Emotional Speech Classification for Romanian Language - Preliminary Results

Silvia Monica FERARU
Romanian Academy of Iaşi
str. T. Codrescu nr.2, RO-700481 Iaşi
monica.feraru@gmail.com

***Abstract* — The paper reports preliminary results of emotional classification in speech, with focus on distinguishing the emotions (joy versus neutral speech, joy versus sadness speech, joy versus fury speech). We compared the results given by Weka (SMO, RBF) and Matlab (KNN) software. We use SRoL emotional database which contains recordings from 18 speakers expressing four emotional states. The results indicate that joy and sadness utterances can be distinguished as well as joy and neutral.**

***Index Terms* — classification rate, emotional speech database, Romanian language**

I. INTRODUCTION

Nonverbal information (e.g. the emotional state) is important in communication between humans. Sometimes the verbal information is more important than the emotional states. The communication between humans and computers is becoming more spread. In the last years many efforts have been done in the speech technology, in special to speaker verification, speech recognition, etc. The recognition rates for intercultural recognition increased with the length of the speech samples.

Emotion recognition is a challenging and exciting field, but also a difficult attribution. People can recognize emotions with about 60% accuracy and emotional facial expressions with about 70-98% [1]. The physiological changes according to emotions can be observed on changes of the body surface and/or autonomic nervous system: e.g., skin conductivity, electrocardiogram, electromyogram, and blood volume pressure [2]. The researchers have investigated emotion recognition from multi-modal physiological features [3].

The recent studies in emotion recognition have used the acoustic information from vocal signal. In emotion classification of speech, the most used parameters are statistics of fundamental frequency, energy contour, duration of silence and voice quality. In this study, we used the formats values (F1 and F2) and features specified by emobase.conf from 'opensmile' (describe in section III).

In the literature, the focus has been on the most spoken languages for example English, German, Spanish, French and very little is known about the Romanian language [4-8].

The paper is organized as follows: section II offers information regarding the emotional speech database for Romanian language; in section III are presented the obtained results of the emotional classification, and in the end we draw the conclusions.

II. EMOTIONAL SPEECH DATABASE

Prof. H-N. Teodorescu has taken the initiative to build a sound archive, a "dictionary of sounds" for the Romanian language. This corpus is located at the address: http://iit.iit.tuiasi.ro/romanain_spoken_language/index.htm.

In this way, the Romanian language has a web-based, freely accessible sound archive - SROL, as most European language do, which is used already in research, in teaching and laboratory activities in the "Speech Technology" class given for the master degree in "Computational Linguistics". We hoped that the database will be helpful also to other universities from Romania where foreign students learn the Romanian language, moreover in other academic media or as an online tool by foreign students and teachers.

The content of the site can be used for educational purposes, such as analysis of sounds, analysis of specificities of the Romanian language pronunciation compared to other languages, Romanian language learning aided by computer, for improving voice recognition systems based on acoustical features as well as for research purposes [9]. The site has been optimized algorithmically, at the graph level, after an analysis of the graph and of the paths followed by various categories of plausible users [10]. New recordings with the corresponding annotations and documentations are continuously added.

A large range of tools are provided, including extensive documentation on the subject condition, peculiarities, spelling conditions, recording conditions, etc., as well as annotations, moreover speech processing software tools.

In the figure 1, we exemplify the sitemap of the SRoL speech database.

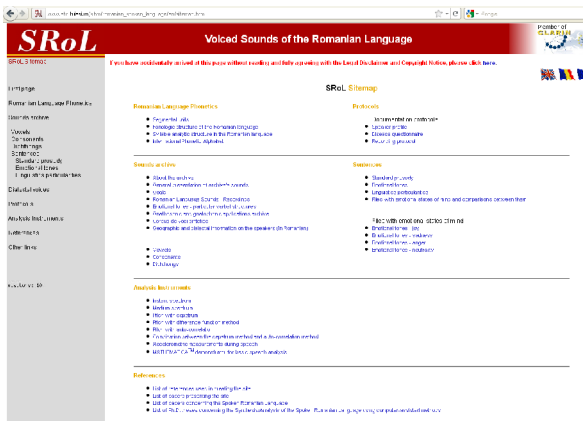

Fig. 1. A screenshot of the sitemap of the SRoL database

The persons who made the recordings were previously

informed about the objective of the project. The speaker signed an informed consent according to the Protection of Human Subjects Protocol to the U.S. Food and Drug Administration and with Ethical Principles of the Acoustical Society of America for Research Involving Human Subjects.

