A NOTE ON TESTING THE RECOGNITION OF EMOTIONAL STATES AND TONES IN SPEECH

LAURA PISTOL[1] and HORIA-NICOLAI TEODORESCU[1,2]
Corresponding member of the Romanian Academy

[1] *Institute of Computer Science of the Romanian Academy, Iaşi Branch*
[2] *"Gheorghe Asachi" Technical University of Iaşi*
hteodor@etc.tuiasi.ro; laura.pistol@iit.tuiasi.ro

We report on a web-based tool for assessing the interpretation degree of the common listeners of the SRoL emotional speech and, separately, tone database for short sentences, in the Romanian language. We designed the tool for allowing us a statistical characterization of the interpretation degree of common, anonymous listeners of the emotional state and tone that characterize the recorded utterances. The intended emotional contents of the sentences fall into one of the categories: *neutral* (relaxed, no emotion expressed), *joy*, *anger* and *sadness*. The intended tones are *neutral* (relaxed tone), *exclamatory*, and *interrogative*. Pre-recorded spoken sentences are used. The main question addressed refers to the reliability of emotion and tonality communication, based only on speech fragments, in the Romanian language. The results show that the tonality and the emotional states are communicated with variable degrees of success; the accuracy was between 51% and 62.30% for emotional states and between 77.29% and 87.70% for tones.

Key words: *emotional speech recognition, tone recognition.*

1. INTRODUCTION

The speech conveys both linguistic information and paralinguistic information (intonation, state, emotion etc.) in human-to-human interaction. In contrast to text messages, in speech the same linguistic content may carry several messages, depending on the speaker emotion and/or tone. The emotional voice contributes to increase the semantic content and to make the linguistic content more comprehensible to the listeners. The listeners, in turn, can react in different ways to the same linguistic content, according to the perceived paralinguistic message. Hence, assessing the comprehensibility of emotions in voice is important for understanding voice communication.

The emotional states and tones in speech have several potential applications, from medical diagnosis help to entertainment (for example, intelligent toys). Earlier studies on the interpretation of emotional states and tones are [1, 4, 5, 7–12, 14]. The reported recognitions rates vary widely, depending on the methodology, language, voice database, type of speakers (typically, actors, but other categories are also analyzed), and type of listeners (expert listeners or not, native or nonnative listeners), level of noise, type of tone, and type of emotion. For example, the emotional states, for the Danish recordings database were recognized in 67% of the cases [2], for the German recordings were recognized in 80% of the cases [16].

Two goals of the ongoing research of our team that develops the language resource (annotated and documented database) SRoL are to determine: i) to what degree are the emotional states and tonalities comprehensible and recognizable in the Romanian language, for “common speakers” and “common listeners”; ii) to what degree people have the ability of perceiving prosody (tonality) in brief messages uttered by persons that are not familiar to the listeners.

While this is not the first attempt to determine the degree of emotion recognition in voice for the Romanian language – the first such attempts are believed to be those reported in [4, 9–12] – those preliminary studies were made on a small number of listeners and for fewer sentences, or addressed other issues of interest related to emotional representation and perception. The results reported herein are statistically more significant than in our previously reported results, because of the increased number of responding listeners. The tool reported here complements the SRoL language resource.

To prevent any confusion that might arise, we put forth some clarifications on the aim and scope of this research. Communication being a two-end process, studies may address the communication in its entirety, or the abilities of the listeners, or the abilities of the speakers to convey information. When speakers’ ability is analyzed, the listening evaluators must be experts; the vice versa requirement should be enforced when listeners’ abilities are assessed. Here, we address the “average” communication as a whole, meaning that we mix professional speakers with common ones (yet, all speakers have higher education), that is, listeners who are experts in speech analysis. Moreover, we are interested in communication through brief sentences, all syntactically correct, but not all having an easy to grasp semantic content.

The paper is structured as follows. In Section 2 we summarize the research methodology, in Section 3 we describe the Web-based validation tool, in Section 4 we present briefly the results, while in Section 5 contains discussion and conclusions.

