‘You are sooo cool, Valentina!’: recognizing social attitude in speech-based dialogues with an ECA

Fiorella de Rosis, Anton Batliner, Nicole Novielli, Stefan Steidl

Angaben zur Veröffentlichung / Publication details:

Rosis, Fiorella de, Anton Batliner, Nicole Novielli, and Stefan Steidl. 2007. “‘You are sooo cool, Valentina!’: recognizing social attitude in speech-based dialogues with an ECA.” In *Affective computing and intelligent interaction: Second International Conference, ACII 2007 Lisbon, Portugal, September 12-14, 2007*, edited by Ana C. R. Paiva, Rui Prada, and Rosalind W. Picard, 179–90. Berlin: Springer. https://doi.org/10.1007/978-3-540-74889-2_17.

Ana Paiva, Rui Prada, Rosalind W. Picard (Eds.): Affective Computing and Intelligent Interaction.
Proc. ACII 2007, Lisbon, Springer, Berlin-Heidelberg, pp. 179-190.

'You are sooo cool, Valentina!' Recognizing social attitude in speech-based dialogues with an ECA.

Fiorella de Rosis[1], *Anton Batliner*[2], *Nicole Novielli*[1], *Stefan Steidl*[2]

[1]Intelligent Interfaces, Department of Informatics, University of Bari
Via Orabona 4, 70126 Bari, Italy
{derosis, novielli}@di.uniba.it

[2] Universitaet Erlangen-Nuernberg
Martensstrasse 3, 91058 Erlangen - F.R. of Germany
{batliner, steidl}@informatik.uni-erlangen.de

Abstract. We propose a method to recognize the 'social attitude' of users towards an Embodied Conversational Agent from a combination of linguistic and prosodic features. After describing the method and the results of applying it to a corpus of dialogues collected with a Wizard of Oz study, we discuss the advantages and disadvantages of statistical and machine learning methods if compared with other knowledge-based methods

1 Introduction

This work is part of a research project that is aimed at adapting the behavior of an advice-giving ECA (that we named Valentina) to the 'social' attitude of its users. To make suggestions effective, knowledge of the user characteristics (preferences, values, beliefs) is needed: this knowledge may be acquired by observing the users' behavior during the dialogue to infer a dynamic, consistent model of their mind. Affect proved to be a key component of such a model (Bickmore and Cassell, 2005). Adaptation may be beneficial if the user characteristics are recognized properly but detrimental in case of misrecognition; this is especially true for affective features, for which consequences of misrecognition may be dangerous. An example:

The user: *"You are not very competent Valentina!"* (by smiling)
The ECA: *"Thanks!"* (by reciprocating smile).

Recognition of the affective state should therefore consider on the one hand the aspects that may improve interaction if properly recognized and, on the other hand, the features that available methods enable recognizing with an acceptable level of accuracy. Affective states vary in their degree of stability, ranging from long-standing features (personality traits) to more transient ones (emotions). Other states, such as *interpersonal stance*[1], are in a middle of this scale: they are initially influenced by

[1] To Scherer, *interpersonal stance* is *"characteristic of an affective style that spontaneously develops or is strategically employed in the interaction with a person or a group of persons, coloring the interpersonal exchange in this situation (e.g. being polite, distant, cold, warm, supportive, contemptuous":* http://emotion-research.net/deliverables/D3e%20final.pdf

individual features like personality, social role and relationship between the interacting people but may be changed, in valence and intensity, by episodes occurring during interaction. This general concept was named differently in recent research projects, each considering one of its aspects: *empathy* (Paiva, 2004), *engagement, involvement, sympathy* (Hoorn and Konijn, 2003, Yu et al, 2004). A popular term among e-learning researchers is *social presence* (or *co-presence*), which received several definitions, from the general one "*the extent to which the communicator is perceived as 'real'*" (Polhemus et al, 2001) to the more ECA-specific one "*the extent to which individuals treat embodied agents as if they were other real human beings*" (Blascovich, 2002). The concept of social presence refers to the nature of interaction with other people in a technologically mediated communication (Rettie, 2003). In reasoning about the social response of users to ECAs, we prefer to employ the term *social attitude.* To distinguish warm from cold social attitude, we will refer to Andersen and Guerrero's definition of *interpersonal warmth* (1998) as "*the pleasant, contented, intimate feeling that occurs during positive interactions with friends, family, colleagues and romantic partners...[and]... can be conceptualized as... a type of relational experience, and a dimension that underlines many positive experiences.*"

