A Study on Speech with Manifest Emotions

Horia-Nicolai Teodorescu[1,2], Silvia Monica Feraru[2]

[1] Institute for Computer Science, Romanian Academy, Bd. Carol I nr 8, Iasi, Romania
[2] Technical University of Iasi, Iasi, Romania
{hteodor, mferaru}@etc.tuiasi.ro

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

Keywords: Romanian spoken language, emotional states, formant analysis

1 Introduction

The study of spoken languages has significantly progressed during the last decades, due to the dramatic increase of interest fuelled by applications like virtual reality, video-games, human-computer speech interaction, security, and medical applications. 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-related answer adapted to the social environment. The prosody is a communication means which includes the attitude and the emotions [7]; it also contains information about the speaker and about the environment. In Section 2, we review the state of the art in emotional speech analysis. In Section 3, we present the essentials of our research methodology. In Section 4, we summarize the results. The final section includes a brief discussion and conclusions.

2 Existing Approaches

For comparison with our approach, we briefly present in this section several researches on emotional voice and related databases, for the languages Greek [3], German [6], Danish [5], and Spanish [4]. Numerous other databases, for other languages, exist, but we limit our presentation to the ones described below because they are quite different and illustrate well the variety of approaches in the literature.

There is no general agreement on the classification of emotions. According to [3], emotions are classified in “basic” emotions, with different intensity levels, and in “non-basic” emotions (the “mixed” emotions). Without entering details, we chose four “basic” emotions for our analysis, happiness, anger, sadness, and neutral tone. At least some of these emotions (furry/anger and joy/happiness) are known to be produced at the level of the limbic system; this means that they are “elementary” emotions, not states of the mind produced together by several mind processes occurring in various sites of the brain.

The creation of an emotional database requires a number of speakers who simulate the emotions in different contexts [2]. A different set of subjects (the evaluators) listen to the recordings and seek 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, joy, and the neutral tone. The recognition rate was 92% for the neutral tone, 89% for sorrow, 89.7% for sadness, and 67.3% for happiness. The recordings have been made by actors, in most cases, or by persons with professional voices [12].

The recordings for the Greek database reported in [3] were made by actors in a professional studio. The goal of that research was to improve the naturalness of synthesized voice. The recordings were made in three different contexts, namely i) in order to reflect the reaction of the speaker to a concrete stimulus (authentic emotion); ii) preparing the environment in order to help psychologically the speaker to simulate the indicated emotion; iii) simulating the emotions only by imagining a context. The study was oriented towards the evaluation of the simulated emotional states by free answers (86.9%) and false answers (89.6%).

The German emotional database [6] contains six basic emotions: anger, happiness, fear, sadness, disgust, boredom and neutral tone. The recordings have been realized by professional actors. The validation commission recognized 80% of the simulated emotional states. The database contains files with sentences and words, the results of the perception tests, 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 professional actors. The emotional states were recorded in a theatre room with an excellent acoustic. The emotions were correctly recognized in 67% of the cases. The happiness state was mostly confused with surprise; 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. Each voice recording has attached video information. The database also contains information about the profile of the speakers. To increase the consistency of the assessment of the emotion in the speech, the researchers used a questionnaire for the assessing persons; questions such as “how the emotions identification seems like”, “what are the factors which bring to the correct identification of the emotions”, etc. help fine-tuning the assessment.

The Spanish database [4] contains recordings with seven emotions: happiness, desire, fear, fury, surprise, sadness, and disgust; eight actors have pronounced the sentences. Every speaker recorded the every sentence for 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 goal of the study was to describe a useful methodology in the validation of the simulated emotional states. The quoted researchers derived a set of rules describing the behavior of the main parameters of the emotional speech, in view of synthesizing emotional speech. The analyzed parameters are the fundamental frequency trajectory, time, and rhythm. They obtained the following characteristics of the emotional modulation [4]: i) for the joy 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; […]"; ii) for 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; […]"; iii) 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 above-quoted analysis leaves many unanswered questions on the variation of objective parameters, like formants, from one emotion to another. In the 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 Methodology, Database, and Analysis Tools

We place a high emphasis on the methodology of acquiring and analyzing emotional speech; this justifies the length of this section. In the first place, while we value the use of dramatic actors to produce emotional speech, we argue that speech by "normal" people should be the primary focus of a sound research in the field, if it were to obtain results for everyday applications. The database contains short sentences or phrases fragments, with different emotional states. We recall that the emotions investigated are sadness, joy, fury and neutral tone. The files are classified in class A (feminine voice) and class B (masculine voice). The speakers are persons aged between 25-35 years, born and educated (higher education) in the middle area of Moldova (Romania), without manifested pathologies. The recordings use a sampling frequency of 22050 Hz, 24 bits. Every speaker pronounced the sentence for three times, following the recording protocol. The persons have been previously informed about the objective of the project and they signed an informed consent in accordance with to the Protection of Human Subjects Protocol of the U.S. Food and Drug Administration and with Ethical Principles of the Acoustical Society of America.