2. THE METHOD

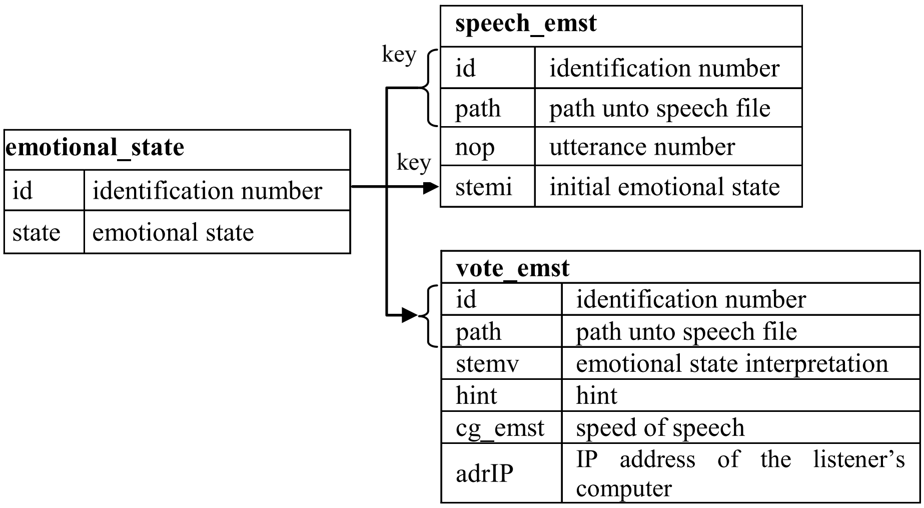
The emotional speech database for the Romanian language is due to a joint effort of teams from the “Gheorghe Asachi” Technical University of Iaşi (UT Iaşi), Institute for Computer Science of the Romanian Academy (IIT), and “Alexandru Ioan Cuza” University of Iaşi. The speech files are recorded in the laboratory for Bio-Medical Engineering Laboratory (UT Iaşi). The laboratory is a room with reduced noise, but not an anechoic room. A rigorous signal recording procedure was observed. The sounds have been recorded using a SONIC Stereo Dynamic Headphones HP-259 with a response frequency of 20-20.000 Hz, microphone sensibility 58dB±2, and headphones sensibility of 100dB/mW. To avoid the noise and distortions that might appear due to the recording process, the microphone placement was under the mouth, approximately at the chin level, a few centimeters from it. The recordings were performed using Goldwave^{TM}5.0 at a sampling frequency of 22050 Hz with PCM signed, 16 bit resolution, and single channel. We informed the speakers prior to recording on the project objectives and assured the confidentiality of the personal information. The speakers signed an informed consent in accordance with the US Food and Drug Administration's Protocol of Protection of Human Subjects [18] and the Ethic Principles of the American Acoustic Association regarding Research Involving Human Subjects [17]. Common people (not actors) uttered the recorded sentences. The speakers group comprises five persons for feminine voice and nine persons for masculine voice; they are all healthy persons aged between 25–45 years, born and educated in the middle area of Moldova, Romania. Importantly, the expert listeners consider that three of the male speakers have low ability to express emotions [13(a)]. More details can be found on SRoL site, on pages related to sentences with tonalities and emotions, *e.g.* at [13(b)].

In the sequel, we summarize a few explanations on SRoL given in [13(a)]. The speakers were instructed to utter the sentences with one of the emotional states, expressing *neutral*, *joy*, *anger*, and *sadness*, but with no specific tone, or with a given tonality and no emotion [13(a), 13(b)]. The four uttered sentences with emotional content but with no specific tonality (no exclamatory or interrogative tones) are *Aseară*, meaning *Yesterday evening*, *Vine mama*, meaning *Mother is coming*, *Cine a făcut asta*, with two possible interpretations, namely *That who made it*, or *Who made this*, and *Ai venit iar la mine*, meaning *You came back to me again*. The three spoken sentences with various possible tonalities but without emotions are *Aseară*, *Vine mama*, *Cine a făcut asta*. In the research reported in [10–12], the recorded utterances were presented to native Romanian speakers for emotional states and tones assessment. In the research reported here, the evaluators were required to access the web page [13(c), 13(d)] and to choose and save their option after each examined record.