Researchers proposed a large variety of markers of social presence related to nonverbal behavior, such as body distance, memory, likability, physiological data, task performance and self-report (Bailenson et al, 2005). We studied this attitude and the factors affecting it by observing the verbal behavior of subjects conversating with an ECA in a Wizard of Oz (WoZ) simulation study. In a previous work (de Rosis et al, 2006), we described how social attitude can be recognized from language and how its evolution during the dialogue can be modeled with dynamic oriented graphs. In that context, we described the 'signs' through which, according to psycholinguistic theories, social attitude may be displayed, and discussed the difficulty of recognizing them by means of simple keyword analysis. The corpus of dialogues on which the methods were developed and tested were collected with studies in which users interacted with the ECA with a graphical interface, by means of keyboard and mouse. Subsequently, we decided to study whether and how changing the interaction mode to speech and touch-screen influenced the user attitude towards the ECA. We collected a new corpus of dialogues with a new set of WoZ studies and extended our method of social attitude recognition in two directions:

- by refining our method of language analysis with a bayesian classifier rather than a keyword analyzer,
- by incorporating acoustic analysis.

In this paper, we will describe the results of this research. In addition to proposing a method for recognizing social attitude from speech and language, we will discuss the advantages and disadvantages of statistical and machine learning methods (now prevailing in affect recognition) if compared with other knowledge-based methods.

2. Corpus Description

We collected, with a WoZ study, thirty speech-based dialogues (with 907 moves overall) from subjects between 21 and 28 years of age, equidistributed by gender and background (in computer science or in humanities). After a first analysis of the data, we noticed that different signs of social attitude could be observed by looking at their prosodic or their linguistic characteristics. Especially when the moves were long, they could be differentiated into several parts (segments), each showing different combinations of acoustic and linguistic signs, cf. the discussion on adequate units of analysis in (Batliner et al, 2003). Some examples:

> *“Vabbé (‘Come on’*, with a neutral intonation), *meglio così insomma’* (*‘So much the better, all in all!’*, with a light laughter)
>
> *“Mmm* (with an intonation of ‘I’m thinking’) *caffè d’orzo, biscotti e cornetto vuoto* (*‘barley coffee, biscuits and an empty croissant’*, with a neutral intonation)*”*

Table 1: Our markup language for signs of social attitude

Sign with definition
Linguistic signs
Friendly self-introduction The subjects introduce themselves with a friendly attitude (e.g. by giving their name or by explaining the reasons why they are participating in the dialogue)
Colloquial style The subject employs a current language, dialectal forms, proverbs etc
Talks about self The subjects provide more personal information about themselves than requested by the agent
Personal questions to the agent. The subject tries to know something about the agent's preferences, lifestyle etc, or to give it suggestions in the domain.
Humor and irony The subjects make some kind of verbal joke in their move
Positive or negative comments The subjects comment the agent's behavior in the dialogue: its experience, its domain knowledge, etc.
Friendly farewell This may consist in using a friendly farewell form or in asking to carry-on the dialogue.
Acoustic signs
Agreement The dialogue segment displays an intonation of agreement with the system
Friendly intonation The dialogue segment displays a friendly intonation
Laughter The dialogue segment displays a smile or laughter
Neutral The dialogue segment does not display any affective intonation
Negative intonation The dialogue segment displays a negative intonation
I’m thinking The dialogue segment displays, in its intonation, a reflection attitude

We therefore decided to preprocess the corpus as follows:

a. we isolated the subset of moves which appeared to be clearly 'neutral' from both the linguistic and the acoustic viewpoint (e.g., a *'yes'* or *'no'* with no particular prosody);

b. we extracted a sample of moves from this subset, to include it in the analysis as representative of 'neutral' cases;

c. we integrated this sample of neutral moves with all 'potentially non neutral' moves in the dialogues; this combination of the two sets was the corpus we used in our analysis;

d. we segmented manually the moves in the corpus into 'acoustically homogeneous' segments;

e. three PhD students labelled independently the corpus with an acoustic-oriented and a linguistic-oriented markup language. While acoustic labelling was done at the segment-level, linguistic labelling was done at the move-level.

