'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] Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg
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 (ECA) 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 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 (and misreactions) 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 individual features like personality, social role and relationship between the

---

[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

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*, 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 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, likeability, physiological data, task performance and self-report (Bailenson et al., 2005). Polhemus et al. (2001) suggested various text-based indicators: *personal address and acknowledgement* (using the name of the persons to which one is responding, restating their name etc.), *feeling* (using descriptive words about how one feels), *paralanguage* (features of language which are used outside of formal grammar and syntax), *humor*, *social sharing* (of information non related to the discussion), *social motivators* (offering praise, reinforcement and encouragement), *negative responses* (disagreement with the other's comment), *self-disclosure* (sharing personal information). Additional dimensions have been suggested by Andersen e Guerrero (1998): *sense of intimacy* (use of a common jargon), attempt to establish a *common ground, benevolent or polemic attitude* towards the system failure, *interest* to protract or close interaction.

We studied social attitude and the factors affecting it by observing the verbal behavior of subjects interacting 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 the psycholinguistic theories proposed above, 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 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 (de Rosis et al., submitted). 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 language analysis with a bayesian classifier rather than a keyword analyzer and by incorporating acoustic analysis.

In this paper, we will describe the results of this research. After proposing a method to recognize 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; Nicholas et al., 2006, Liscombe et al., 2005). Some examples:

"Vabbé (Come on, with a neutral prosody), meglio così insomma" (So much the better, all in all!, with a light laughter)

"Mmm (with a prosody of 'I'm thinking') caffè d'orzo, biscotti e cornetto vuoto" (barley coffee, biscuits and an empty croissant, with a neutral prosody)

Table 1: Our markup language for signs of social attitude

Signs 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, experience, 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 a prosody of agreement with the system.
Friendly prosody: The dialogue segment displays a friendly prosody.
Laughter: The dialogue segment displays a smile or laughter.
Neutral: The dialogue segment does not display any affective prosody.
Negative prosody: The dialogue segment displays a negative prosody.
I'm thinking: The dialogue segment displays, in its prosody, 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. as 'neutral' moves were prevailing in the corpus, we extracted a sample set, 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, by adopting an 'optimal chunking' method: a segment was isolated from the move when a clear switch in prosody was detected. Very long segments resulting from this preliminary chunking were then segmented again according to pauses which were identified, in the transcript, with full points, semicolons and dots;

e. three PhD students labelled independently the corpus with an acoustic-oriented and a linguistic-oriented markup language (Table 1). Acoustic labelling was done at the segment-level, linguistic labelling at the move-level; overlapping between linguistic signs was allowed.

These 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 information to employ in selecting 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 a dynamic image of the social attitude of the user towards the ECA that is used, as well, in adapting the dialogue (de Rosis et al., 2006).

b. Acoustic analysis is aimed at enriching the linguistic connotation of moves with information about their prosody (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 prosody 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 omit from this table the linguistic sign of 'humour and irony' (with a low frequency) and will illustrate columns 3 and 4 of this table in the next Section. In last coloumn we provide Cohen's Kappa statistics as an estimates of interrater reliability. Overall, our markup language proved to be quite reliable (Di Eugenio, 2000), although some of the signs we wanted to label were rather fuzzy from the conceptual viewpoint.

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

Linguistic labels	Frequency	Recall	Precision	Kappa
Friendly self-introduction	2%	99.5	37.5	0.87
Friendly farewell	3%	99.5	38.9	0.65
Colloquial style	3%	75.9	11.7	0.70
Question about the agent	6%	85.2	30.9	0.56
Talks about self	16%	78.5	48.9	0.64
Positive comment	5%	4.3	66.7	0.42
Neutral-l	56%	48.4	94.9	0.53
Negative comment-l	3%	24.0	60.0	0.42
Acoustic labels				
Agreement	5 %	47.1	21.4	0.96
Friendly prosody	14 %	24.5	20.9	0.83
Laughter	9 %	44.7	23.8	0.98
I'm thinking	21 %	57.5	62.4	0.91
Neutral-p	43 %	32.6	58.8	0.76
Negative comment-p	9 %	19.6	12.4	0.94

