Simpozionul național
“Bazele performanței și inventică”
organizat în cadrul “Zilele Academice Ieșene”
9 Septembrie 2006

Towards the Emotional Annotation of a Corpus for the Romanian Spoken Language

MONICA FERARU[1,2], DIANA TRANDABĂȚ[3]
[1] CERFS Center for Research, Technical University “Gh. Asachi”, Iasi, Romania
[2]National Institute of Inventics, Iasi, Romania
3Institute for Theoretical Informatics of the
Romanian Academy

Abstract. We present the steps of the annotation process for some phrases of the Romanian Spoken Language. The recordings have been made from a set of young subjects and the results are presented.

Keywords: corpus, annotation, emotional states, spoken language

1. Introduction

The aim of this paper is to describe the first steps of the annotation process carried out within the frame of the realization of a database of sounds for the Romanian language. The sounds database was made by the members of CERFS and Institute of Theoretical Informatics (of Romanian Academy) and the work is part of the project of Voiced Sounds of Romanian Language [1].

The term *language resources* refers to a set of speech or language data and descriptions in machine readable form, used e.g. for building, improving or evaluating natural language and speech algorithms or systems, or, as core resources for the software localization and language services industries, for language studies, electronic publishing, international transactions, subject-area specialists and end users. Examples of language resources are written and spoken corpora, computational lexicons, terminology databases, speech collection, etc.

As language resources are usually intended to capture the behavior of human productions for its automatic computer-aided reproduction, the data is processed and all information is “squeezed” and marked as a meta-level of annotations.

After a brief presentation of the archive of sounds of the Romanian language in section 2, we begin the annotation description in section 3 with the first annotation levels, the phonetic and linguistic level. Section 4 describes the emotional annotation which intends to capture parameters that distinguish between different emotional states, while the last section presents some concluding remarks and further work.

2. The archive of Romanian spoken language[1]

The Romanian Language has no web-based, freely accessible sound archive so far, as most European language do, lack felt both in research and teaching activities. Thus, Prof. H-N Teodorescu has taken the initiative to build a sound archive, a "dictionary of sounds" for the Romanian language. The web-site page has been created though the cooperation of The Institute for Computer Science of the Romanian Academy (the Group for Spoken Language Processing), the "Al. I. Cuza" University of Iaşi (Faculty of Computer Science) and The Technical University "Gh. Asachi" Iaşi (Center for Excellence in research "CERFS").

The main goals of the spoken language corpus for the Romanian language are:

1. The database will include both professional voices ("perfect" pronunciations), and non-professional voices – the "voices of the people in the street". For the moment, it concentrates on voices from the Iasi region (East Romania, region of Moldova); following that in the next phase, it will include voices from all the regions of Romania and from other regions where Romanian language is spoken (R. Moldova, Ukraine, Yugoslavia etc.).

2. Based on this corpus, a vast systematical study of the currently spoken Romanian language can be carried out. The study will include a statistical vowel triangle, statistical characteristics of the spectra, regional statistical characteristics etc.

3. The database is also aimed to serve as a basis for building concatenate voice synthesizers.

4. The database may be helpful in improving voice recognition systems based on acoustical features, thus becoming a benchmark database.

The data consists of approx. 500 recordings (in different formats: .wav 16 bits, .wav 24 bits, .ogg and .txt) of the basic Romanian sounds and of some sentences coming form 6 speakers. The basis sounds are:

- vowels, recorded in a sustained way;
- consonants, recorded according to the IPA standard in VCV format (V=vowel "a", C=consonant);
- diphthongs, triphthongs and hiatus;
- Romanian specific sounds (ce, ci, che , chi , ge, gi, ghe, ghi).

Besides isolated sounds, short sentences were also recorded. The speakers have been asked to try to simulate the following emotional states: happiness, sadness, joy, hate, optimism, pessimism, exclamation, interrogation, neutral tone, madness.

