THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 8, Number 2/2007, pp. 000-000

ANALYZING EMOTIONS IN SPOKEN ROMANIAN

Horia-Nicolai TEODORESCU [*], Monica FERARU [**]

[*] Institute for Computer Science, Romanian Academy, Bd. Carol I nr 8, Iasi, Romania
[**] Technical University of Iasi, Iasi, Romania
Corresponding author: Horia-Nicolai Teodorescu, hteodor@etc.tuiasi.ro

We present a study of the prosody – seen in a broader sense – that supports the theory of the interrelationship function of speech. "Pure emotions" show a relationship of the speaker with the general context. The analysis goes beyond the basic prosody, as related to pitch values, pitch trajectory, sound duration and pauses; the analysis also aims to determine the change in higher formants. The refinement in the analysis asks for finer tools. Methodological aspects are discussed, including limitations of the currently available tools. Some conclusions are drawn.

Key words: Romanian spoken language, prosody, emotions, formant analysis

1. BASES OF OUR APPROACH

A challenge in artificial intelligence development is the emotional interpretation of the interaction between human and computer. Speech is a subtle and rich communication; it transfers not only the linguistic information, but also information about the personality and the emotional state of the speaker. The emotion is a motivation answer adapted to the social environment. The prosody is defined like "the rhythm and intonation aspects of a language" while the intonation represents "the expression manner; raising and lowering of the pitch in voice" (Merriam-Webster). Containing information about the speaker and about the environment, the prosody is a communication manner which includes the attitude, the emotions etc. [8].

According to Fakotakis [3], emotions are classified in "basic" emotions, with different intensity levels, and in "non-basic" emotions (the "mixed" emotions). In 1992, Johnson-Laird and Oatley stated that there are five basic emotions; in 1998, after other studies, they concluded that there are four basic emotions: happiness, anger, sadness and fear.

In order to create an emotional database, it is necessary to have a number of speakers who try to simulate the emotions in different contexts [2]. A different set of subjects listen to the recordings and try to identify the emotion that the speaker has tried to simulate. The experimental analysis of Buluti, Narayanan and Syrdal [1] showed that the emotion's recognition is not perfect for emotions like sorrow, sadness, happiness, and the neutral tone. Their recognition rate was 92.1% for the neutral tone, 89.1% for sorrow, 89.7% for sadness, and 67.3% for happiness. It is important to notice that the recordings have been made by professionals (actors in most cases) with professional voices [9].

The research, as envisioned by the first author, has been aimed to verify several speculative hypotheses, namely:

- The emotions are represented by a complex characteristic of the voice; that complex goes beyond the accepted set of characteristics: duration of phonemes, duration of pauses, and pitch trajectory, including the first formants, moreover the higher formants and the subtle mixture of linear and nonlinear processes of speech generation.
- The same emotion is represented differently in presence of different interlocutors, depending on the relationship the speaker has with the interlocutor.
- If the inter-relationship theory of the prosody proposed in [8] were to be true, some of the syntactical constructions of the language should be specifically aimed to the inter-relationship promotion; therefore, such constructions have to be looked for, determined and analyzed. One of these constructions,

identified by the first author as possibly related in a specific way to the emotion and inter-relationship representation, is the double-subject construction.

2. EXISTING APPROACHES

The emotional speech analysis is a very challenging research field, as shown by the increasing number of databases dealing with the voice characteristics in different emotional contexts. We briefly present in this section several databases of emotional voice: the Greek database [3], the German archive [6], the Danish [5] and the Spanish databases [4].

The recordings from the Greek emotional database [3] were made in a professional studio, in Athens, and the speakers were actors. The recordings were made on 16 bits, mono, with a sampling frequency of 44,1 kHz. The goal of this research was to improve the naturalness of synthesized voice. The recordings were made in three different contexts:

- in order to reflect the reaction of the speaker to a concrete stimulus (authentic emotion);
- preparing the environment in order to help psychologically the speaker to simulate the indicated emotion;
- simulating the emotions only by imagining a context.

Several sentences were selected from the press by linguists. The corpus included 10 words, 20 short phrases, 25 long phrases, and 12 fragments of fluent speech. The study was oriented towards the evaluation of the simulated emotional states by free answers (86.88%) and false answers (89.63%).