The database contains two types of protocols, namely the recording technical protocol and the recording documentation one. The recording protocol contains information about the noise, the microphone used, the soundboard etc. The documentation protocol contains information on the speaker's profile – linguistic, ethnic, medical, educational, and professional information; for details, see [10].

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. *Aseară*. (Yesterday evening). The consistency of the emotional content in the speech recordings has been verified by the evaluators; the emotion confusion matrix has proved that all emotions are identified with a rate of more than 67% by the evaluators.

Today, there is no standard model for the emotional annotation process [13]. The sentences have been annotated using the Praat™ software (www.praat.org) at several levels: phoneme, syllable, word, and sentence. In this paper, the analysis refers only to the sentences "Aseară" and "Vine mama", as pronounced by eight persons, three times each, i.e., a total of 192 recordings. The values of the formants were determined using four tools: Praat™, Klatt analyzer™ (www.speech.cs.cmu.edu/comp.speech/Section5/Synth/klatt.kpe80.html), GoldWave™ (www.goldwave.com), and Wasp™ (www.wasp.dk). Every tool produces a value for each formant. We compared the obtained values for the emotional states. In case where three out of four values are increased, the conclusion is that the values of the formants increase. Where there are clear discrepancies between subjects or between results obtained with different tools, we cannot draw any conclusion and we say that the values of the formants are fluctuant. There are cases when the tools (e.g. Praat) cannot determine the value of the formant (see the table 1).

Table 1. The values of F0 [Hz] obtained with several tools, for the one-word utterance „Aseară", for the states *happiness* and *sadness* (person # 77777m)

Tools	Happiness		
	F0 / a	F0 / ea	F0 / ă
GoldWave™	100-150	100-200	400-500
Wasp™	117	166	467-490
Klatt analyzer™	106	176	476
Praat™	112	165	481

Tools	Sadness		
	F0 / a	F0 / ea	F0 / ă
GoldWave™	80-150	100-150	100-150
Wasp™	78	77	74
Klatt analyzer™	88	123	106
Praat™	undefined	81	479

We have been confronted with several problems in the determination of the formants, namely with large disagreements between values provided by different applications. According to Klatt Analyzer™, the F1 formant for vowel *i* is "missing". Notice that, sometimes, it is difficult to determine visually the formants in the spectrograms using the GoldWave™. 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 is well documented in the medical literature as well as in the recent info-linguistic literature, [8], [9], determines a lack of significance of the parameters defined in the frame of the linear theory of speech analysis.

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.

With respect to other voice databases, we are insisting on some methodological aspects, namely the using of "natural voices", i.e. non-artist speakers which show "everyday emotions", and the use of inter-validated tools to determinate the formants and intra-validated (per speaker) emotional utterance.

4 Results of the Analysis of Speech with Manifest Emotions

The main results obtained in the analysis are listed in the Tables 2-5. The main rules we obtained, based on the results on the analyzed eight subjects, are listed at the end of this Section. In the tables, "-" means a decrease of the obtained values in first emotion, compared to the second emotional state; these couples of states can be happiness compared with sadness, fury compared with sadness, happiness with fury, and any emotion compared to normal tone. Also, "+" means an increase, while "±" means fluctuant, i.e. no conclusion can be derived. The "a", "ea", and "ă" represent the first vowel, the diphthong, respectively the last vowel in the word "aseară".

As a general conclusion, the states happiness and sadness, on one side, and fury from sadness, on the other side can be easily distinguished in all cases. It is more difficult to distinguish between happiness and fury. The utterance analyzed in the tables 2, 3, and 4 is "Aseară"; in table 5, the utterance is "Vine mama".