Here, we report on the emotion validation performed by 16 listeners (13 of them, colleagues from our Institute, not all researchers; one of the users is editor at Radio Iaşi; for the other two, we have information from the Internet Protocol address, and we only know that one is from Romania and the other is from Japan; the comments written by them are in Romanian, so we believe that they are Romanian natives, or at least Romanian-language speakers). The listeners whom we know are grouped according to two criteria: knowing / not knowing the recordings database, respectively expert / non expert in the field. According to these criteria, moreover taking into account if the listeners have intimate knowledge of the database, the listeners fall into three categories: i) six listeners who have not worked for the database and who previously have not worked in the area of speech signal processing; ii) three experts in speech processing but not having contributed to our database (who previously worked for the speech database); iii) four expert evaluators (who worked on the creation of the recordings database). All the voting took place during the morning; therefor the listeners were not tired during the tests.

3. DESCRIPTION OF THE WEB-BASED VALIDATION TOOL

The recognition tests for speech emotional states and the tones recognition tests were carried out using a web page tool we specified and implemented. The tool is available at [13(c)] for emotions and at [13(d)] for tones. The page is available only in the Romanian version of the site, because only fluent Romanian speakers are expected to validate emotions in voice. We sketch in Figure 1 the structure of the database tables used for the public emotional states and tone validation pages. We use six tables, namely: i) the *emotional_state* table contains the list of coding emotional states; ii) the *speech_emst* table contains the emotional states recordings; iii) the *vote_emst* table contains the votes for the emotional state; iv) the *tones* table contains the list of coding tones; v) the *speech_tone* table contains the tones recordings; vi) the *vote_tone* table contains the votes for the tones.



speech_emst	
id	identification number
path	path unto speech file
nop	utterance number
stemi	initial emotional state

emotional_state	
id	identification number
state	emotional state

vote_emst	
id	identification number
path	path unto speech file
stemv	emotional state interpretation
hint	hint
cg_emst	speed of speech
adrIP	IP address of the listener's computer

Fig. 1. The relations between the database table used for emotional states and the table for tones recognition tests.

The tool is developed under the PHP scripting language and uses MySQL5.0 database for preserving and manipulating the voting (assessment) results. The structure of the database created during the vote is shown in Figure 1, for emotions; a similar structure is used for tones. The tool operates as follows:

- Files in the SRoL database of emotional and tonality recordings are randomly selected and presented to the user (listener).
- The listeners examine the records randomly selected and presented to them from the database and save their options by clicking the corresponding buttons.
- The listener options are saved either in the *vote_emst* table, if the page displayed corresponds to the emotional states validation, or in the *vote_tone* table, if the page displayed corresponds to the tone validation.
- After the vote, the program automatically displays the updated confusion matrices, computing the absolute frequencies for all emotional states and tones separately, for each spoken sentence and feminine or masculine speaker.

We took advantage, starting with the MySQL4.1 version, of the subquery option in order to compute the confusion matrix:

```
SELECT e.id, SUM(IF(v.stemv = ‘0’, 1, 0)), SUM(IF(v.stemv = ‘1’,
     1, 0)), SUM(IF(v.stemv = ‘2’, 1, 0)), SUM(IF(v.stemv = ‘3’,
     1, 0)), SUM(IF(v.stemv = ‘7’, 1, 0)), count(*) as total
FROM emotional_state e, vote_emst v, speech_emst s
WHERE e.id = s.stemi AND (v.stemv, s.stemi)
       IN (SELECT ee.id, eee.id
           FROM emotional_state ee, emotional_state eee
           WHERE s.id=v.id AND ee.id=v.stemv AND eee.id=s.stemi)
GROUP BY e.id
```

The *SELECT* statement computes the confusion matrix for the emotional states, meaning: for each emotional state, from the *emotional_state* table is calculated the total number of votes and the number of votes of the expressed class perceived as one of the following emotional states: *neutral tone*, *joy*, *anger*, *sadness, ambiguous*.