The German emotional database [6] contains six basic emotions: anger, happiness, fear, sadness, disgust, boredom and neutral tone. The recordings in that study included five short sentences and five long sentences; the recordings have been realized by 10 professional actors, 5 women and 5 men in a special room. Over 800 recordings have been made (7 emotions × 10 sentence × 10 actors × 2 versions). The validation commission, composed by 20-30 persons, listen the phrases and recognized 80% of the simulated emotional states. The recordings were made on 16 bits, mono, with a sampling frequency of 16 kHz. The database contains files with sentences and words, files with syllables and phonemes, in wave format, the information about the results of the perception tests (recognized emotion, evaluation of natural language, the power of emotion), and the results of measuring the fundamental frequency, the energy, the duration, the intensity and the rhythm.

The Danish database [5] contains recordings of two words, nine sentences and two fragments of fluent speech, simulating happiness, surprise, sadness, anger and neutral tone, spelled by four professional actors. The emotional states were recorded in a room of the Aarhus theatre. The emotions were correctly recognized in a proportion of 67%. The happiness state was mostly confused with surprise and the sadness state was confused with the neutral tone. 75% of the people listening to the recordings said that it was difficult to identify the recorded emotions. The recordings were made on 16 bits, mono, with a sampling frequency of 16 kHz. Each recording has attached video information along the voice signal. The database also contains information about the profile of the speaker, like weight, height, sex, how long he/she worked like an actor etc. To increase the consistency of the assessment of the emotion conveyed by the speech, researchers at the University of Aolborg use a questionnaire for the assessing persons. The questionnaire includes questions such as “how the emotions identification seems like”, “what are the factors which bring to the correct identification of the emotions”, and “additional remarks regarding the recordings”.

The Spanish database [4] contains recordings with the following emotional states: happiness, desire, fear, fury, surprise, sadness and disgust; the sentences have been pronounced by eight actors. Every speaker recorded the text three times, with various levels of intensity of the emotions. The validation of the recorded emotions was made by a test based on the questions: “mark the emotion which was recognized in each recording”, “mark the credibility level of the speaker”, and “specify if the emotional state was recognized and at what level”. The speakers have made 336 recordings; only 34 were selected, analyzed, and validated through the tests made by 1054 persons. The goal of the study was to describe a useful methodology in the validation of the simulated emotional states. They obtained a set of rules which describe the behavior of the important parameters of the speech associated with emotions. The obtained results are useful in generating synthesized speech with emotion. The analyzed parameters were fundamental frequency, time and rhythm.

The researchers represented graphically the wave form, the pitch contour and the energy. They obtained the following characteristics of the emotional modulation [4]:
"- in happiness state: increase of the average tone, increase of the variability of the tone, quick modulations of the tone, […] stable intensity, decrease [of] the silence time; […]
- fury state: variation of the emotional intonation structure, short number of pauses, increase of the intensity from beginning till the end, variation of timber, increase of the energy; […]
- sadness state: decrease of the average tone, decrease of the variability of the tone, no inflexions of the intonation, decrease of the average intensity."

The analysis quoted above is somewhat subjective and leaves unanswered many questions on the variation of objective parameters, like formants, from one emotion to another. In our research reported here, we specifically address the characterization of emotions using the objective parameters for the states reported in [4]. We also contrast the characteristics of the voice for the above emotions with the normal (i.e., no emotion) speech. The comparison of our results with the results reported in [4] may help identify inter-language variations for the emotional speech.

3. THE METHODOLOGY

3.1. Recording protocol

The database contains short sentences or phrases fragments, with different emotional states. The emotions are: sadness, happiness, anger and detaching state. The files are classified in A class (feminine voice) and in B class (masculine voice). The speakers are persons with age between 25-35 years, born and educated in the middle area of Moldova (Iasi, Vaslui, Bacau), with higher education and without manifested pathologies.

The recordings were made using the GoldWave™ application, with a sampling frequency of 22050 Hz. Every speaker pronounced the sentence for three times, following the recording protocol. The sound was saved in .wav, .ogg, .txt format on 16 and 24 bits. The persons 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 database [10]contains two types of protocols:

- the recording protocol, containing information about the noise, the microphone used, the soundboard, and the corresponded drivers;
- the documentation protocol, which contains the speaker profile (here we find linguistic, ethnic, medical, educational, professional information about the speaker), and a questionnaire regarding the healthy state of the speaker.