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 of linguistic (word) information: we do not need any word segmentation, and we do not use acoustic features such as Mel Frequency Cepstral Coefficients (MFCCs). These features on the one hand have proved to be competitive for classifying affective speech, on the other hand implicitly contain word information, so that 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 third and fourth column of Table 2, lower part. I'm thinking seems to be the best sign to

recognize; Negative comment-p, Agreement and Friendly prosody 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 prosody), the 'non positive' ones (Neutral-p and Negative comment-p) 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 prosody 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. 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: the set S of signs of social attitude that may be displayed in the language, with $S = \{s_1,..., s_j,...,s_n\}$; a set C of semantic categories of word sequences in the language, with $C = \{c_1,..., c_h,...,c_m\}$; a mapping between signs and categories, according to which the categories $c_h, c_k,...,c_z$ 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 $V(c_h, c_k,...,c_z)$ of truth values for the categories $c_h, c_k,...,c_z$, denoting their presence in a given sentence. (E.g., the combination (0,1,1) for the set $\{c_1, c_2,...,c_3\}$ denotes that 'Greetings' is absent while 'Self-introduction' and 'Ciao' are present in a sentence, like in "Hi, my name is Carlo"); prior probabilities $P(s_j)$ of the sign s_j in the sentences of the language and $P(V(c_h, c_k,...,c_z))$ for the combination of truth values $V(c_h, c_k,...,c_z)$ in the language (E.g. 4 % of sentences in the language include a 'Self-introduction' and a 'Ciao' and no 'Greetings'); a conditional probability $P(V(c_h, c_k,...,c_z) | s_j)$ for the combination $V(c_h, c_k,...,c_z)$ 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: result of the lexical analysis of the string m_h, as a combination of truth values for all the elements in $(c_1,...,c_h,...,c_m)$, the probability that the string m_h displays the sign s_j may be computed as $P(s_j|V(c_h, c_k,...,c_z)) = P(V(c_h, c_k,...,c_z) | s_j) * P(s_j) / P(V(c_h, c_k,...,c_z))$.

Notice that this formula does not assume the conditional independence of semantic categories given a sign. All parameters (prior and conditional probabilities) are extimated as observed frequencies in the annotated corpus.

The recognition performance of the various signs in our corpus are shown, again in terms of recall and precision, in the third and fourth column 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 are working at improving the recognition of these features.

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-p is the main source of reduction of recall for all signs; negative prosody (NegativeComment-p) is often confounded also with Friendly prosody and Laughter.

Table 3: confusion matrix for acoustic signs

	Agr	FrInt	Laughter	I'mThinking	Neutral-p	NegativeComment-p
Agreement	**47.1**	7.8	11.8	3.9	17.6	11.8
Friendly prosody	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-p	9.1	19.4	13.7	10.3	**32.6**	15.0
Negativecomment-p	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: the following move: "No! / La frutta... qualche frutta / Ma non tutte."(No! / Fruits... some fruits / But not all fruits) was divided into three segments, all labelled as Negative comment-l and Familiar style, as the whole move was labelled.

Differently from acoustic analysis, our bayesian classifier does not force us to select only one sign, but enables us to consider cases of 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-l, with a Familiar style.

However, to produce a confusion matrix for linguistic analysis (in Table 4) to compare 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.

We analysed, in particular, the segments belonging to the most problematic category: negative prosody. An accurate analysis of these segments enabled us to understand the nature of this data.

Table 4: confusion matrix for linguistic signs

	Fsi	Ffwell	Collst	Qagt	Talks	PosC	Neut-I	NegC-I
Friendly self-introduction	**64.3**	28.6	0.0	0.0	7.1	0.0	0.0	0.0
Friendly farewell	0.0	**70.8**	12.5	0.0	4.2	0.0	12.5	0.0
Colloquial style	0.0	0.0	**57.1**	14.3	14.3	0.0	14.3	0.0
Question about the agent	0.0	0.0	0.0	**72.2**	22.2	0.0	5.6	0.0
Talks about self	0.0	0.0	5.4	1.2	**74.7**	1.2	12.7	4.8
Positive comment	0.0	7.1	25.0	14.3	10.7	**25.0**	17.9	0.0
Neutral-I	0.0	1.1	9.8	6.6	24.4	3.4	**49.5**	5.1
Negative comment-I	0.0	0.0	20.4	16.7	24.1	1.9	20.4	**16.7**

As displayed in Table 3, the recognition rate of these segments was quite low (less than 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 slight 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 prosody; 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 prosody; linguistically: a Negative comment-I and a Colloquial style.
Comment: in this case, the subject expresses his negative evaluation of the ECA's behavior linguistically, but with a neutral prosody 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 prosody (in the form of laughter or smiling), or in the language (in the form 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, suited 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 compacted the raters' acoustic and linguistic labelling into four-categories, according to the following rules (rules are applied subsequently until one of them is satisfied):