To enforce the statistical validity of the tests, the random selection of the recording to be examined by the listeners is based on a random number generator. We used the *mt_rand*(), which is the PHP random function. This function uses a random number generator with known characteristics, based on the “Mersenne Twister” algorithm [15].

4. RESULTS

The number of assessments per sentence and per emotional state or tone is shown in the Table 1. Because of the random character of sentence attribution to a voting listener, some sentences may have a significant larger number of votes than other sentences.

Table 1

Number of assessments for the corresponding file (file numbers have been freely assigned)

Current number of the recording	1	2	3	4	5	6	7	8	9	10	11
Number of votes (assessments)	9	19	23	28	30	25	18	4	6	6	6

The results of recognition of the emotional states and tones are presented in the confusion matrices constituting the Tables 2 to 4. The tables show the number of votes for each emotional state/tone separately. The number of correct predictions is along the main diagonal. The lines contain the number of votes of actual classes, while the columns contain the number of votes perceived for each class. The voting protocol enforces a limited choice of emotions and tones. The listeners may consider that none of the offered voting choices (possibilities determined by the names of the buttons) is correct, in their opinion. In that case, the users have available a box where they can comment. Also, users may consider that the pronunciation is ambiguous. To cope with this situation, the listeners are offered an "escape choice, named "*Ambiguous*". In this way, listeners are not requested to make a forced choice, from a list of predetermined ones.

The voting results show that the *neutral tone* is sometimes confused with the *exclamatory tone*, and sometimes is considered *ambiguous*. The *interrogative tone* is often confused with *exclamatory tone*. Yet, according to Table 2, the *neutral tone* and *interrogative tone* have a large identification degree. Notably, the confusion is by far greater for the sentence *Cine a făcut asta*, a sentence that may be interpreted as either *neutral tone* (in a larger context) or *interrogative tone*. This may imply that the syntax and the semantic information may improve human tones recognition, because the tones of the sentences that are semantically less ambiguous are better recognized.

Table 2

Confusion matrix tone validation

		Perceived class					
		Neutral tone	Exclamatory tone	Interrogative tone	Ambiguous	Total votes	% error
Expressed classes	Neutral tone	**219**	18	3	10	250	12.40
	Exclamatory tone	19	**203**	25	2	249	18.47
	Interrogative tone	22	33	**194**	2	251	22.71

Tables 3 and 4 show that the *neutral tone* is most frequently confused with *sadness* (the total error confusion percentage is 46.32%, from which 33.33% error is *sadness*), but rarely with a*nger, joy, ambiguous*. The emotional state of *joy* is frequently confused with the *neutral tone* (18.14%)/*anger* (14.77%)/*sadness* (9.71%) with a total confusion error of 46%. *Anger* is confused frequently with *neutral tone* (21.10%) and with *joy* (15.61%) and *sadness* (10.55%).

Table 3

Confusion matrix for emotional states validation – absolute frequencies

		Perceived class						
		Neutral tone	Joy	Anger	Sadness	Ambiguous	Total votes	% error
Expressed classes	Neutral tone	**124**	10	8	77	12	231	46.32
	Joy	43	**128**	35	23	8	237	45.99
	Anger	50	37	**121**	25	4	237	48.95
	Sadness	53	19	**9**	**152**	11	244	37.70

According to the results gathered on the web page, *sadness* is most often confused with the *neutral tone*, rarely with *anger/ambiguous*. Also, *sadness* has the highest score of recognition among the emotional states (62.3%). While the emotional state of *joy* and *sadness* are in an antonymic emotional relation, confusion quite often occurs between these two emotional states.