3.2. The emotional speech database and processing tools used in the analysis

We recorded a set of sentences with emotional states. The sentences are: 1. *Vine mama*. (Mother is coming) 2. *Cine a facut asta*. (Who did that?) 3. *Ai venit iar la mine*. (You came back to me) 4. *Aseara*. (Yesterday evening). Only two emotions are discussed here, namely happiness and sadness. The consistency of the emotional content in the speech recordings has been verified by several listeners; the emotion confusion matrix has proved that all emotions are correctly identified, with a rate of more than 80% by the listeners.

Today, there is no standard model for the emotional annotation process [7]. The sentences have been annotated using the Praat software [13] at several levels: phoneme, syllable, word and sentence. In this paper, the analysis was made only for the sentence "Vine mama" pronounced by five persons, three times each. Our goal has been to make the difference between happiness and sadness emotional states. For that purpose, we computed the values for the formants and the duration of the vowels using several tools: Praat [14], Klatt analyzer [15], GoldWave [16], and Wasp [17]. The purpose is to discriminate based on these values, the emotional states of happiness and sadness. We obtained general and particular rules which are

discussed in the section on results. We have been confronted with several problems in determination of the formants, namely with large disagreements between values provided by different applications. For example, there were cases where according to Praat™, on some segments there the fundamental frequent is not defined (see figure 1) while Wasp™ or Klatt analyzer™ identifies a pitch on those segments. Using the Klatt analyzer™, we could not see the F1 formant for vowel "*i*", as presented in figure 2.

Notice that, in figures 3 and 4, it is difficult to visually determine the formants in these spectrograms using the GoldWave™ and Wasp™ applications. The difficulties are largely due to the imprecision of the definitions of the pitch and of the formants, especially for non-stationary signals. The nonlinear behavior of the phonatory organ, which are well documented in the medical literature as well as in the recent info-linguistic literature, [11] [12], determines a lack of significance of the parameters defined in the frame of the linear theory of speech analysis.