Linguistic and acoustic labels are listed in Table 1, with their meaning. The following are the reasons why we decided to label our corpus according to linguistic and acoustic criteria at different grain size:

a. Linguistic analysis is aimed at recognizing in a user move the signs that may be employed to adapt the next system move.

Example 1a:

S: *Hi' my name is Valentina…*
U: *What should I do?*(neutral)
S: *To start, I'm going to ask you some information about your eating habits. What did you eat at dinner, yesterday?*

Example 1b:

S: *Hi' my name is Valentina.*
U: *Hi Valentina, my name is Carlo and I'm happy to interact with you* (friendly self-introduction)
S: *Hi Carlo, good to hear your interest! I hope you'll find our dialogue useful. I'll, first of all, ask you some information about your eating habits. OK?*

Example 2a:

S: *Do you believe your weight is correct?*
U: *No* (neutral)
S: *Will you tell me something more? Which is your weight and height?*

Example 2b:

S: *Do you believe your weight is correct?*
U: *No, I am a bit overweight, I would like to loose some kilos* (talks about self)
S: *Good that you care for your fitness!* …

At the same time, as far as the dialogue goes on, linguistic signs discovered in the dialogue history contribute to build an overall, dynamic image of the social attitude of the user towards the advice-giving ECA (de Rosis et al, 2006).

b. Acoustic analysis is aimed at enriching the linguistic connotation of moves with information about their intonation. When the segment corresponds to an entire move, acoustic parameters just refine the linguistic description. When several acoustically different segments are isolated in a single move, the variation of intonation within a move may help in interpreting its meaning and reducing the risk of errors. In the next Section, we will see some examples of this kind of recognition. Our corpus includes 1020 segments overall, with the frequency of labels (majority agreement among raters) that is shown in the second column of Table 2. We will illustrate columns 3 and 4 of this table in the next Section.

Table 2: prevalence of linguistic and acoustic signs of social attitude in our corpus

Linguistic labels	Frequency	Recall	Precision
Friendly self-introduction	2%	99.5	37.5
Friendly farewell	3%	99.5	38.9
Colloquial style	3%	75.9	11.7
Question about the agent	6%	85.2	30.9
Talks about self	16%	78.5	48.9
Positive comment	5%	4.3	66.7
Neutral	56%	48.4	94.9
Negative comment	3%	24.0	60.0
Acoustic labels			
Agreement	5 %	47.1	21.4
Friendly intonation	14 %	24.5	20.9
Laughter	9 %	44.7	23.8
I'm thinking	21 %	57.5	62.4
neutral	43 %	32.6	58.8
Negative comment	9 %	19.6	12.4

3. Sign Recognition Method

3.1. Acoustic Analysis of Segments

For each segment, we first computed a voiced-unvoiced decision. For each voiced sub-segment, a prosodic feature vector consisting of 73 features (69 for duration, energy, and pitch, and 4 for jitter/shimmer) was computed; subsequently, minimum, maximum, and mean values were calculated for each segment, resulting in a total of 219 acoustic features. This approach is fully independent from linguistic (word) information: we do not need any word segmentation, and we do not use acoustic features such a Mel Frequency Cepstral Coefficients (MFCCs) which on the one hand have proved to be competitive for classifying affective speech; on the other hand, as they implicitly contain word information, a strict separation of linguistic and acoustic modelling would no longer have been possible.

As classifier, we used Linear Discriminant Analysis; with Principal Component Analysis, the 219 features were reduced to 50 features. As we are faced with a strong sparse data problem - very few speakers, and some of the classes could be observed only for some of the speakers - we decided in favour of leave-one-case-out; our classification is thus not speaker-independent.

