Interpretation of User's Feedback in Human-Robot Interaction

B. De Carolis and G. Cozzolongo

***Abstract*— In this paper we will propose the use of social robots as interface between users and services in a Smart Environment. We will focus on the need for a robot to recognize the user's feedback, in order to respond and revise its behaviour according to user's needs. As we believe speech is a natural and immediate input channel in human-robot interaction, we will discuss the importance of recognising, besides the linguistic content of the spoken sentence, the attitude of the user towards the robot and the environment. In this way, the meaning of the user dialog will be made clear when hardly recognisable by the analysis of the utterance structure. Then, we will present the results of the application of a potential approach used for integrating the linguistic analysis with the recognition of the valence and arousal of the user's utterance. In order to achieve this goal, we collected and analysed a corpus of data to build an interpretation model based on a Bayesian network. Then we tested the accuracy of the model using a test dataset. Results will show that the integration of the linguistic content with the recognition of some acoustic features of spoken sentences perform better in recognising the key aspects of user's feedback.**

***Index Terms*— Intelligent robots applications, Man-Robot interaction.**

I. INTRODUCTION

FOLLOWING Ambient Intelligence (AmI) vision [1], a Smart Environment (SE) has the main aim of facilitating users in interacting with its services by making their fruition easy, natural and suitable to their needs. However, most of the times, user interfaces for handling functions and services of SE require navigation through menu options just to switch off the lights [2] or a complex setting procedure in order to change the behaviour of the environment in typical scenarios.

Smart Environments should assist their users in a proactive and responsive way, trying to recognise user's behaviours and needs so to respond as it is expected from them. Moreover, they should learn from user's feedback and progressively revise their interaction and behavioural rules.

To this aim in a previous project we developed an agent-based system for controlling a smart home environment [3]. In this system, the proactive response of the environment was mainly triggered by sensor data obtained from the user and other aspects in the surrounding environment (e.g. location, noise, temperature, light conditions) reasonably combined with explicit user actions in certain contexts. An intelligent agent acting "behind the scene", was able to infer user needs from his/her actions and, with reference to the current context, to answer the user in an appropriate way by changing the state of the environment [4].

After the first evaluation phase of the project, 80% of subjects declared to feel uncomfortable in interacting with an invisible presence and without explicit control over the home services. They declared to prefer to have an explicit physical interface as a counterpart to the environment, in order to request services, clarify some potential misunderstanding about task execution, express their approval and disapproval and change the behaviour of the environment according to their needs, as for the system illustrated in [5].

Afterwards, according to several research studies about the topic [6] [7], we decided to introduce the figure of an intermediary between the user and the environment; in particular we decided to employ a social robot as interface between the two participants.

Social robots can be thought, on one hand, as a mobile and intelligent interface to the environment system. On the other hand, they embody the role of friendly companions [8][9] improving the level of robots acceptance by humans. Sony AIBO and iCat, PaPeRo for instance, have been created with this purpose [10]-[12].

These kinds of robots are able to communicate and interact with users following a social behaviour. To this aim the robot could express personality and emotions but, most of all, it has to understand the social cues in interacting with the users [13] [14]. However, if we do not want users to consider robots as useless toys besides interacting, entertaining and engaging users, then they should support people in many aspects of daily life.

For instance, it could provide useful information for decision making, remind tasks and scheduled activities and warn users on possible dangers. In this case, it is necessary to consider some requirements such as the awareness about the user and environment situation, the recognition of his/her intentions, the generation of strategies and plans for satisfying the recognized user goals and the monitoring of effects of plans execution, in order to recover from errors.

As a result, the key goal of our research is to develop a SBDI (Social, Belief, Desire and Intention [15]) mind for this type of robot. Therefore, after having designed the basic architecture of a SBDI [16], we started to develop its behavioural models.

Berardina De Carolis is with University of Bari.
E-mail: decarolis@di.uniba.it

Giovanni Cozzolongo is with University of Bari.
E-mail: cozzolongo@di.uniba.it.

Since speech is a natural way for humans to interact with robots [17] [18], we initially faced the problem of interpreting the user's utterances during the interaction dialogue, in order to understand the valence and therefore the effect that the user wants to achieve through the provided feedback as to revise the reasoning behaviour accordingly.

