,
- “Open Instance File” - call test phrases from the “TEST” folder.

For individual training of emotional intonation, the user can apply an extended or built-in microphone by pressing button “Rec”.

The user has also the ability to visually compare melodic portraits of various emotions using test phrases from the “TEST” folder by pressing the button “Open Instance File”. In Fig. 5 there are presented for comparison 2 curves of NMP (neutral emotion and emotion of anger) for the same speaker and phrase.

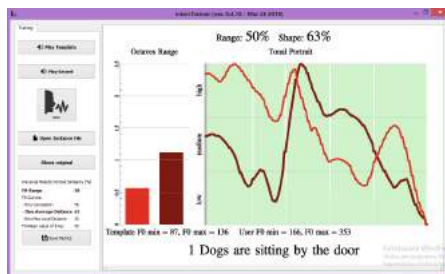


Fig. 5. NMPs of neutral (red line) and anger (dark line) emotions (Color figure online)



### 3 Numerical Evaluation of Signs of Emotional Intonation

The numerical estimates of the signs of emotional intonation are based mainly on the results of measuring various parameters of the NMP presented in Figs. 3, 4, 5. Based on the NMP, the following set of features is calculated:

- Mean Value of the NMP (in % of the maximal value of the NMP);
- Center of the NMP position (in % of the normalized length of the NMP);
- Width of the NMP (in % of the normalized length of the NMP);
- Mean Value of the NMP derivative (in % of the maximal value of the NMP derivative);
- Center of the NMP derivative (in % of the normalized length of the NMP derivative);
- Width of the NMP derivative in % of the normalized length of the NMP derivative).
- Additionally, the following features are calculated from the source signal:
  - F0-Diapason (in octaves):  $D = (F0_{max}/F0_{min}) - 1$ ;
  - F0-Mean Register (in Hz):  $R = (F0_{max} - F0_{min})/2$ ;
  - Voiced Sounds Level (in %, as the average value of the signal amplitudes relative to the maximum value;
  - Voiced Sounds Duration [in seconds], as the total duration of voice sections.

In the left part at the bottom of Figs. 3, 4, 5 control buttons are shown for “Save Metrics” functions. When you click the “Save Metrics” button appears and a page opens in EXCEL, on which a complete set of 10 prosodic features of the reference phrase is written (see Table 1). The obtained data is stored in the same folder where the reference phrase being studied is stored.

**Table 1.** Prosodic features of the phrase “Dogs are sitting by the door” (Neutral emotion)

#	Names of prosodic features	Results
1	F0-Diapason [Octaves]	0,56
2	F0-Register [Hz]	111,50
3	Mean Value of the NMP [%]	46,14
4	Center of the NMP [%]	42
5	Width of the NMP [%]	33,03
6	Mean Value of the NMP derivative [%]	54,65
7	Center of the NMP derivative [%]	5,77
8	Width of the NMP derivative [%]	54,71
9	Voiced Sounds Level [%]	24
10	Voiced Sounds Duration [Sec]	1,71

Table 2 shows an example of the results of calculating the numerical values of the prosodic signs of a phrase expressing the emotion “Anger”.

**Table 2.** Prosodic features of the phrase “Dogs are sitting by the door” (Anger emotion)

#	Names of prosodic features	Results
1	F0-Diapason [Octaves]	1,17
2	F0-Register [Hz]	258,5
3	Mean Value of the NMP [%]	36,94
4	Center of the NMP [%]	42,69
5	Width of the NMP [%]	36,91
6	Mean Value of the NMP derivative [%]	44,26
7	Center of the NMP derivative [%]	44,92
8	Width of the NMP derivative [%]	47,34
9	Voiced Sounds Level [%]	34
10	Voiced Sounds Duration [Sec]	2,43

At the time, when the user makes a visual comparison of NMPs of reference and tests phrases (see Fig. 5), it is also possible to calculate and to click the “Save Metrics” button to store values of ratios for each prosodic signs of this couple of phrases in dB scale. The results of a calculation based on the data given in Tables 1 and 2 (a pair of phrases with “Anger/Neutrality” emotions) is shown in Table 3.

The use of ratios in dB scale allows the comparison of a pair of phrases with different emotions, using prosodic signs of different nature and in various units of measurement.