Results of this analysis are described, in terms of recall and precision, in the last two columns of Table 2, lower part. `I'm thinking` seems to be the best sign to recognize; `Negative comment`, `Agreement` and `Friendly intonation` the most difficult ones. However, I'm thinking is not a specific sign of social attitude: it is rather a sign of 'doubt' or of a reflexive personality trait. We thought how to possibly compact the six signs, to increase the recall rate. A plausible combination might assemble all 'positive' signs (Agreement, Friendly intonation), the 'non positive' ones (Neutral-i and Negative comments) and leave separate the sign of doubt (I'm thinking): this would produce a 42% recall for the'positive' signs and a 62% for the'non positive' ones. This idea was confirmed by a careful analysis of results of acoustic analysis of individual moves, in which we could notice that the distinction between Agreement and Friendly intonation was quite fuzzy.

3.2. Linguistic Analysis of Moves

As we anticipated in the Introduction, we improved our original keyword-based recognition method by applying a bayesian classifier in which an input text is categorized as 'showing a particular sign of social attitude' if it includes some word sequences belonging to *semantic categories* which are defined as 'salient' for the considered sign. More in detail: bayesian classification enables associating with every string (segment or full move) a value of a-posteriori probability for every sign of social attitude. Given:

- a set S of signs of social attitude that may be displayed in the language, with $S = \{s_1,\ldots, s_j,\ldots,s_n\}$;
- a set C of semantic categories of word sequences in the language, with $C = \{c_1,\ldots, c_h,\ldots,c_m\}$;
- a mapping between signs and categories, according to which the categories $c_h, c_k,\ldots,c_z$ are considered 'salient' for the sign s_j. E.g., the categories 'Greetings', 'Self-introduction', and 'Ciao' are defined as salient for the Friendly self-introduction sign;
- a combination $V(c_h, c_k,\ldots,c_z)$ of truth values for the categories $c_h, c_k,\ldots,c_z$, denoting their presence in a given sentence. E.g., the combination (0,1,1) for the set $\{c_1, c_2,\ldots,c_3\}$ denotes that 'Greetings' is absent while 'Self-introduction' and 'Ciao' are present in a sentence, like in *"Hi, my name is Carlo";*
- a prior probability $P(s_j)$ of the sign s_j in the sentences of the language;
- a prior probability $P(V(c_h, c_k,\ldots,c_z))$ for the combination of truth values $V(c_h, c_k,\ldots,c_z)$ in the language. E.g. 4 % of sentences in the language include a 'Self-introduction' and a 'Ciao' and no 'Greetings';
- a conditional probability $P(V(c_h, c_k,\ldots,c_z)| s_j)$ for the combination $V(c_h, c_k,\ldots,c_z)$ in the sentences displaying the sign s_j. E.g., 85 % of the sentences showing a sign of Friendly self-introduction include a 'Self-introduction' and a 'Ciao' and no 'Greetings';

and given:

- a result of the lexical analysis of the string m_h, as a combination of truth values for all the elements in $(c_1,\ldots,c_h,\ldots,c_m)$;

the probability that the string m_h displays the sign s_j may be computed as follows:

$$P(s_j|V(c_h, c_k,\ldots,c_z)) = P(V(c_h, c_k,\ldots,c_z)| s_j) * P(s_j) / P(V(c_h, c_k,\ldots,c_z))$$

Notice that this formula does not assume the conditional independence of semantic categories given a sign.

The recognition performance of the various signs in our corpus are shown, again in terms of recall and precision, in the last two columns of Table 2, upper part. This table clearly shows that Positive and Negative comments are the most difficult signs to recognize, while the recall for the other signs is quite good: we will come back to this problem in the next session and will describe, in the Discussion, how we plan to improve recognition of these features in the longer term.

4. Integration Of Acoustic and Linguistic Features

We did two types of integration: a) combination of both features at the segment level, and b) linguistic analysis at the move level, integrated with acoustic features at the segment level. Let us describe the two methods in more detail.