Table 2. The tendency for the F0, F1, F2 formants for the eight persons (- =increase, + =decrease, ± =fluctuant) for the states *happiness* and *sadness* versus *happiness* and *fury*

Subject	F0			F1			F2		
	a	ea	ă	a	ea	ă	a	ea	ă
20048f	-/±	+/+	+/±	±/±	±/+	±/-	+/+	+/+	+/+
01312f	±/+	+/+	+/±	+/±	+/±	+/±	+/-	+/+	+/+
55555f	±/-	+/±	-/-	+/-	+/-	-/-	+/±	+/±	±/±
123456f	-/-	+/+	+/+	+/±	+/+	+/+	+/+	+/+	+/+
77777m	+/±	+/±	+/+	+/+	+/-	+/+	±/+	+/+	+/+
263315m	+/±	+/±	±/±	+/-	+/±	-/-	+/+	+/-	+/±
14411f	±/±	-/-	-/±	-/-	-/-	+/-	-/-	±/±	-/-
26653m	±/±	-/±	±/±	-/±	-/±	-/-	-/-	-/-	±/-
General	±/ ±	+/ ±	partly+/ ±	partly+/ ±	partly+/ ±	±/ -	partly+/ partly+	+/ partly+	partly+/ partly +

For the utterance "Aseară", the obtained values for the F0, F1, F2 formants of the diphthong "ea" for all the persons increase in the happiness compared with sadness state; the values for the F1, F2 formants of the first vowel "a" and for the F2 formant of the last vowel "ă" in the word "aseară" increase too (see table 2).

Table 3. The tendency for the F0, F1, F2 formants for the seven persons, for *fury* and *sadness*

Subject	F0			F1			F2		
	a	ea	ă	a	ea	ă	a	ea	ă
20048f	-	±	±	±	-	+	-	+	±
01312f	-	+	+	+	+	+	+	+	+
55555f	+	+	-	+	+	±	+	+	±
123456f	±	±	+	+	+	±	-	+	+
77777m	+	+	+	+	+	±	±	+	±
263315m	+	+	±	+	+	±	+	+	+
14411f	+	-	±	-	±	+	-	±	-
General	+	partly+	±	partly+	partly+	±	±	+	±

For the utterance “Aseară”, the obtained values for the F0 formant of the diphthong “ea” for all the persons increase in fury state compared with neutral tone; the values for the F1 formant of the first vowel “a”, of the diphthong “ea” in the word “aseară” increase too (table 3). Notice that, from table 2, no significant conclusions can be derived related to (joy; furry), except that the values for the formants are fluctuant (table 2); thus, the states (joy; furry) can not be reliably distinguished.

For the sentence “Aseară”, the obtained values (table 4) for the F0 and F2 formants of the diphthong “ea” and, partly, of the last vowel “ă”, for all the persons, increase in the couple (happiness; neutral tone); the values for the F1 and, partly, F2 formants of the first vowel “a” respectively of the diphthong “ea” increase too. Table 5 shows, for the sentence “Vine mama”, the same problems as table 2, with regard to the fluctuations of the formants values.

Table 4. The tendencies for the F0, F1, F2 formants for the seven persons, for *happiness* and neutral tone

Subject	F0			F1			F2		
	a	ea	ă	a	ea	ă	a	ea	ă
20048f	-	+	+	+	±	±	+	+	+
01312f	±	+	±	±	+	+	-	+	+
55555f	-	+	±	+	+	±	±	±	+
123456f	-	+	+	+	+	±	+	+	+
77777m	+	+	+	+	±	+	+	+	+
263315m	±	+	-	±	±	-	+	±	+
14411f	±	-	±	-	-	-	-	±	-
General	±	+	±	+	±	±	+	+	+

Notice in table 5, that the obtained values for the F0 formant of the vowels “e” in the word “vine”, the first “a” (a1) and the last “a” (a2) in the word “mama”, increase in the happiness state compared with sadness state, for all the persons.

Comparing the results obtained on both phrases, the additional rules are derived.

- The obtained value for the F0, F1, and F2 formants in the couple (happiness; sadness state) increase in both situations, but the increasing tendency is more evident in the case of the sentence “Vine mama”. The values for the F0 formant increase for all subjects, while for the sentence “Aseară” there is a subject not obeying the rule (see tables 2 and 5).

- In the couple (happiness; neutral tone), the fluctuations of the values for the formants are similar for the “Aseară” and “Vine mama” sentences.
- In the couple (fury; sadness), the obtained results in the case of “Vine mama” are more fluctuant compared with the results for the sentence “Aseară”.
- In the couple (fury; neutral tone), the obtained results in the case of “Aseară” are more fluctuant compared with the results for the sentence “Vine mama”.