In our system we consider two sources of knowledge coming from the user spoken input: the **linguistic information content** and the **acoustic features** of the utterance. Here we present results of an empirical study aiming at demonstrating the feasibility of this approach for clarifying the user's feedback intent in spoken utterances when interacting with a social robot acting as a mediator in smart environments.

The paper is structured as follows: in Section II the results of the first experiment for evaluating the impacts of social robots in the interaction with an SE are presented. In Section III we explain the motivations for using both linguistic content and acoustic features of the spoken sentence for disambiguating the user's feedback. Then we present how we annotated a corpus of human-robots spoken dialogues, collected during some experiments. Starting from the analysis of this corpus, we built a model aiming to interpreting the feedback intention of the user. This model is described in Section V. Then, we tested the accuracy of the model in properly recognizing the category of user's feedback. Conclusions and future work directions are discussed in the last Section.

II. UNDERSTANDING COMMUNICATIVE INTENT

In order to express their intentions humans use words and transfer emotions and emphasis by modulating their voice tone. In fact speech conveys two main types of information: it carries linguistic information according to the rules of the used language and paralinguistic information that are related to acoustic features such as variations in pitch, intensity and energy [19]-[23].

Usually the first component conveys information about the content of the communication and the second one about the user's attitude or affective state.

According to several studies [18][22] the linguistic analysis is not enough to properly interpret the real user's communicative intent towards the robot and the environment behaviour. For instance, the user can pronounce the same sentence with different emotional attitudes in order to convey different communicative intents. In fact, if the sentence *"where are you going?"* is pronounced with a negative attitude, it should not be interpreted as "the user want to know where the robot is going" but as "the user is probably disapproving my behaviour" and therefore it should be considered as a negative feedback. In this case, the robot should stop and revise its belief set accordingly.

Another example, typical of domestic environments, is the following: "It is hot in here". According to [24][25] this sentence can be interpreted in different ways:

- the user is *informing* the hearer about his/her perception of the room temperature;
- the user is *requesting* indirectly for someone to open the window;
- the user is *complaining* about the temperature (usually expressed *emotionally*);
- the user is *complaining* about the temperature implying that someone should better keep the windows closed (usually expressed using *sarcasm*).

In this case, the effects that the user wants to achieve are completely different: in one interpretation the user wants someone to open the window, in another one he/she wants to achieve the opposite effect.

Then, while words still play an important role in the recognition of communicative intents, taking into account the user attitude while speaking adds another source of knowledge that is important for understanding the interpretation of the utterance.

Integrating more than one modality should, in fact, help to resolve ambiguities and compensate for errors [26].

Starting from the work of Breazeal and Aryananda [27] and taking into account some theories that state the importance of the voice tone in interacting for educative and training purposes with pre-verbal infant [28] and domestic pets [29] we decided to start investigating how paralinguistic features of the spoken input can be used to improve the recognition of the user's communicative intent.

Research in emotional speech has shown that acoustic and prosodic features can be extracted from the speech signal and used to develop models for recognising emotions. Much of this research has used acted corpus of speech as training data and their research did not take into account the semantic content of what being conveyed [21]-[23].

According to Litman [30], in natural interactions users convey emotions by combining several signals. Therefore, acoustic-prosodic features should not be considered alone but combined with other information sources such as the linguistic content of the spoken sentence. Indeed, while acoustic-prosodic features address how something is said, lexical features represent what is said and, together, these features have shown to be useful for recognising intentions in human communication.

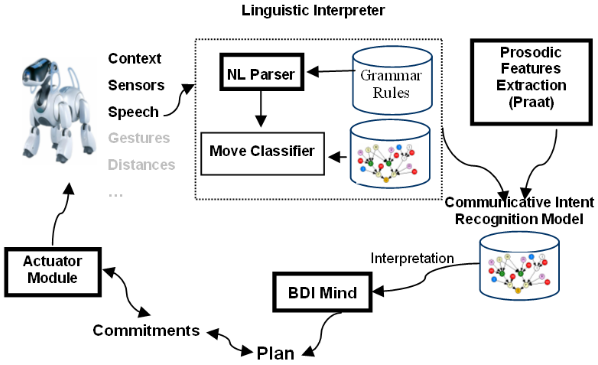


Fig. 1. Reasoning Schema of the Environment Mediator Robot

To this aim we have coupled a linguistic parser that transforms the sentence in a speech act with an acoustic analyzer able to extract the prosodic features of the user spoken input. Then, using a probabilistic model based on a Bayesian Network (BN), our system infers the user's intention by combining these two knowledge sources. BN allows handling uncertainty and incomplete world representation typical of this kind of domain [31]. In fact, interaction with smart environments contains various uncertainties that are mostly caused by the imperfectness and incompleteness of data or by the difficulty of a certain correct interpretation of human behaviour.