4.1. Linguistic and Acoustic Analysis at the Segment Level.

Prior to describing how we combined the two sets of features we show, in Table 3, the confusion matrix for acoustic analysis. This table shows that confounding with Neutral-i is the main source of reduction of recall for all signs; negative intonation (Negative-i) is often confounded also with Friendly intonation and Laughter.

Table 3: confusion matrix for acoustic signs

	Agr	FrInt	Laughter	I'mThinking	Neutral-i	Negative-I
Agreement	47.1	7.8	11.8	3.9	17.6	11.8
Friendly intonation	12.9	24.5	12.2	7.2	26.6	16.5
Laughter	10.6	9.4	44.7	7.1	17.6	10.6
I'm thinking	5.1	5.6	11.2	57.5	9.8	10.7
Neutral-i	9.1	19.4	13.7	10.3	32.6	15.0
Negative-i	10.9	21.7	16.3	12.0	19.6	19.6

To integrate acoustic with linguistic features, we assigned to the segments the same linguistic labels that were assigned by raters to the whole move. An example: *"No! / La frutta... qualche frutta / Ma non tutte."("No! / Fruits... some fruits / But not all fruits).* This is a move, divided into three segments, all labelled as Negative comment and Familiar style as the whole move was labelled.

Differently from acoustic analysis, our bayesian classifier does not force us to select only one of them, but enables us to consider this as an index of the presence of multiple signs; as a matter of fact, some segments displayed several linguistic signs of social attitude at the same time: see the previous example, but also the following one:

> *"Vabbé, ma non mangio cose fritte ogni giorno!" ("OK, but I don't eat fried food every day!"):* a Talk about self and a Negative comment, with Familiar style.

However, to produce a confusion matrix for lingustic analysis (in Table 4) that might be compared with the matrix for the acoustic one we selected, for every segment, only the sign with maximum probability value. As a consequence, if data in the diagonal of this table are compared with the recall data in Table 2, one may notice a reduction of recall for all signs. Notice that, for several reasons, the comparison between acoustic and linguistic analysis is not conceived so as to be in favour of the second one: as we said in Section 2, both kinds of analysis were performed on the segments that were subjectively considered as ‘acoustically significant’. This way, we left out of the analysis some potentially (from the language viewpoint) significant segments. In addition, the examined segments, while being homogeneous from the acoustic viewpoint, were often not homogeneous from the linguistic one. Finally, as we said, we excluded the case of multiple signs for comparability reasons.

Table 4: confusion matrix for linguistic signs

	Fsi	Ffwell	Collst	Qagt	Talks	PosC	Neut-l	NegC
Friendly self-introduction	0.64	0.29	0.00	0.00	0.07	0.00	0.00	0.00
Friendly farewell	0.00	0.71	0.13	0.00	0.04	0.00	0.13	0.00
Colloquial style	0.00	0.00	0.57	0.14	0.14	0.00	0.14	0.00
Question about the agent	0.00	0.00	0.00	0.72	0.22	0.00	0.06	0.00
Talks about self	0.00	0.00	0.05	0.01	0.75	0.01	0.13	0.05
Positive comment	0.00	0.07	0.25	0.14	0.11	0.25	0.18	0.00
Neutral-language	0.00	0.01	0.10	0.07	0.24	0.03	0.50	0.05
Negative comment	0.00	0.00	0.20	0.17	0.24	0.02	0.20	0.17

We analysed, in particular, the segments belonging to the most problematic category: negative intonation. An accurate analysis of these segments enabled us to understand the nature of this data. As displayed in Table 3, the recognition rate of these segments was quite low (less that 20%). If the result of linguistic analysis was added to the acoustic one, the recognition rate of ‘acoustically and linguistically negative’ cases increased to 31 %: a sligth increase, then. But, by looking deeper into the segments, we found that cases in which the subjects expressed their negative attitude both linguistically and acoustically were really ‘extreme’ cases. An example:

“Madò, ma ci metti di tempo a rispondere!” (My god, it takes you a lot to answer!): acoustically and linguistically negative.
Comment: the subject seems to be really bored by the ECA’s behavior.