The recordings were made using the GoldWave software, and the sampling frequency being 22050 Hz. Every speaker pronounced the sentence minimum for three times, maximum for seven times. The sound was saved in different format. In the analysis we used only the wav. file on 16 bits.

The database has the recording protocol (which contains information about the noise, the soundboard, the microphone used and the corresponded drivers) and the documentation protocol which contains the speaker profile (regarding the linguistic, ethnic, medical, educational, professional information about the speaker), and a questionnaire (regarding the healthy state of the speaker).

The consistency of the emotional content in the speech recordings has been verified by several listeners; the confusion matrix has proved that all emotions are identified, with accuracy around 43-80% by the listeners. The emotional database for the Romanian language contains short sentences or phrases fragments with different emotional states. The emotional states analyzed are happiness, sadness, fury, and the neutral tone. We choose these emotions to be analyzed because they are the basic ones. All chosen sentences contain all the vowels from Romanian language. The registered sentences are: Mother is coming (Vine mama, in Romanian), Who did that? (Cine a facut asta, in Romanian), Last night (Aseara, in Romanian), Anyway, you can win / get the desired place, anyway (Oricum îți poți câștiga locul dorit), You will win / get the desired place (Îți vei câştiga locul dorit) and My man done / manufactured it / sapped him (Omul meu îl lucră, in Romanian).

III. RESULTS AND DISCUSSIONS

For the experiments reported in the paper, we used Weka toolkit (http://www.cs.waikato.ac.nz/~ml/weka/). We use 10-folds cross validation technique and in the preprocess step, we selected the option RandomSubset.

The extracted features was made using OpenSmile -1.0.1. The feature set specified by emobase.conf contains the following low-level descriptors (LLD): Intensity, Loudness, 12 MFCC, Pitch (F0), Probability of voicing, F0 envelope, 8 LSF (Line Spectral Frequencies), Zero-Crossing Rate. Delta regression coefficients are computed from these LLD, and the following functionals are applied to the LLD and the delta coefficients: Max./Min. value and respective relative position within input, range, arithmetic mean, 2 linear regression coefficients and linear and quadratic error, standard deviation, skewness, kurtosis, quartile 1-3, and 3 inter-quartile ranges.

The results by comparing two emotional states from SRoL database are presented in Table I. The results are better when it is used SMO algorithm (Sequential Minimal Optimization) than RBF network (Radial Basis Function). The sadness state is the most recognized and the less recognized is fury state for the Romanian language.

TABLE I. THE CORRECTLY CLASSIFIED INSTANCES (56) FOR THE PAIRS OF EMOTIONS FROM SROL DATABASE

Type of classification	joy-neutral	joy-sadness	joy-fury
SMO -10folds	79%	88%	55%
SMO -7folds	79%	88%	63%
RBF -10folds	63%	68%	46%
RBF -7folds	61%	64%	48%

The recognition rates using SMO and RBF with 7 folds for each emotional state are given in Table II. The best recognition rates are by comparing the joy with sadness state and the less recognition rate are comparing the fury with joy state. This happens because the frequency is in the same range of variation.

TABLE II. THE RECOGNITION RATES OF THE EMOTIONS PAIRS USING WEKA

Type of classification	Joy	Neutral	Joy	Sadness	Joy	Fury
SMO -7folds	75%	82%	82%	92%	63%	37%
RBF -7folds	71%	50%	75%	54%	48%	52%

In the fig. 2 and 3, we exemplify a screenshot of the results of emotions classification and a way of visualization using Weka.