3. The linguistic annotation

The first step of the annotation of the Romanian spoken corpus is the segmentation of the voice signal at phonetic level (phoneme and syllable). Then, starting to merge syllables into words and sentences, a second level, the linguistic

[1] This section presents the archive objectives, transcribed from the paper [2].

one, is constructed. For the annotation of the corpus of the Romanian spoken language, we have used Praat, due to its efficiency and facility. For the annotation, we used the signal wave form, the pitch, and the energy curve. Figure 1 presents the annotation result for the sentence "Vine mama!" ("Mother is coming!").

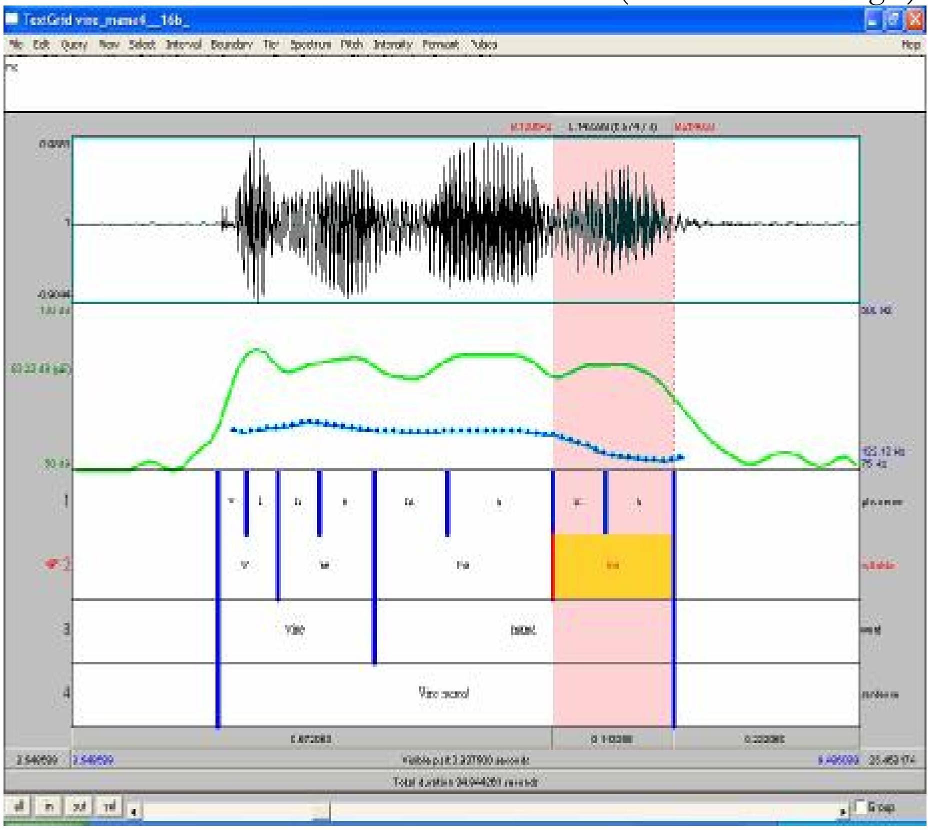
TextGrid vine_mama4__16b_
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Video
Help
1
2
3
4
Vine
Vine mama!
Group

Figura 1. *Annotation example for the sentence Vine mama! (Mother is coming! in romanian)*

It was sometimes difficult to clearly distinguish where the real boundary between phonemes is, especially at the contact between two vowels.

4. The emotional annotation

There is no general agreement on how to annotate emotional content in a natural database. To have an emotional database it is necessary to have a number of speakers which try to simulate emotions in various contexts [3]. The listeners must perceive the intended emotions afterwards and they can make the difference between the simulated and authentic vocal expressions of emotion. It is better for recordings to have professional voice for example actors, professors etc [4].