Figure 1 illustrates the schema architecture of the speech based intention recognition module integrated in a general BDI reasoning model [32].

The transcript of the spoken input is analyzed by the linguistic parser in order to recognize linguistically the type of user's move, then contextual features and the presence of cue words are used to provide a better classification of the move (i.e. inform, ask, etc.) using a BN model that will be illustrated later in the paper. At the same time the audio file relative to the user‘s move is analyzed using Praat functions [33].

Once we have extracted the parameters relative to pitch, intensity and energy of the sound, these, together with the interpreted category of move become the evidences or the initial distribution (in case of uncertainty in parsing the sentence) of some of the variables of the BN that allows inferring the valence and intention relative to the user's move.

Since the main aim of this research does not specifically concern the prosodic features analyser but the one of understanding how this integration can improve intention recognition, we will focus the rest of the description on model building and on its validation in the context of the interaction with a smart environment.

III. Corpus Annotation

There is a substantial body of literature on emotion recognition from speech [19]. These works show how certain emotional states are often correlated with particular acoustic features such as pitch (fundamental frequency F0), energy, timing and voice quality [34].

Obviously, the quantity and types of features to consider depend on how many emotions or attitudes are relevant for the purpose of the system.

For instance, in the work of [27] which is also aimed at recognising affective intent in human-robot interaction, it was important to distinguish among four categories of affective intents: approval, prohibition, attention, comfort. These categories were selected since the main aim of the work was to "educate" the robot like parents do with their children [28]. According to these results, pitch mean and energy variation seem to work with a good accuracy in distinguishing between the mentioned categories.

An experiment similar to the one presented in this paper was performed by Batliner et al [35] for recognizing emotions in children speech. Their system based on Neural Networks was able to classify speech with an overall recognition rate of about 70% for all four classes: *neutral* without any labels, *laughter, vocative* and *marked* for any other combination of prosodic peculiarities.

We did not consider children speech in our experiment because the potential users of our domotic application were adults. However, in the optic of extending the use of this application to families, children speech should be considered.

Another important research work aiming at recognizing the "social attitude" of users towards an Embodied Conversational Agent is represented by [36]. The adopted approach is also based on corpus annotation and analysis of linguistic and prosodic features in order to classify the user's attitude towards the agent.

In our research, as a first step for building the recognition model, the following classes of variables were considered, as they were deemed to be very important for achieving our goal:

– the typical *input moves* used to provide feedback during the interaction with the robot in a smart environment;
– the *feedback intention* to be recognized;
– the user *attitudes* that could influence intention interpretation.

A. Data Collection

In the data collection phase, our aim was to identify the basic set of user moves in terms of speech acts, communicative intents, attitudes, and to discover the relations among them.

In collecting the corpus dataset, we could use two approaches based on "acted" or "natural" moves. Obviously natural collected corpus is 'pure' and 'real' in terms of emotional content when compared with elicited or simulated content. However, it is difficult to collect these data in a controlled experiment setting. In fact, as far as concerns naturally-occurring behaviour, simply observing and recording the user's moves "in the wild" would produce a collection of data in which cause-effects and other types of relations are difficult to be generalized in machine understandable patterns [37].

A way to avoid this problem and to keep the naturalness in the user behaviour could consists in eliciting data in particular scenario that allow to control as many as relevant factors as possible [38]. Therefore we decided to collect data using this approach.

For collecting scenario-driven data we performed a set of Wizard of Oz (WOZ) studies [39]-[41].

In this experimental setting, we considered our research laboratory as a smart environment. We thought it was an appropriate place since it was well equipped (computers, microphones, cameras, wireless connection, bluetooth, etc.) for performing these types of experiments.