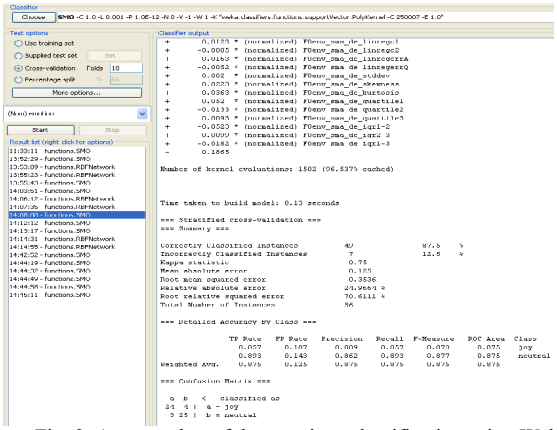

Fig. 2. A screenshot of the emotions classification using Weka

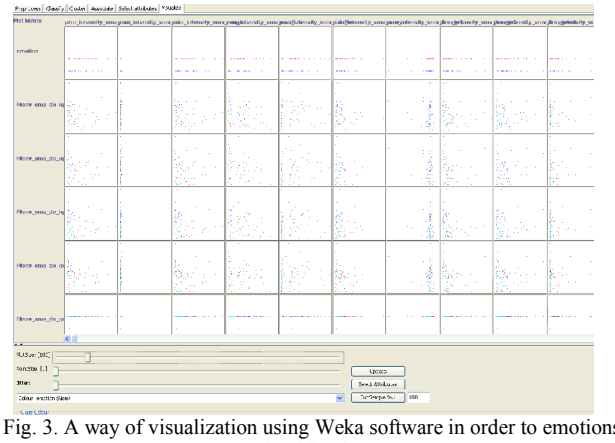

Fig. 3. A way of visualization using Weka software in order to emotions classification

In order to compare the obtained results by two different methods of emotional classification, it was used also the Matlab software. Based on the audio recordings files, we annotated the sentences at four levels using Praat software (www.praat.org.). Then it was computed the formants values (F0, F1, F2, F3) of the all vowels from the Romanian

language on the entire duration of the vowel. Based on the files of the formants values and on each vowel (/a/, /e/, /i/, /o/, /u/, /ă/) we run two algorithms (K-means and K-NN) from Matlab. In this study, we used the fundamental frequency, F0 and the F1 and F2 formants values.

Unsupervised learning allows the identification of the completely new concepts based on the known data - the algorithm K-means. Supervised learning is a type of inductive learning which is based on a set of examples of the problem and forms an evaluation function in order to allow the classification of new data sets - the algorithm k-NN (k-nearest neighbor). We choose the distance from the clusters to be simple that means the minimum distance between objects in the two clusters.

Fig. 4 and 5 exemplify the spatial representation of the /a/, /e/, /i/ vowels, for the sadness and neutral tone, in the Romanian language. We note with red points the formats values of the /a/ vowel, with pink the values of the /e/ vowel, and with blue the values of the /i/ vowel.

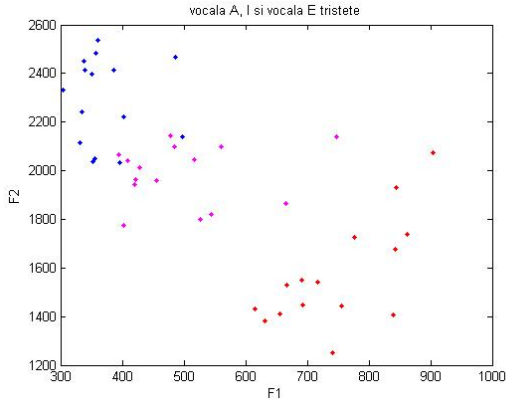

Fig. 4. The spatial representation of the formants values of the /a/, /e/, and /i/ in sadness state for Romanian language

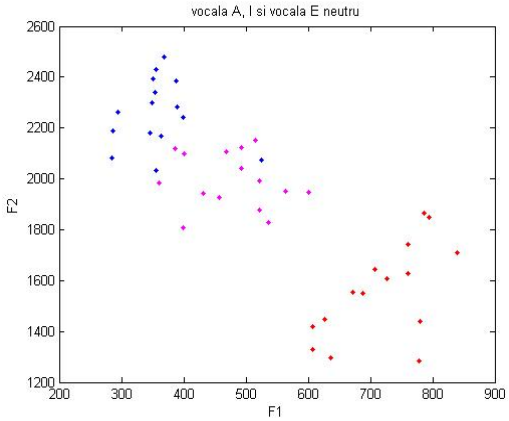

Fig. 5. Spatial representation of the formants values of the /a/, /e/, and /i/ in neutral tone for Romanian language