In the majority of cases, on the contrary, the segments that were annotated as ‘showing acoustic signs of negative attitude’ displayed multiple (and apparently inconsistent) results of acoustic and linguistic analysis. This was not an inconsistency though, but rather a realistic description of the subjects’ behavior when reacting negatively to an ECA’s move. Some examples:

“Cioè, ma non c’entra con quello che ti ho detto!” (But this has’n got anything to do with what I said!): acoustically: a `Laughter`; linguistically: a `Negative comment` and a `Talk about self`.
Comment: the subject expresses his negative evaluation of the ECA’s behavior with a bit of irony and politeness.

“Eh però, quando tu parli di frutta secca non mi parli di dosi!” (Hey, but when you talk about dried fruits, you don’t say anything about doses!); acoustically: a

Friendly intonation; linguistically: a Negative comment and a Question about the agent.

Comment: again, the subject expresses friendly his negative evaluation of the ECA's behavior.

"No, mi auguro di no!" (No, I hope no!); acoustically: neutral intonation; linguistically: a Negative comment and a Colloquial style.

Comment: in this case, the subject expresses his negative evaluation of the ECA's behavior linguistically, but with a neutral intonation and by smoothing it with a colloquial style.

To summarise: apparently, our subjects tended to express their negative attitude towards an ECA's move by avoiding to be rude: they smoothed their negative comments by introducing some bit of politeness in the intonation (in the form of laughter or smiling), or in the language (in the form of of colloquial style or other).

To integrate acoustic with linguistic signs, we then decided to compact the 8x6 combinations of labels into a lower number of categories, defined according to adaptation purposes. The first need of adaptation is to distinguish, as accurately as possible, between a 'negative', 'neutral' or 'warm' attitude of the user. We labelled the corpus of segments with an automatic rule-based annotation which combined the raters' acoustic and linguistic labelling into four-categories, according to the following rules:

IF (Neutral-i or I'mThinking) and Neutral-l THEN NEUTRAL

A segment is labelled as Neutral *if it was acoustically labelled as* Neutral*–l or*I'm thinking*, and linguistically as* Neutral-l*;*

IF (Negative-i or NegativeComment) THEN NEGATIVE

A segment is labelled as Negative *if it was labelled as such either acoustically or linguistically;*

IF (¬Neutral-i ∧ ¬Thinking ∧¬Negative-i) ∨ (¬Neutral-l ∧ ¬Negative-l) THEN LIGHT-WARM

A segment is labelled as Light-war*m if it was annotated either acoustically or linguistically as displaying some positive sign*

IF (¬Neutral-i ∧ ¬Thinking ∧ ¬Negative-i ∧ ¬Neutral-l ∧ ¬Negative-l) THEN WARM

A segment is labelled as 'warm' if it was annotated both acoustically and linguistically as displaying some positive sign (not neutral, not I'm thinking and not negative)

For every segment, we had a 'probability value' for each of the 8+6 signs. We processed this dataset with K2 learning algorithm (k-fold cross validation, with k=number of segments with WEKA) and got a 90.05 % of recall; results of this analysis are displayed in Table 5.

Table 5: confusion matrix for the combination of acoustic and linguistic features

	Negative	Neutral	Light-warm	Warm	Recall	Precision
Negative	232 (94 %)	11 (4 %)	1 (.5 %)	4 (1.5 %)	.94	.94
Neutral	2 (1 %)	174 (95 %)	8 (4 %)	0	.95	.84
Light-warm	10 (3 %)	23 (6 %)	317 (85 %)	*21 (6 %)*	.85	.92
Warm	3 (1 %)	0	*19 (9 %)*	201 (90 %)	.90	.89

A positive aspect of this recognition method is that the only non negligible confusion is between `light-warm` and `warm` attitude: a kind of confounding that is not very dangerous for adaptation. Note that again, due to sparse data, this cross-validation was not done speaker-independently.