In this experiment we used two groups of 8 subjects each with an age between 20-28 years old equally distributed in gender and background. For both groups the lab room conditions were the same: the air condition was off, the windows were closed and the room was quite hot, the light was on. The Wizard was performed by the same person in all the experiments.

We assigned to each group the same goal:

finding a paper a professor left in the lab and complete the set of references within 10 minutes using a computer connected to the Internet. When finished, the subject had to send an email to the professor with the paper attached.

However, the behaviour of the robot was different for the two groups.

For the first group of subjects AIBO was behaving in a *cooperative* way: the robot was helping the user in achieving his/her goal, providing explanations. In particular AIBO was greeting the subject, showing him/her where the professor left the paper and which computer could be used to search for the missing references. During this subtask AIBO was near the subject observing what he/she was doing, ready to provide help on request without disturbing the subject during task execution.

For the second group AIBO was behaving *not cooperatively*. In particular, AIBO was not helping the user in achieving his/her goal by ignoring the subject requests for three times before answering. Moreover, when the subject was busy in completing the missing references of the paper AIBO was trying to distract him/her from task execution.

In our opinion, this difference would elicit two sets of users' behaviours that could emphasise interaction situations in which the subject approved or disapproved the robot behaviour.

In all the experiments the robot was controlled through an interface by the same person who guided its behaviour following a working script to evoke the previously mentioned situations. To this aim we have developed a tool that helps not only to drive the robot, but also to record the history of the interaction, included the user voice and the robot moves, beside all the data received from the robot sensors.

Before starting with the experiments we administered to each subject a simple questionnaire aiming at collecting some personal data (age and gender) and at understanding their background (department, year of course, artificial intelligence background).

Following the described approach we collected a corpus of 592 moves (25 moves on average for subjects belonging to the first group, 49 for those belonging to the second one).

Each move was recorded using a wireless microphone whose output was sent to the speech processing system. We assigned to each utterance, a unique identifier, the correspondent transcript and the related 16-bit single channel, 8 kHz signal (in a .wav format).

B.Data Annotation

This corpus was annotated in a previous phase of the project as reported in [42], but with two annotators only for each step. In order to improve the reliability of the annotation process, we decided to use more people for each step. Moreover, we added a new annotation step related to the presence of cue words that are typically used when conveying a particular communicative intent. Then, we re-analyse all the data.

Since we are interested in finding the relations between the type of user move, the use of cue words, the attitude and the feedback intention, we divided the annotation process into 4 steps using different annotators for each step so that the labelling of each component was not influenced by the others. Then we completed the annotation by adding the acoustic features related to each sentence.

Examples of collected user moves are listed in Table I.

TABLE I
SAMPLE OF COLLECTED FEEDBACK MOVES TRANSLATED IN ENGLISH.

User spoken sentences
Hi AIBO.
It's hot in here.
What are you doing?
Which is the computer I'm supposed to use?
Where are you going?
OK.
Yes/No.
Thanks.
You are stupid.
Well done.
Good.
Can you show me where is the computer?
Show me where is the computer.
Speed up.
I told you several time don't bother me.
Please stop bothering me.
AIBO.
…

Step 1.

As far as the type of linguistic content is concerned, our aim was to formalize the spoken sentence as a Speech Act.

This formalization was chosen since its semantics is related to effects that the speech act may achieve in the user mental state and therefore it can be used to let our robot reasoning about the beliefs that induced the user to say something or about the effects that he/she wanted to achieve in the mental state of the robot.

For this annotation step we chose the set of categorical labels listed in Table II. This list is a subset of the Speech Act family.

In this annotation phase we needed people that were familiar with the speech act theory. Therefore, we involved five annotators distributed as follows: a professor teaching computational linguistic, two Ph.D. students researching about these topics and two under-graduated students of the computational linguistic course.

TABLE II.
CATEGORIES OF USER MOVES

Move Type	Function
Greet(U,AIBO)	The user U greets AIBO
Call(U,AIBO)	U calls AIBO
Request(U,AIBO,a)	U asks AIBO to perform an action a
Order(U,AIBO,a)	U orders to AIBO to perform an action a
Inform(U,AIBO,f)	The user informs AIBO about a fact f.
Ask(U,AIBO,f)	U makes a question to AIBO about a fact f.
Thank(U,AIBO)	U thanks AIBO.
Reproach(U,AIBO)	U reproaches AIBO's behaviour
Compliment(U,AIBO)	U makes a compliment to AIBO.
Acknowledge(U,AIBO)	U acknowledges AIBO's behaviour

We provided the annotators with the set of human written transcripts of all subjects moves collected during the WOZ study in both modalities and we asked them to use the labels in Table II to annotate them. They could also introduce new labels if they did not recognize in the move any of the listed speech acts.