Table 4

Confusion matrix for emotional state validation – percentage representation

		Perceived class				
		Neutral tone	Joy	Anger	Sadness	Ambiguous
Expressed classes	Neutral tone	**53.67**	4.34	3.47	33.33	5.19
	Joy	18.14	**54.00**	14.77	9.71	3.38
	Anger	21.10	15.61	**51.05**	10.55	1.69
	Sadness	21.72	7.79	3.69	**62.30**	4.50

For emotional states the confusion is greater for the sentences *Cine a făcut asta*, and *Ai venit iar la mine*. Our interpretation is that sentences that do not convey enough context for the understanding of the meaning make the emotional interpretation difficult. Indeed, the first sentence can be interpreted as either an interrogative one or a segment from an affirmative sentence; what it actually represents has not been hinted to in the questionnaire. Also, the second sentence, mentioned above, conveys a partial meaning – the listener would expect some consequence, a semantically reasonable continuation of the sentence. The lack of semantic context apparently induces a decrease in the ability of emotion interpretation. For tone recognition, the confusion was grater for the sentence *Cine a făcut asta*.

While we believe that all the respondents to the test have been benevolent and fair in the evaluations, the accuracy of the voting and the statistical significance for a larger population of listeners cannot be guaranteed at any time, because of the public accessibility to the voting Web page format.

Based on the responses gathered until now, the results are as follows. The emotional states and tonalities are recognized with a variable degree of success in Romanian language, similarly to the other languages [1, 4, 5, 7–12, 14]. High recognition scores are obtained for all tones (*neutral*, *exclamatory* and *interrogative*), but only for a single emotional state, namely for *sadness*.

The results also point to the fact that more specific syntactic and the semantic information may improve the human emotion recognition.

5. DISCUSSION AND CONCLUSIONS

We presented a web-based tool to collect results of emotion recognition in the Romanian language, based on recordings in the SRoL voice database. The tool consists in a voting application allowing users at any location to freely listen at randomly selected recordings and to insert their emotion evaluation in a database. The database automatically generates assessment tables and emotion recognition scores and shows them on the screen. The evaluation method has the advantage of using a potentially large number of evaluators to create statistically significant results. This advantage comes with several drawbacks, among others we are unable to check the reliability of the evaluators, we have no knowledge on their ability to assess emotions etc.

The results are relevant for applications like speech synthesis, psychological studies, speech therapy, and phonetic analysis. Assuming that a machine cannot perform much better than humans, at this rate of human speech emotional state recognition (51%–62.3%), creating a reliable automatic speech emotion recognition system is a difficult task. However, tonality information is more reliably recognized, and machines have a better chance to interpret tonality than emotions in the near future.

Acknowledgements. We thank our colleague Monica Feraru for her essential contribution to the development of the recordings database we used and for comments on a preliminary version of this paper. We also thank to all colleagues who agreed to participate in the test, to our colleague Nicolai Curteanu for corrections, and to reviewers for comments.

A preliminary version of this paper has been presented as a poster at the SpeD 2009 – The 5th Conference on Speech Technology and Human Dialogue, Constanța, România, June 18–21, 2009.