**Table 3.** Relative values for the prosodic features of a pair “Anger/Neutrality” emotions

#	Names of prosodic features	Results
1	F0-Diapason	1,33
2	F0-Register	3,66
3	Mean Value of the NMP	-1,22
4	Center of the NMP	0,21
5	Width of the NMP	-0,11
6	Mean Value of the NMP derivative	-2,69
7	Center of the NMP derivative	0,30
8	Width of the NMP derivative	-0,16
9	Voiced Sounds Level	2,06
10	Voiced Sounds Duration	1,27

## 4 Preliminary Testing of the Developed Signs of Emotional Intonation

For testing of the developed signs, we used the RAVDESS emotion database [8]. It is a validated multimodal database of emotional speech. The database is

gender balanced consisting of 24 professional actors and actresses, vocalizing lexically-matched statements in a neutral North American accent. Speech includes Neutral, Happy, Sad, Angry, Fearful, Surprise, and Disgust expressions of emotions. The common number of emotional speech samples available for testing is 1534. The testing and analysis of such a big database become possible only with the involvement of special programs for processing big data, for example, using neural network algorithms that we are planning to realize in the future.

Below on Fig. 6 we present comparative graphs of the NMP and the tables with data in the logarithmic ratio for each of the 10 signs (see: Table 3) for three pairs of emotions expressed by one of the male and female actors.

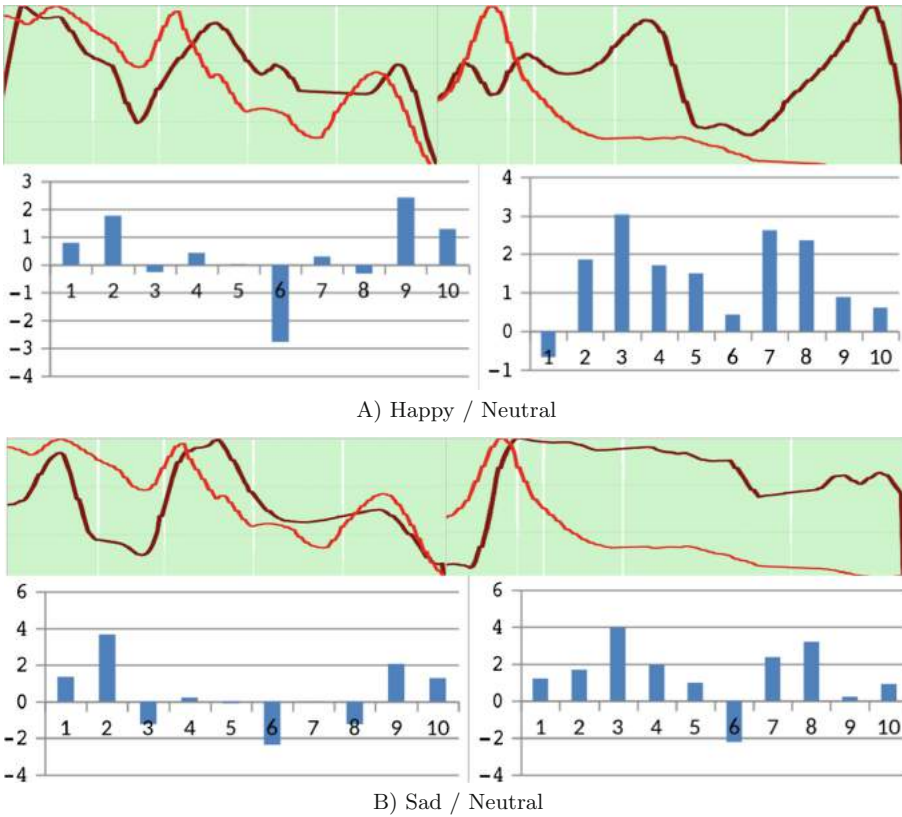
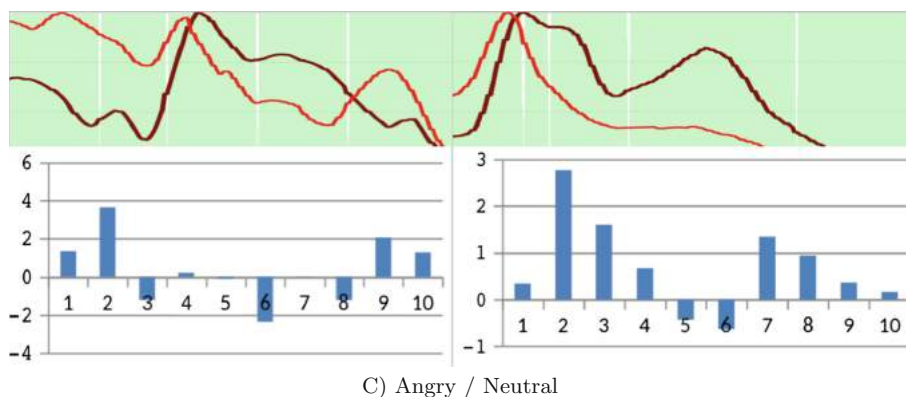