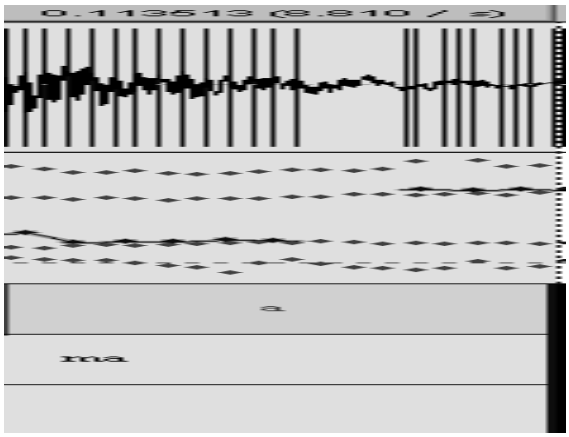

Figure 1. Determination of the F0 with Praat™ application

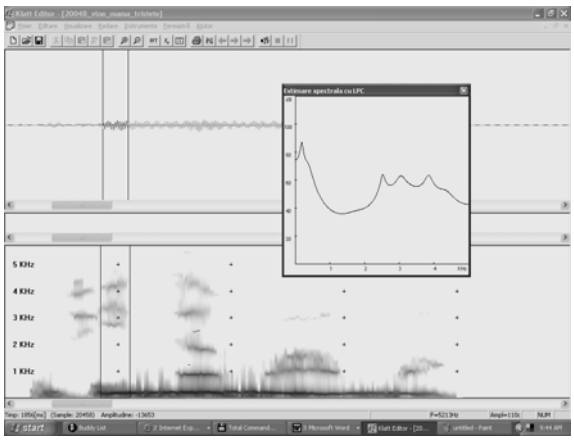

Figure 2. Determination of the formants with Klatt analyzer™

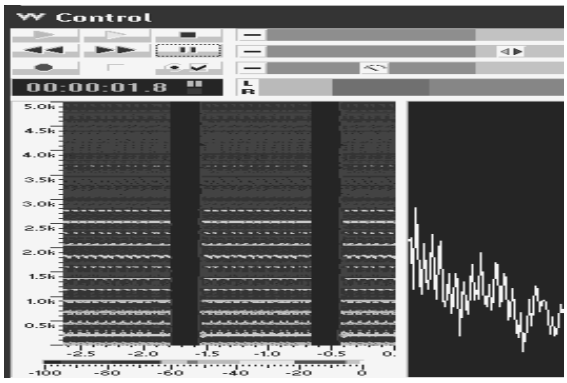

Figure 3. Determination of the F0 according to GoldWave™

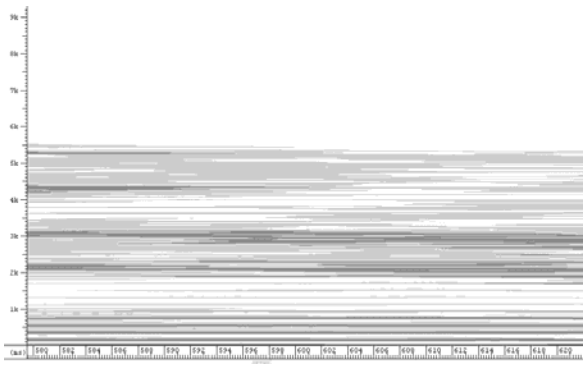

Figure 4. Determination of the formants according to Wasp™

The differences in the results obtained with various tools reflect the theoretical limits of the formant parameters, as well as the capabilities of the various approximation methods used in the tools. These inconsistencies are one reason why the results we report should be considered preliminary, although we made every effort to obtain the results according to the best present knowledge.

3. RESULTS OF THE ANALYSIS OF SPEECH WITH MANIFEST EMOTIONS

The main general rules that we obtained based on the reduced number of cases we analyzed are listed below. The results are shown in the next tables and our notations used are: "-" means decrease the obtained values in sadness compared with happiness, "+" means increase, "±" means fluctuant, i.e., no conclusion can be derived. The symbol "a1" represents the first "a" vowel in the word "mama" and "a2" represent the second "a" in the same word.

- The obtained values for the F0 formant for all the persons decrease in sadness state compared with the happiness state (table 1). We notice that the Klatt analyzer™ application "is not seeing" the F1 formant, we can distinguish easily with the application GoldWave™ the F0 and F1 formants, and with the application Wasp™, the F2 and F3 formants.

Table 1. The variations of the formants F0, F1, and F2 for the five persons (- =increase, + =decrease, ± =fluctuant)

Subject	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	a1	a2
20048f	-	-	-	-	-	-	-	±	-
01312f	-	-	-	+	-	±	-	-	-
55555f	-	-	-	±	±	-	+	-	-
123456f	-	-	-	-	-	-	-	-	±
77777m	-	-	-	-	-	-	-	-	-

- The accentuated vowels (like the vowel "i" from the word "vine" and the vowel "a", first "a" from the word "mama") don't offer important information compared with non-accentuated vowels (vowel "e" from the word "vine' and vowel "a", the last from the word "mama").
- The vowel "i" from the word "vine" has random values of the formants; therefore, it does not help in the emotion recognition.
- The obtained values of the formant F2, for the vowel "a" (the last "a" from the word "mama") decrease in sadness compared with happiness and the values of the F1 formant have the tendency to decrease too.
- The obtained values of the formants F1 and F2, for the vowel "a" (the first "a" from the word "mama") have the tendency to decrease in sadness compared with happiness states.
- The obtained values of the formants F2, for the vowel "e" (from the word "vine") have the tendency to decrease in sadness compared with happiness states.

The particular rules obtained are:

- For the person 20048f (table 2 and table 3), the values for the F1 formant have the tendency to decrease in the sadness state compared with the happiness state, for all the considered vowels.

Table 2. The values of formants, in Hz, using Wasp™ application for sentence "Vine mama", person #20048f

Parameters	Happiness			
	i	e	a1	a2
F0	100-200	300-400	200-300	200-400
F1	400-600	600-800	600-700	800-1000
F2	2400-2600	900-1000	900-1000	1400-1600
F3	3300-3500	1800-2200	2500-2600	3000-3300

Parameters	Sadness			
	i	e	a1	a2
F0	100-200	200-300	150-250	200-300
F1	300-400	600-700	600-700	800-900
F2	2200-2400	1800-2000	1200-1400	1300-1500
F3	3000-3200	3000-3200	2900-3200	3200-3400

- The values of F2 formant for the vowel "e" (from the word "vine") and for the vowel "a" (the last "a" from the word "mama") have the tendency to decrease in the sadness state compared with the happiness state.
- For the "a" vowel (the first "a" from the word "mama") the values of F2 formant decrease according to GoldWave™, increase according to Praat™, and are constant with Wasp™ and Klatt analyzer™.