4.2. Acoustic Analysis as Complementary to the Linguistic One.

This is an ongoing work that we performed, so far, on a subset of the moves. Every move was first analysed to recognize linguistic signs of social attitude; this information was then integrated with the recognized prosodic signs in every 'acoustically significant' segment of the move. This analysis, together with possible information about the context in which the move was uttered by the subject (previous ECA's move) enabled us to have a deeper insight into the subject's attitude towards the ECA and its suggestions. Some examples:

"E i dolci? Fanno proprio male i dolci?" ("How about sweets? Do sweets harm?"). This is a linguistically neutral move which, in its first segment, does not show any particular affective intonation. In the second one, however, some light laughter is shown. This variation of intonation seems to display a little embarrassment of the subject in admitting her preferences.

"No, finora non ho avuto questi problemi; il fegato funziona, e i reni pure". (*"No, so far I had no problem; my liver works, my kidneys too.").* This move comes after a system's information of the possible negative consequences of the kind of dietary habits declared by the subject. In the move, the subject talks about self, initially with a negative intonation, then with a neutral one, and finally with a friendly intonation. Overall, this change of intonation during the move seems to display the subject's intention to smooth her objection to the system's remark.

"Vabbé, ma non mangio cose fritte ogni giorno: ogni tanto, una volta a settimana!" (*"OK, but I don't eat fried food every day: from time to time, once a week!"*). The context of this move is similar to that of the previous example: information about negative effects of fried food. The subject replies by describing his eating habits with a colloquial style but by introducing, at the same time, a negative intonation in the beginning of the move, probably to show his disagreement with the ECA's evaluation.

These examples demonstrate that analysis at the move level which integrates linguistic interpretation of the utterance with recognition of the variation of intonation during the utterance itself might provide more information than a simple integration of the two kinds of features at the segment level. Rather than machine learning methods, rule-based recognition criteria (including consideration of the context) seem to be more appropriate to this task.

5. Discussion

As we said in the Introduction, recognition of the affective state should consider on one hand the aspects that may improve interaction and, on the other hand, those that available methods enable recognizing with a reasonable level of accuracy. In this paper, we proposed two methods for recognizing social attitude of users in speech-based human-ECA dialogues; in the first one, we showed how integrating linguistic

Ana Paiva, Rui Prada, Rosalind W. Picard (Eds.): Affective Computing and Intelligent Interaction.
Proc. ACII 2007, Lisbon, Springer, Berlin-Heidelberg, pp. 179-190.

and acoustic features at the segment level enables distinguishing between ‘levels of social attitude’ (negative, neutral, light or strong warm) with a good level of accuracy (90 %). In the second one we proposed, with some examples, how combining language analysis at the move level with acoustic analysis at the segment level might enable deeper and more refined understanding of the user attitude towards the ECA. Research about this second method is still ongoing, and we plan to extend it in the near future.

Our research builds upon a consolidated experience in the domain. Several studies investigated how to assess affective situations from spoken language, by combining prosodic information with language features: in all these studies, language features had a supporting role to prosodic ones, which were the main recognition factors. Lee et al (2002) found that, by adding language features to acoustic information, the recognition rate of ‘negative’ and ‘non negative’ emotions increased considerably. Ang et al (2002) integrated prosodic features with a trigram model to discriminate ‘neutral’ from ‘annoyed and frustrated’ conditions in call center dialogues. Litman and Forbes-Riley (2003) combined prosodic features with lexical items to recognize the valence of emotions in spoken tutoring dialogues, by finding that the combined feature set outperformed the speech-only set. In attempting to recognize fear, anger, relief and sadness in human-human medical dialogues, Devillers and Vidrascu (2006) separated linguistic analysis from paralinguistic one, by obtaining a better performance with lexical cues than with acoustic features. In working with WoZ data, Batliner et al (2003) demonstrated that the combination of prosodic with linguistic and conversational data yielded better results than the use of prosody only, for recognizing ‘troubles in communication’, that is the beginning of emotionally critical phases in a dialogue.