Emotions have been categorized in "basic" or "non-basic" [7]. The second class, "non-basic", are classified variously as "blends", "combinations", "mixed" or "secondary". In 1992, Johnson-Laird and Oatley said that there are five basic emotions and in 1998 they revised this number to four: happiness, anger, sadness and fear.

In our study, we have young people, students and PhD-students. The age of the speakers varied between 23 and 28. The speakers were asked to simulate the following emotions: happiness, sadness, joy, hate, optimism, pessimism, exclamation, interrogation, neutral tone, madness. The simulated effective states in a database can be considered to represent basic emotion. They have recorded three sentences ("Aseara", "Cine a facut asta?", "Vine mama"). The recordings were monophonic with 22 050Hz. We made the annotation using Praat software like in example from Figure 1. Other persons tried to identify the emotional states hearing the recordings. In the next figure we represented four emotional states and in the center we put the neutral ton.

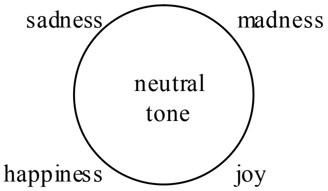

Figura 2*. A emotional states diagram*

The people which confirmed us the emotional states made the following observations:

- there is confusion between the happiness and the joy, between sadness and madness;
- there is confusion also between exclamation and happiness, between pessimism and sadness.

The next table presents, for one person who made the recordings with the emotional states, the results of confirmation to the emotional states.

Table 1. *Confirmation the emotional states for one person*

Emotional states	**Aseara (Last night)**	**Cine a facut asta (Who did this?)**	**Vine mama (Mother is coming)**
happiness	joy	interrogation	yes
sadness	madness	yes	yes
joy	happiness	happiness	yes
hate	sadness	yes	yes
optimism	yes	yes	yes
pessimism	yes	yes	yes
madness	yes	yes	yes

Table 2. *Confirmation of special situations*

Other Situations	**Aseara (Last night)**	**Cine a facut asta (Who did this?)**	**Vine mama (Mother is coming)**
exclamation	joy	joy	happiness
interrogation	yes	exclamation	yes
neutral tone	yes	yes	yes

One person identified the neutral tone with pessimistic state and other person the interrogative as optimistic state. The persons which identified the emotional states complained that it is hard to distinguish between happiness and joy, between sadness and madness and said that it is easy to make difference between hate and happiness, between exclamation and sadness.

We calculated the confusion matrix guided by professor H.N. Teodorescu and we obtained the next results of the phrases "Vine mama" (Mother is coming) and "Cine a facut asta?" (Who did this?)" for the subject 4, which are shown in the table 2 and 3.

Table 3. *The confusion matrix for the phrase "Vine mama" (Mother is coming)*

Vine mama	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	66.66	0	0	0	0	0	0	0	33.33	0
S2	0	66.66	0	0	0	0	0	0	0	33.33
S3	33.33	0	66.66	0	0	0	0	0	0	0
S4	0	0	0	100	0	0	0	0	0	0
S5	0	0	0	0	100	0	0	0	0	0
S6	0	0	0	0	0	100	0	0	0	0
S7	33.33	0	0	0	0	0	66.66	0	0	0
S8	0	0	0	0	0	0	0	100	0	0
S9	0	0	33.33	0	0	0	0	0	66.66	0
S10	0	0	0	0	0	0	0	0	0	100

Table 4. *The confusion matrix for the phrase "Cine a facut asta?" (Who did this?)*

Cine a facut asta	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
S1	0	0	0	0	0	0	0	66.66	33.33	0
S2	0	66.66	0	0	0	0	0	0	0	33.33
S3	33.33	0	33.33	0	0	0	0	33.33	0	0
S4	0	0	0	100	0	0	0	0	0	0
S5	0	0	33.33	0	66.66	0	0	0	0	0
S6	0	33.33	0	0	0	66.66	0	0	0	0
S7	0	0	33.33	0	0	0	66.66	0	0	0
S8	33.33	0	0	0	0	0	33.33	33.33	0	0
S9	0	0	0	0	0	0	0	0	100	0
S10	0	0	0	0	0	0	0	0	0	100

In some annotation, there is information from video regarding the facial expressions, movements of the head, of the shoulders, the gaze, the hands gestures etc [5].