Fig. 6 exemplifies the spatial representation of the fundamental frequency, F0 and the F1, F2 formants values for happiness state. We note with red points the formats values of the /ă/ vowel, with blue the values of the /i/ vowel, with magenta the values of the /o/ vowel, with cyan the values of the /a/ vowel, with green the values of the /e/ vowel, and with black the values of the /u/ vowel.

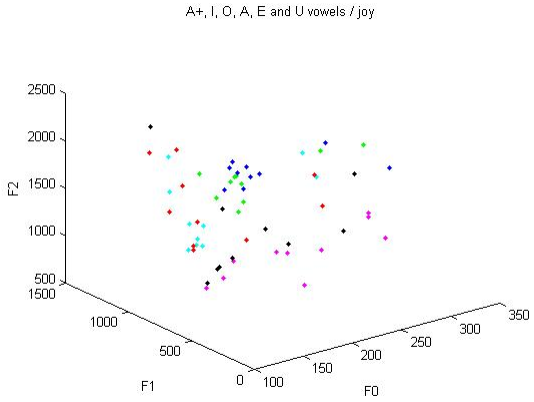

Fig. 6. Spatial representation of the formants values of the vowels in joy state for Romanian language

Using KNN algorithm for the /a/ vowel (18 speakers, only the F1 and F2 formants values, k=5), we obtained for the k=1, the classification rates for neutral tone 75% and for happiness 30% and for k=3, the classification rate for sadness is 33% and for the fury 44%.

For the /e/ vowel (18 speakers, only the F1 and F2 formants values, k=5), we obtained for the k=1, the classification rates for neutral tone 60% and for happiness 42%.

For the /o/ vowel (15 speakers, only the F1 and F2 formants values, k=5), we obtained for the k=3, the classification rates for neutral tone 66% and for sadness 50%.

For the /u/ vowel (15 speakers, only the F1 and F2 formants values, k=5), we obtained for the k=1, the classification rates for neutral tone 66% and for sadness 25%.

For the Spanish language the classification rates using also prosodic features they obtained for joy 44.4%, for fury 48.9%, for sadness 75.6% and for neutral tone 66.7%. The happiness state was the most difficult emotion to identify. The fury state is the second best identified emotion. The sadness state has proven with a high identification accuracy and high precision [11].

For Danish language [12] the classification rates of a Bayes classifier for persons of both genders are 51% for neutral, 36% for joy, 70% for sadness and 31% for fury.

For German language [13] the classification rates for the emotions are: neutral 88%, fury 79%, happiness 48% and sadness 80%. For the Indonesia language the percentages of the recognition rates are: neutral 70%, fury 64%, happiness 28% and sadness 58%.

For French language [14] the emotion recognition rates using a multilayer perceptron (with AHL features) are: neutral 57.86%, fury 50.71%, happiness 49.64% and sadness 58.27%.

The distribution of emotion classes among speech after [15] is the next one: neutral is by far the most common emotion, after it follows the happiness state; fury and impatient states are the next negative emotions.

In his studies, Scherer [16] said that the languages of the countries with the highest accuracy rates are Germanic origin (Dutch and English), followed by Romanic group (Italian, French and Spanish). The lowest recognition rate belongs to the Indo-European language family.

IV. CONCLUSIONS

The reported research presents some preliminary results regarding the emotions identification and classification in Romanian language. The classification rates for emotional states are: sadness state 92%, neutral tone 82%, happiness state 78.5% and fury state 44.5%.

We used Weka - Data Mining Software (http://www.cs.waikato.ac.nz/ml/weka/) using the functions RBF Network and SMO with different number of folds and Matlab software. The values of formants values we computed with Praat (http://www.fon.hum.uva.nl/praat/) and the recordings where made with Goldwave (http://www.goldwave.com/) and belongs to SRoL emotional database.