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

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

IF (NegativeComment-p or NegativeComment-l) THEN NEGATIVE

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

IF ((¬Neutral-p ∧ ¬I'mThinking) xor (¬Neutral-l)) ∧¬NegativeComment-p ∧ ¬NegativeComment-l THEN LIGHT-WARM

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

IF (¬Neutral-p ∧ ¬I'mThinking ∧ ¬NegativeComment-p ∧ ¬Neutral-l ∧ ¬NegativeComment-l) THEN WARM

A segment is labelled as Warm *if it was annotated both acoustically and linguistically as displaying some positive sign.*

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 recall of 90.05%; results of this analysis are displayed in Table 5. The higher level of accuracy in the recognition of the four categories (if compared with tables 3 and 4) is due, on one hand, to reduction of the number of features from 14 to 4 and, on the other hand, to integration of linguistic and acoustic analysis. Due to space issues, we only provide results for the combination of the two categories of signs, while we omit the separate confusion matrices. 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. Notice that again, due to sparse data, this cross-validation was not performed speaker-independently.

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

4.2 Acoustic Analysis as Complementary to the Linguistic One

This is an ongoing work that we performed, so far, only on a subset of the moves. Every move was first analyzed 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 prosody. In the second one, however, some light laughter is shown. This variation of prosody 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 dietary habits declared by the subject. In the move, the subject talks about self, initially with a negative prosody, then with a neutral one, and finally with a friendly prosody. Overall, this change of prosody 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 introduces, at the same time, a negative prosody 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 prosody* 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 and Future Work

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 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 produce some results 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 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). Statistical methods have their advantages in enabling a quick analysis of the data distributions. However, 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, may insure more careful adaptation. Patterns discovered may be formalized, again in terms of decision rules. In the near future, we plan to continue this work by collecting more dialogues, to overcome the sparse data problem. In addition, we are focusing our present activity on the recognition of positive and negative comments with sentiment analysis methods. The main idea is to consider the language processing methods which have been applied to opinion extraction, to reflect on their limits and on how beliefs may be inferred gradually, in conditions of uncertainty and by carefully considering various forms of context (de Rosis and Novielli, 2007).

References

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

Ang, J., Dhillon, R., Krupsky, A., Shriberg, E. and Stolcke, A.: Prosody-based automatic detection of annoyance and frustration in human-computer dialog. In: ICSLP, pp. 2037-2040 (2002)

Bailenson, J.N., Swinth, K.R., Hoyt, C.L., Persky, S., Dimov, A.and Blascovich, J.: 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 vol. 14 (4) pp. 379-393 (2005)

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

Bickmore, T., Cassell, J.: 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, pp.1-32 (2005)

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

Carberry, S., Lambert, L., and Schroeder, L.: Towards recognizing and conveying an attitude of doubt via natural language. Applied Artificial Intelligence 16 (7), pp. 495-517 (2002)

de Rosis, F. and Novielli, N.: From language to thought: inferring opinions and beliefs from verbal behavior. In: AISB '07, Mindful Environments Workshop (2007)

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

de Rosis, F., Novielli, N. and Mazzotta, I., Factors affecting the social attitude of users towards an ECA and how it is Worded. Submitted.

Devillers, L. and Vidrascu, L.: Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. In: INTERSPEECH, pp. 801-804 (2006)

Di Eugenio, B.: On the usage of Kappa to evaluate agreement on coding tasks. In: LREC2000: Second International Conference on Language Resources and Evaluation, 441-444 (2000)

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

Lee, C.M., Narayanan, S. S., and Pieraccini, R.: Combining acoustic and language information for emotion recognition. In: ICSPL , pp. 873-876 (2002)

Liscombe, J., Riccardi, G., and Hakkani-Tür, D.: Using context to improve emotion detection in spoken dialogue systems. In: Interspeech (2005)

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

Nicholas, G., Rotaru, M. and Litman, D. J.: Exploiting word-level features for emotion prediction. In: IEEE/ACL Workshop on Spoken Language Technology (SLT) (2006)

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

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

Rettie, R.: Connectedness, awareness and social presence, in: PRESENCE, online proceedings (2003)

Yu, C., Aoki, P.M. and Woodruff, A.: Detecting user engagement in everyday conversations. In: ICSLP, pp. 1329-1332 (2004)