Hartman et al. said that exist a set of six dimensions to describe expressivity like: spatial extent, temporal extent, power, fluidity, repetition and overall activity [6]. Montepare et al. analyzed in the annotation process hand position, gait, fluidity, stiffness, strength, speed, spatial expansion, and activity. Wallbott in 1998 analyzed the information from upper body, shoulders (up, backward, forward), head (downward, backward, turned sideways, bent sideways), arms, hands, movement quality (activity, spatial expansion, movement dynamics, energy, power) and symmetry.

Vincent Wan, from University of Sheffield said that the most frequently annotated behaviors are facial expressions (78.6%), gestures (11.3%) and postures (10%) and the most frequent attributes are gaze direction (26.8%), head movements (23.5%), blinking (15.8%), eyebrows movements 10%. Serenity involve no gestures whereas exaltation is accompanied by fast and energetic gestures. Anger is correlated with fast and intensive whereas irritation involves slow and low intensive gestures.

5. Conclusions and further work

In the vocal emotion literature, the focus has so far been on major language for example English, French, German and very little is known about the vocal correlates of emotion in Romanian Spoken Language.

We have presented the annotation process of a corpus for the Romanian Spoken Language, using Praat software. We have created a large corpus of phonemes, words and phrases spelled by a significantly number of subjects within various contexts (standard, emotional, pathological, etc.).

More than 30 subjects have contributed to speech recordings. The recordings have not been yet annotated and new recordings with the corresponding annotations will be added.

We received the confirmation of the emotional states from the persons which are PhD-students, from the persons who finished a faculty and from the students and in the future will make a validation team.

The database maybe a useful tool in basic research. The statistical classification of emotional speech and the estimation of the degree of emotion can offer new opportunities for the future Internet database technologies. This database can be considered a step towards computer understanding and responding to human emotions. In this way, we hope that this database will be helpful to the researchers on this domain and it'll be a basic point of departure.

Acknowledgments. The Grant CEEX - 3140P / 10.10.2005 – "The automatic system for paraclinic diagnosis in dysfunctional syndrome of stomatognat system" has partly supported of the research for this paper.

References

[1]. H.N.Teodorescu, D. Trandabat, M. Feraru, R. Ganea, A. Verbuta, M. Zbancioc, Voiced Sounds of Romanian Language Project,

http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/ro/arhiva_sunete.htm

[2]. H.N. Teodorescu, D. Tandabat, M. Feraru, M. Zbancioc, R. Luca, A corpus of the sounds in the Romanian spoken language for language-related education, Human and material resources in foreign language learning, 12-14 july 2006, Murcia, Spain

[3]. E. Douglas-Cowie, N. Campbell, R. Cowie and P. Roach, Towards a new generation of databases, Speech Communication, vol. 40, pp. 33-60, 2003

[4]. T. Seppannen, J. Toivanen and E. Vayrynen, Media team speech corpus: a first large Finnish emotion speech database

[5]. J.C. Martin, G. Caridakis, L. Devillers, K. Karpouzis, S. Abrilian, Manual annotation and automatic image processing of multimodal emotional behaviours: validating the annotation of TV interviews, Proceedings of Intelligent Virtual agents 2005

[6]. B. Hartmann, M. Mancini, C. Pelachaud, 2005, Implementing expressive gesture synthesis for embodied conversational agents, Gesture Workshop, LNAI, Springer, may 2005 [http://www.iut.univ-paris8.fr/~pelachaud/AllPapers/gw05-expre.pdf]

[7]. N. Fakotakis, Corpus design, recording and phonetic analysis of Greek emotional database