Table 3. The values of formants, in Hz, using Klatt™ application for sentence "Vine mama", person #20048f

Parameters	Happiness			
	i	e	a1	a2
F0	264	352	264	176
F1	???	951	951	1021
F2	2659	2290	1444	1462
F3	3293	3311	3311	3223

Parameters	Sadness			
	i	e	a1	a2
F0	176	229	176	141
F1	???	845	933	1021
F2	2747	1955	1374	1444
F3	3399	3082	1744	3276

- For the person 55555f (table 4), the values of the F1 formant for the vowel "a" (the last "a" from the word "mama") decrease in sadness compared with happiness, except for the values obtained with Praat™.
- For vowel "e" and the first "a" (from the word "vine") the values of F1 seem constant according to GoldWave™ and Klatt analyzer™, decrease according to Wasp™ and increase in Praat™.
- For the last "a", the values of F1 formant decrease according to all applications except for Praat™, where the values for F1 increase.

Table 4. The variations of the formants F0, F1, and F2 for the person #55555f

55555f	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	A1	a2
GoldWave™	-	-	-	±	±	-	+	±	-
Wasp™	-	-	-	-	-	-	+	-	-
Klatt analyzer™	-	-	-	±	±	-	±	-	-
Praat™	-	-	-	+	+	+	-	-	+

- The values of F2 formant for the vowels "a" (the first and the last "a" from the word "mama") decrease in the sadness state compared with happiness state according to Wasp™ and Klatt analyzer™ and increase in Praat™ for the last "a".
- For vowel "e", the values of F2 formant increase according to GoldWave™ and Wasp™ application and decrease according to Praat™ application.
- For the person 01312f (table 5), for the vowel "e", the values of F1 formant increase according to GoldWave™, Praat™ and Klatt analyzer™ application and decrease according to Wasp™ application.

Table 5. The variations of the formants F0, F1, and F2 for the person #01312f

01312f	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	a1	a2
GoldWave™	-	-	-	+	-	±	-	-	-
Wasp™	-	-	-	-	±	±	+	-	-
Klatt analyzer™	-	-	-	+	-	-	-	+	-
Praat™	-	-	-	+	-	-	-	-	-

- The values of F2 formant for vowel "e" have the tendency to increase according to Wasp™ application and to decrease according to GoldWave™, Praat™ and Klatt analyzer™ application.
- For the vowel "a" (the first and the last "a" from the word "mama"), the values of F2 formant have the tendency to decrease in the sadness state compared with happiness state, except for the values obtained with Klatt analyzer™ for the first "a".
- For the person 123456f (table 6), the values of F1 formant for vowel "e" and the two vowels "a" decrease according to all applications.

Table 6. The variations of the formants F0, F1, and F2 for the person #123456f

123456f	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	a1	a2
GoldWave™	-	-	-	-	-	-	+	-	-
Wasp™	-	-	-	-	-	-	-	±	-
Klatt analyzer™	-	-	-	-	-	±	-	-	+
Praat™	-	-	-	-	-	-	-	-	+

- For the first "a" and the vowel "e", the values of F2 formant have the tendency to decrease in the sadness state compared with happiness state, except for the values obtained with GoldWave™ for the vowel "e" and the values obtained with Wasp™ for "a".
- The values for the last "a" decrease with GoldWave™ and Wasp™ and increase with Praat™ and Klatt analyzer™.
- For the person 77777m (table 7), the values of F1 formant for all the vowels decrease with all four applications except for the values obtained with Praat™ for the second "a".
- The values of F2 formant for all the vowels decrease with all four applications except for the values obtained with Praat™ for the second "a".
- The values of the F0 formant for the second "a" is undefined according with Praat™.