According to our experiments, the classification rates are better in the neutral tone compared with sadness state. The happiness and fury state are not distinguished very well. The happiness state is the more easily confused emotion. The recognition systems of emotional states must be trained by speaker in order to distinguish the fury state.

The emotional intra-speaker states can be distinguished, but we cannot specify the emotional inter-speaker states.

In the future we want to make other type of classifications and to select the significant features extracted from voice signal. We think that also the numbers of features that are used in classification are important too.

ACKNOWLEDGMENTS

This paper is supported by the Sectoral Operational Programme Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number POSDRU ID 56185.

REFERENCES

[1] R.W. Picard, Affective computing, The MIT Press, 1997

[2] Picard R.W., Vyzas E., Healey J., Toward machine emotional intelligence: Analysis of affective physiological state, IEEE Transactions on Pattern Analysis and Machine Intelligence, 23(10):1175–1191, 2001

[3] Takahashi K., Comparison of emotion recognition methods from bio-potential signals, The Japanese Journal of Ergonomics, 40(2):90–98, 2004

[4] H.N. Teodorescu, M. Zbancioc, M. Feraru, "The analysis of the vowel triangle variation for Romanian language depending on emotional states", ISSCS Conference, Iasi, Romania 30June-1Jul.2011, ISBN 978-1-4577-0201-3, pp. 331-334

[5] H.N. Teodorescu, M. Zbancioc, M. Feraru, "Statistical characteristics of the formants of the Romanian vowels in emotional states", International Conference on Speech Technology and Human-Computer Dialogue SPeD 2011, 18-21 may 2011 Brasov, Romania, ISBN 978-1-4577-0439-0, pp. 13-22, IEEE publication, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=5940725

[6] S.M. Feraru, "The comparisons between the formants values in French and Romanian Languages", Proceedings of the International Conference on Languages, E-Learning and Romanian Studies, 3-5 Iunie 2011, Isle of Marstrand, Sweden– to be published

[7] S.M. Feraru, "Emotional expressiveness in the Romanian and German language", Gr. T. Popa University of Medicine and Pharmacy, Publishing House, 2011, ISBN: 978-606-544-078-4, pp. 61-65, http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=6150420

[8] S.M. Feraru, "Emotional expressiveness in Romanian language", WSEAS/INEEE International Conferences, 2012, April 18-20, Rovaniemi, Finland, ISBN: 978-1-61804-085-5, pp. 208-212

[9] Feraru S.M., Teodorescu H.N., The Emotional Speech Section of the Romanian Spoken Language Archive, 5th European Conference on Intelligent Systems and Technologies, Iaşi, Romania, 2008

[10] Teodorescu H.N., Pistol L., Using Graphs to Improve the Structure of a Web Site, 2st International Conference on Electronics, Computers and Artificial Intelligence, Piteşti, România, 2007

[11] R. Barra, J.M. Montero, J. Macías-Guarasa, L.F. D'Haro, R. San-Segundo, R. Córdoba, Prosodic and segmental rubrics in emotion identification, ICASSP, 2006

[12] D. Ververidis, C. Kotropoulos, Automatic speech classification to five emotional states based on gender information, European Signal Processing - Eusipco, 2004, pp. 341-344

[13] Klaus R. Scherer, R. Banse, Harald G. Wallbott, Emotion inferences from vocal expression correlate across languages and cultures, Journal of cross-cultural psychology, vol. 32 no. 1, pp. 76-92, 2001

[14] V. Hozjan, Z. Kacic, Context-Independent Multilingual Emotion Recognition from Speech Signals, International journal of speech technology, no. 6, pp. 311–320, 2003

[15] S. Can, B. Schuller, M. Kranzfelder, H. Feussner, Emotional factors in speech based human-machine interaction in the operating room, IJCARS, Vol. 5 (Suppl. 1) 2010

[16] Klaus R. Scherer, A cross-cultural investigation of emotion inferences from voice and speech: implications for speech technology, ICSLP 2000 (http://www.isca-speech.org/archive/icslp_2000/i00_2379.html)