Language analysis methods that may be applied in the recognition of affective features range from simple keyword recognition to more sophisticated approaches. Statistical machine learning methods are now becoming a very popular approach in this domain, after the initial rule-based methods that were applied, e.g., to recognize doubt (Carberry et al, 2002). We claim that, although statistical methods have their advantages in enabling a quick analysis of the data distributions, in building criteria that may be applied to adapt conversational systems to the user attitude a deeper inspection of the corpus, with some reasoning on the patterns they display (and a possible formalization of these patterns into decision rules) may insure more careful adaptation. In the near future, we plan to continue the work by collecting more dialogues, to overcome the sparse data problem. In addition, we will focus our study on the recognition of positive and negative comments with sentiment analysis methods. This is not a easy task, considering the kind of complex sentences our recognition methods have to deal with. For instance, “*You are not very competent Valentina!*” would not be recognized as a negative comment by just looking at the polarity of the adjective ‘competent’.

Acknowledgements

This work was financed, in part, by HUMAINE, the European Human-Machine Interaction Network on Emotion (EC Contract 507422). We sincerely acknowledge

Irene Mazzotta for cooperating to the Wizard of Oz studies in which the corpus of dialogues analysed in this paper was collected.

Main References

Andersen, P.A. and Guerrero, L.K., 1998. Handbook of Communication and Emotions. Research, theory, applications and contexts, Academic Press, New York.

Ang, J., Dhillon, R., Krupsky, A., Shriberg, E. and Stolcke, A., 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog, in: Proceedings of ICSLP, pp. 2037.

Bailenson, J.N., Swinth, K.R., Hoyt, C.L., Persky, S., Dimov, A.and Blascovich, J, 2005. The independent and interactive effects of embodied agents appearance and behavior on self-report, cognitive and behavioral markers of copresence in Immersive Virtual Environments. PRESENCE. 14, 4, 379-393.

Batliner, A., K. Fischer, R. Huber, J. Spilker, and E. Nöth: 2003. How to Find Trouble in Communication. Speech Communication, 40, 117-143.

Bickmore, T., Cassell, J., 2005. Social Dialogue with Embodied Conversational Agents, in: J. van Kuppevelt, L. Dybkjaer, & N. Bernsen (Eds.), Advances in Natural, Multimodal Dialogue Systems. New York: Kluwer Academic.

Blascovich, J., 2002. Social influences within immersive virtual environments, in: R. Schroeder (Eds.), The social life of avatars. Springer-Verlag, London, 127-145.

Carberry, S. Lambert, L. And Schroeder, L. Towards recognizing and conveying an attitude of doubt via natural language. Applied Artificial Intelligence, 2002.

de Rosis, F., Novielli, N., Carofiglio, V., Cavalluzzi, A. and De Carolis, B., 2006. User modeling and adaptation in health promotion dialogs with an animated character. Journal of Biomedical Informatics, Special Issue on ‘Dialog systems for health communications’. 39(5), 514-531.

Devillers, L. and Vidrascu, L., 2006. Real-life emotion detection with lexical and paralinguistic cues on human-human call center dialogues, in: Proceedings of INTERSPEECH.

Hoorn, J.F. and Konijn, E.A., 2003. Perceiving and Experiencing Fictional Characters: An integrative account. Japanese Psychological Research, 45 (4), 250-268.

Lee, C.M., Narayanan, S. S., and Pieraccini, R., 2002. Combining acoustic and language information for emotion recognition, in: Proceedings of ICSPL , 873-876.

Litman, D., Forbes-Riley, K., Silliman, S., 2003. Towards emotion prediction in spoken tutoring dialogues, in: Proceedings of HLT/NAACL, 52-54.

Paiva, A. (Ed), 2004. Empathic Agents. Workshop in conjunction with AAMAS.

Polhemus, L., Shih, L-F and Swan, K., 2001. Virtual interactivity: the representation of social presence in an on line discussion. Annual Meeting of the American Educational Research Association.

Rettie, R., 2003. Connectedness, awareness and social presence, in: Proceedings of PRESENCE, online proceedings.

Yu, C., Aoki, P.M. and Woodruff, A., 2004. Detecting user engagement in everyday conversations, in: Proceedings of ICSLP, 1329-1332.