Table 7. The variations of the formants F0, F1, and F2 for the person #77777m

77777m	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	a1	a2
GoldWave™	-	-	-	-	-	-	-	-	-
Wasp™	-	-	-	-	-	-	-	-	-
Klatt analyzer™	-	-	-	-	-	-	-	-	-
Praat™	-	-	undefined	-	-	+	-	-	+

5. CONCLUSIONS

The reported research had the general but somewhat diffuse aim of determining whether there are prosodic features that support the interrelationship theory of language. The choice of the linguistic and paralinguistic features selected for the analysis has been motivated by the analysis of manifest and intentional emotions

For sentences uttered with manifest emotional charge in the Romanian language, we found that most informative regarding the emotions is the change of the pitch. This conclusion is compatible with some findings reported for other languages. In contrast, we found that the accented vowels do not carry significantly more emotional information than the non-accented vowels; rather, the opposite is true. This conclusion is a departure from findings by other authors, for different languages. We need to further analyze this issue to validate it for a larger number of sentences and subjects. We also found that some higher formants, F1 and F2, in both accented and non-accented vowels, are also essential in conveying emotional information.

Regarding the available speech analysis tools, we conclude that no tool provides irrefutable results. While we used four tools and compared the results, no one is significantly better than the others. We have indicated a methodology to choose a stable section of the vowels for the analysis, to improve consistency in measurements, but even using this methodology, the lack of good formant extractors restricts today possibilities to obtain high confidence in the results.

Overall, the individual findings and conclusions reached in this research support the theory of interrelationship expression through prosodic and paralinguistic information.

ACKNOWLEDGMENTS

We acknowledge the partial support of the Romanian Academy "priority research" grant "Sisteme cognitive" and of the CEEX grant "Sistem automat de diagnostic paraclinic in sindromul disfunctional al sistemului stomatognat".

REFERENCES

1. BULUTI, M., NARAYANAN, S.S., SYRDAL, A.K. *Expressive speech synthesis using a concatenative synthesizer*, Proc. ICSLP, 2002
2. DOUGLAS-COWIE, E. , CAMPBELL, N., COWIE R. and ROACH, P. *Towards a new generation of databases*, Speech Communication, vol. **40**, pp. 33-60, 2003
3. FAKOTAKIS, N., *Corpus design, recording and phonetic analysis of Greek emotional database*, Proc. The Language Resources and Evaluation Conference, LREC'04, 2004.
4. IRIONDRO, I., GAUS, R., RODRIGUES, A., et al., *Validation of an acoustical modeling of emotion expression in Spanish using speech synthesis techniques* (http://serpens.salleurl.edu/intranet/pdf/239.pdf)
5. http://kom.aau.dk/~tb/speech/Emotions/des.pdf
6. http://pascal.kgw.tu-berlin.de/emodb/
7. ODERLMAN, R., POEL, M., HEYLEN, D., *Emotion annotation in the AMI project*, February, 2005
8. TEODORESCU, H.N., *A proposed theory in prosody generation and perception: the multi-dimensional contextual integration principle of prosody*, Proc. The 3rd Conference on Speech Technology and Human-Computer Dialogue "SpeD 2005", Cluj-Napoca, Romania, May 13-14, 2005, Burileanu, C. (Coordinator), Trends in Speech Technology, Romanian Academy Publishing House, Bucharest, Romania, ISBN 973-27-1178-7, pp. 109-118, 2005
9. TEODORESCU, H.N., FERARU, M., TRANDABAT, D., *Nonlinear Assessment of the Professional Voice "Pleasantness"*, Biosignal, 28-30 June 2006, Brno, Czech Republic, pp. 63-66, 2006
10. TEODORESCU, H.-N., TRANDABĂȚ, D., FERARU, M., GANEA, R., A. VERBUȚĂ, M. ZBANCIOC, HNATIUC, M., *Proiectul Sunetele Limbii Române*, http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm
11. LOSCOS, A., BONADA, J., *Emulating Rough and Growl Voice in Spectral Domain*, Proc. of the 7th Int. Conference on Digital Audio Effects (DAFX-04), Naples, Italy, October 5-8, 2004. (http://www.iua.upf.es/mtg/publications/DAFX04-aloscos.pdf, accessed 12 Nov. 2006)
12. SUN, X., *Pitch Determination and Voice Quality Analysis Using Subharmonic-to-Harmonic Ratio*. http://www.ling.northwestern.edu/~jbp/sun/sun02pitch.pdf. (accessed 12 Nov., 2006)
13. www.praat.org
14. www.praat.org
15. www.speech.cs.cmu.edu/comp.speech/Section5/Synth/ klatt.kpe80.html
16. www.goldwave.com
17. www.wasp.dk

Received February 5, 2007