In order to test the validity of our findings we used the method found in [43]. Then, to have a measure of the level of agreement between annotators, we calculated the percentage of cases that were labelled in the same way, we computed the percentage of agreement and then we calculated the Kappa statistics**.**

Kappa is an index which compares the agreement against what might be expected by chance. Kappa can be thought of as the chance-corrected proportional agreement and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement). This index is widely accepted in the field of content analysis and allows different results to be compared.

Table III shows the frequency of every move category in the corpus and summarizes inter-annotator agreement. The annotators agreed in recognising most of the moves from their linguistic content. Only the level of agreement about the "inform" and "request" speech acts was lower than we expected.

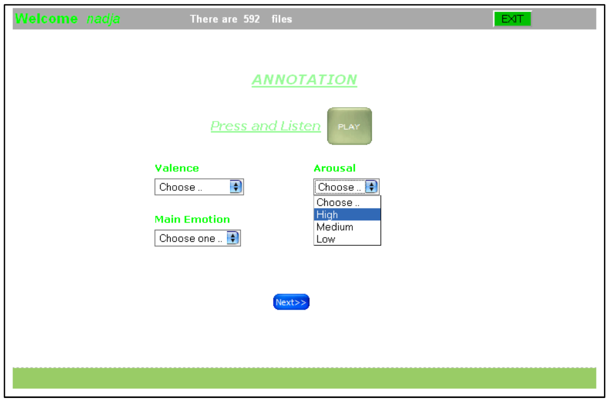


Fig. 2. Annotation web page

Step 2.

In this step our main goal was to identify which was the user's attitude in the feedback sentence that could change the linguistic interpretation of the underlining intention.

In this annotation phase we did not need people with a specific knowledge, as in the *step 1*, therefore we used a website (Figure 2) and we invited 25 people to annotate the corpus.

For our purpose, we did not need a very sophisticated distinction between all the possible user affective states.

In our opinion, in this type of interaction it is important to understand which is the **valence** of the user attitude, since it is related to the achievement of a particular goal and the **arousal**, since it indicates how important was that goal for the user and therefore indicates the urgency to recover that state.

In the case of feedback sentences these two dimensions are more important than recognising the emotion itself.

For instance, a negative valence will indicate a failure in the achievement of the user's goal and if it is correlated with a high arousal will allow to distinguish hanger towards an event (e.g.: a move of the robot), from sadness related to a personal mental state.

TABLE III.
INTER-ANNOTATOR AGREEMENT

Label	% agreement	%frequency	Kappa
Greet	0,9	12%	0.8
Call	0.8	9%	0.6
Request	0.7	15%	0.4
Order	0.8	10%	0.6
Inform	0.6	4%	0.2
Ask	0.8	13%	0.6
Thank	1	9%	1
Reproach	0.8	13%	0.6
Compliment	0.8	7%	0.6
Acknowledge	0.9	8%	0.8

Therefore, for the valence dimension we considered the *positive*, *neutral* and *negative* categories measured along a 5-point scale (from 1- very negative to 5- very positive).

The *positive* valence was important to recognize positive feedback towards the robot and the environment while the *negative* one was important to recognize the disappointment towards the robot behaviour. The *neutral* attitude was considered important for interpreting the user intention only from the linguistic part of the user input.

The arousal was measured in a 3-point scale from high to low. The annotators had to listen to the audio file and set the appropriate value in the drop down menu (Figure 2). While the setting of the arousal and valence was compulsory, annotators could optionally set the main emotion recognized in the sentence. This label, however, is still not considered in building the model. But it could be used in a future phase of the project. The resulting annotation was then stored in a database. Results are shown in Table IV-a and IV-b.

TABLE IV-A.
VALENCE

	% agreement	Kappa
very positive	0.8	0.6
Positive	0.58	0.166
Neutral	0.55	0.117
Negative	0.75	0.5
very negative	0.95	0.9

TABLE IV-B
AROUSAL.