Fig. 6. Graphs of the NMP (top rows) and the tables (bottom rows) for three pairs of emotions expressed by one of male (left column) and female (right column) actors.



**Fig. 6.** (*continued*)

The present study of the effectiveness of the developed signs of emotional intonation showed their significant distinctive power when comparing different pairs of emotions. It should be noted that one group of signs can play a large role in distinguishing the emotions of one speaker and be poorly informative for another speaker. So, for example, from Fig. 6 it is clear that signs 1, 2, 9, 10, calculated from the source signal, play the most significant role in distinguishing emotions in a male speaker. To distinguish the emotions of a female speaker, the most significant role is played by signs of 3, 4, 5, 6 calculated from NMPs.

## 5 Conclusions

The task of upgrading the “IntonTrainer” system wasn’t including the creation of a valid speech emotion recognition model. The ultimate goal of refinement was limited to the creation of such a software tool that would provide analysis and visualization of an extended set of prosodic signs of emotional intonation, and which could be used as a new tool for phonetic studies of speech.

We do not exclude also some applied aspects of the application, for example, in the tasks of teaching the required emotional intonation of actors, as well as people of various professions who are striving to enhance their so-called “emotional intelligence (EQ)”. For this, in the Similarity Measure section (see the Main settings window of the “IntonTrainer”) it is possible choose a method for assessing the intonation proximity of the spoken phrase to the reference one, using various similarity measures. The chosen method of calculating the intonational similarity is then used in calculating the digital or verbal evaluation assessment of the intonation quality of the spoken phrase.

## References

1. Banse, R., Sherer, K.R.: Acoustic profiles in vocal emotion expression. *J. Pers. Soc. Psychol.* **70**(3), 614–636 (1996)
2. Abelin, A., Allwood, J.: Cross-linguistic interpretation of emotional prosody. In: *Proceedings of the ISCA Workshop on Speech and Emotion* (2000)
3. Ververidis, D., Kotropoulos, C., Pitas, I.: Automatic emotional speech classification. In: *Proceedings of 2004 IEEE International Conference on Acoustics, Speech and Signal Processing*, Montreal, vol. 1, pp. 593–596, May 2004
4. Xiao, Z., Dellandrea, E., Dou, W., Chen, L.: Features extraction and selection for emotional speech classification. In: *2005 IEEE Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 411–416, September 2005
5. Pao, T.-L., Chen, Y.-T., Yeh, J.-H., Li, P.-J.: Mandarin emotional speech recognition based on SVM and NN. In: *Proceedings of the 18th International Conference on Pattern Recognition (ICPR 2006)*, vol. 1, pp. 1096–1100, September 2006
6. Sbattella, L., Colombo, L., Rinaldi, C., Tedesco, R., Matteucci, M., Trivilini, A.: Extracting emotions and communication styles from prosody. *Physiological Computing Systems*. LNCS, vol. 8908, pp. 21–42. Springer, Heidelberg (2014). [https://doi.org/10.1007/978-3-662-45686-6\\_2](https://doi.org/10.1007/978-3-662-45686-6_2)
7. Lobanov, B., Zhitko, V., Zahariev, V.: A prototype of the software system for study, training and analysis of speech intonation. In: Karpov, A., Jokisch, O., Potapova, R. (eds.) *SPECOM 2018*. LNCS (LNAI), vol. 11096, pp. 337–346. Springer, Cham (2018). [https://doi.org/10.1007/978-3-319-99579-3\\_36](https://doi.org/10.1007/978-3-319-99579-3_36)
8. Livingstone, S.R., Russo, F.A.: The Ryerson Audio-visual Database of Emotional Speech and Song (RAVD ESS): a dynamic, multimodal set of facial and vocal expressions in North American English (2018). <https://doi.org/10.1371/journal.pone.0196391>