REFERENCES

1. ABELIN A., ALLWOOD J., *Cross linguistic interpretation of emotional prosody*, Proceedings ISCA Workshop (ITRW) on Speech and Emotion: A conceptual framework for research, Belfast, 2000, www.ling.gu.se/~abelin/abelin.pdf (accessed 10.03.2009).
2. ENGBERG I.S., HANSE A.V., *Documentation of the Danish Emotional Speech Database*, http://kom.aau.dk/~tb/speech/Emotions/des.pdf (accessed at 10.03.2009).
3. FAKOTAKIS N., Corpus Design, *Recording and Phonetic Analysis of Greek Emotional Database*, https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2004/ LREC/pdf/41.pdf (accessed at 10.03.2009).
4. FERARU M., TRANDABĂȚ D., *Towards the Emotional Annotation of a Corpus for the Romanian Spoken Language*, Academic Days of Iaşi, September 8–10, 2006, Iaşi, România, Proceedings, ISBN 978-973-730-244-1, Ed. Performantica, 43–49.
5. KOSTOULAS TH., GANCHEV T., MPORAS I., FAKOTAKIS N., *A Real-World Emotional Speech Corpus for Modern Greek*, LREC 2008, Morocco, 28–30 May, 2008.
6. MAKAROVA V., MATUSI J., *The production and perception of potentially ambiguous intonation contours by speakers of Russian and Japanese*, http://www.asel.udel.edu/icslp/cdrom/vol3/264/a264.pdf, (accessed at 10.03.2009).
7. PENG G., WANG W. S.-Y., *Tone recognition of continuous Cantonese speech based on support vector machines*, 2005, www.sciencedirect.com.
8. PREDAWAN S., KIMPAN C., WUTIWIWATCHAI C., *Monosyllabic Thai Tone Recognition Using Ant-Miner Algorithm*, International Journal of Computer Science and Network Security, 2007, **9**, *1*, 227–235.
9. TEODORESCU H.-N., *Traces of emotion, intentions and meaning in spoken Romanian*, ASOS workshop, EUROLAN 2007, Iaşi, România.
10. TEODORESCU H.-N., FERARU M., *A study on Speech with Manifest Emotions*, TSD (Text, Speech, Dialogue), Lecture Notes in Computer Science, **4629**, Springer, 254–262, 2007.
11. TEODORESCU H.-N., FERARU M., *The Emotional Speech Section of the Romanian Spoken Language Archive*, 5th European Conference on Intelligent Systems and Technologies, ECIT 2008, July 10–12, 2008, Iaşi, România, ISBN 978-973-730-497-1.
12. TEODORESCU H.-N., FERARU M., TRANDABĂȚ D., *Studies on the Prosody of the Romanian Language: The Emotional Prosody and the Prosody of Double-Subject Sentences*, SpeD 2007 – 4th Conference on Speech Technology and Human Dialogue, Iaşi, România, May 10–12, 2007, ISBN 978-973-27-1516-1, 171–182.
13. *** TEODORESCU H.-N., TRANDABĂȚ D., FERARU M., GANEA R., VERBUȚĂ A., ZBANCIOC M., PISTOL L., *Voiced Sounds of Romanian Language Project*,
www.etc.tuiasi.ro/sibm/romanian_spoken_language/en/voci_emotionale.htm (a),
ww.etc.tuiasi.ro/sibm/romanian_spoken_language/en/fraze_en.htm (b),
www.etc.tuiasi.ro/sibm/romanian_spoken_language/php/vote_stari_emotionale.php (c),
www.etc.tuiasi.ro/sibm/romanian_spoken_language/php/vote_tonalitati.php (d).
14. *** TÓTH Sz. L., SZTAHÓ D., VICSI K., *Speech Emotion Perception by Human and Machine*, Proceeding of COST Action 2102 International Conference, Patras, Greece, October 29–31, 2007. Revised Papers in Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction Springer 2007 (pp. 213–224), 2008, 978-3-540-70871-1, http: //berber.tmit.bme.hu/final/docs/cikkek/Emotion_Patras.pdf (accessed at 07.08.2009).
15. *** *A very fast random number generator*, Mersenne Twister Home Page, http://www.math.sci.hiroshima-u.ac.jp/ ~m-mat/MT/emt.html (accessed at 10.03.2009).

16. *** *The Berlin Database of Emotional Speech*, http://pascal.kgw.tu-berlin.de/emodb/ (accessed at 10.03.2009).
17. *** *The Ethic Principles of the American Acoustic Association regarding Research Involving Human Subjects*, http://asa.aip.org/ethical.html.
18. *** *The US Food and Drug Administration's Protocol of Protection of Human Subjects*, http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/default.htm.

Author contributions: H.N.T. and L.P. equally contributed to this work. H.N.T. designed research and specified the application; L.P. implemented the web tool; both authors contributed to correcting the tool – also advised by colleagues, to data analysis, and to writing the paper.

Present address: Academia Română, Filiala Iaşi, Carol I nr. 8, Cod 700505, Iaşi, România.

To whom correspondence should be addressed. E-mail: hteodor@etc.tuiasi.ro

This article contains supporting information online at:
http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/voci_emotionale.htm
http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/lucrari_sit.htm

Received August 10, 2009

Institute of Computer Science, Romanian Academy, Iaşi Branch
22A Copou, Iaşi 700505, Romania
E-mail address: laura.pistol@iit.tuiasi.ro, hteodor@etc.tuiasi.ro