Table 5. The tendencies for the F0, F1, F2 formants for the six persons, for *happiness* and *fury* versus for *happiness* and *sadness*

Subject	F0			F1			F2		
	e	a1	a2	e	a1	a2	e	a1	a2
20048f	+/+	+/+	+/+	-/+	-/+	+/+	-/+	-/-	+/+
01312f	+/+	±/+	+/+	-/-	+/+	±/+	±/±	+/+	+/+
55555f	-/+	±/+	±/+	-/±	±/-	-/+	-/±	±/+	±/+
123456f	+/+	+/+	+/+	+/+	+/+	±/+	+/+	-/+	±/±
77777m	±/+	±/+	+/+	±/+	±/+	±/+	+/+	+/+	+/+
263315m	±/+	+/+	±/+	±/+	+/+	-/+	+/+	+/+	+/+
General	±/ +	±/ +	partly+/ +	±/ partly+	±/ partly+	±/ partly+	±/ partly+	±/ partly+	partly+/ partly+

The recognition systems of emotional states must be trained by speaker in order to distinguish the fury. The emotional intra-speaker states can be clearly distinguished, but we cannot specify the emotional inter-speaker states.

Comparing the prosody for happiness with that for fury, in the Romanian language, we noticed amazing similarities with the other European language. Even more remarkable, the other two languages where the same findings where reported on ambiguity between happiness and fury are of different roots than Romanian: while the Romanian is a Latin language with Slavic influence, the other two languages are German and Greek. We hypothesize that all Indo-European languages have a similar representation of emotion and the same resemblance between happiness and fury. This general conclusion on similarity of emotion representation in European languages, disregarding their particular roots, is preserved for all emotions.

5 Discussion and 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 paralinguistic features in prosody, selected for the analysis, has been motivated by the analysis of manifest, intentional emotions.

For sentences uttered with manifest emotional load 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 determine its validity 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, at least from the perspective of the voice personality of some speakers.

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 are. 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 of obtaining high confidence in the results.

Acknowledgments. Research partly performed in relation to the Romanian Academy “priority research” theme “Cognitive Systems” and to a CEEX grant.

References

1. Buluti, M., Narayanan, S.S., Syrdal, A.K. Expressive speech synthesis using a concatenative synthesizer, http://sail.usc.edu/publications/BulutNarayananSyrdal.pdf, accessed 9.may. 06
2. Douglas-Cowie, E. , Campbell, N., Cowie R. and Roach, P. Towards a new generation of databases, Speech Communication, vol. 40, p. 33-60, 2003
3. https://nats-www.informatik.uni-hamburg.de/intern/proceedings/2004/LREC/pdf/41.pdf, (accessed 9.May. 06)
4. Iriondo I. et al, Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques. http://serpens.salleurl.edu/intranet/pdf/239.pdf. (accessed 01 July. 2006)
5. Engberg I.S., Hansen A.V., Documentation of the Danish Emotional Speech Databasehttp://kom.aau.dk/~tb/speech/Emotions/des.pdf. (accessed 01.July.2006)
6. http://pascal.kgw.tu-berlin.de/emodb/ (accessed 01.July. 2006)
7. Teodorescu, H.N., A proposed theory in prosody generation and perception: the multi-dimensional contextual integration principle of prosody". In Burileanu, C. (Coordinator), Trends in Speech Technology, Romanian Academy Publishing House, Bucharest, Romania, ISBN 973-27-1178-7, p. 109-118, 2005
8. 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)
9. 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)
10. http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/index.htm
11. Teodorescu, H.N., Feraru, M., Trandabat D., Studies on the Prosody of the Romanian Language: The Emotional Prosody and the Prosody of Double-Subject Sentences. In Corneliu Burileanu and H.N. Teodorescu (Eds.), Advances in Spoken Language Technology, Editura Academiei Române, ISBN 978-973-27-1516-1, p. 171-182, 2007
12. Teodorescu, H.N., Feraru, M., Trandabat, Nonlinear assessment of the professional voice “pleasantness”, Biosignal, 28-30 June 2006, Brno, Czech Republic, p. 63-66, 2006
13. Ordelman, R., Poel, M., Heylen, D., Emotion annotation in the AMI Project. Extended Abstract - Humaine Workshop, Paris, March 2005, http://wwwhome.cs.utwente.nl/~heylen/Publicaties/emo-anno-REV.pdf. accessed 9.May. 06