	% agreement	Kappa
High	0.9	0.8
Medium	0.5	0
Low	0.7	0.4

Apparently the annotators clearly agreed in recognizing strong attitudes (very positive and very negative valence and high arousal) while we cannot say that for neutral and positive valence there was a different agreement from the one expected by chance. Moreover, a negative attitude was better recognized than a positive one.

Step 3.

This step was aiming at understanding which categories of feedback were relevant in our domain. Another group of 25 annotators were invited to annotate the intention underlying the feedback sentence by listening to the audio on the web site. The labels used in this case represent the effect that a speech act may achieve (its perlocutionary part)[24]. In this way it was in theory possible to relate the recognised move (linguistic part) and its prosodic characteristic with the move effect.

We identified the ones reported in the first column of Table V together with the percentage of agreement and the results of the Kappa statistics.

TABLE V.
RESULTS OF THE INTENTION ANNOTATION PHASE.

Feedback Type	***Function***	***agreement***	***kappa***
WantToDo(U,AIBO,*a*)[a]	The user U want the robot to do some action	0,75	0,5
WantToKnow(U,AIBO,*f*)	The user U wants to know a fact.	0,8	0,6
KnowAbout(AIBO,*f*)	The user U wants that AIBO knows about a fact.	0,7	0,4
Approve(U,AIBO)	Positive feedback corresponding to the intention to approve;	0,75	0,5
Disapprove(U,AIBO)	Negative feedback corresponding to the intention to disapprove;	0,8	0,6
GetAttention(U,AIBO)	The user U wants to get the attention of the robot;	0,9	0,8
SocialCue(U,AIBO,*c*)	The user U performs a social communicative act such as greeting, soothing, etc..	0,75	0,5

The overall agreement is good, showing that listening to the sentences (linguistic content and acoustic features) allows interpreting its meaning and the user's feedback intention with a good accuracy.

Moreover, in order to understand the correspondence

TABLE VI.
RELATION BETWEEN SPEECH ACTS AND FEEDBACK INTENTIONS.

	Greet (SC)	***Call (GA)***	***Request (WTD)***	***Order (WTD)***	***Inform (KA)***	***Ask (WTK)***	***Thank (SC)***	***Reproach (DA)***	***Compliment (A)***	***Acknowledge (A)***
WantToDo (WTD)	0	0	61	47	0	0	0	0	0	0
WantToKnow (WTK)	0	0	0	0	0	47	0	0	0	0
KnowAbout (KA)	0	0	0	0	7	0	0	0	0	0
Approve (A)	0	0	0	0	10	0	36	0	27	42
Disapprove (DA)	0	13	28	12	11	31	10	59	8	0
GetAttention (GA)	0	32	0	0	2	0	3	18	0	0
SocialCue (SC)	71	8	0	0	1	0	4	0	6	6

between the speech acts annotated on the corpus transcripts and the intentions recognised from the spoken sentences we calculated a matrix showing the relations and the differences between the two annotations.

In this matrix, shown in Table VI, the columns indicates the speech acts and their intended communicative effect as formalized in the Speech Act Theory and the rows the categories of feedback that we are considering in our domain.

The results seem to confirm our hypothesis and therefore the need to integrate this two knowledge sources (the linguistic and the acoustic component of the sentence) for recognising the user intention. In fact, almost all the communicative intentions can be expressed using different speech acts. But, while it is easy to classify a *thank* or a *greet* as the intention to provide a "Social Cue", a "disapprove" can be expressed using almost all the categories of speech acts. As we will see later on, the integration of the linguistic interpretation with the prosodic features allows recognising it with a quite good accuracy.

Step 4.

The goal of this step was to recognize in the transcript of the collected moves the presence of cue words that could help in identifying the nature of the feedback, according to the categories in Table VII. Cue words could belong to an **emotional lexicon** or could be just words with a clear meaning, such as a verb or an object.

We classify as belonging to the **emotional lexicon** all the words, in a language, containing information about an emotional state in their semantic representation. This could be either an emotion or a feeling. Besides exclamations (i.e. "oh, damn!"), cue words could be nouns (fear, thanks, etc.) or verbs (to hate, to love), or adjectives (angry, furious, sad, happy, etc.).

In our domain, cue words are classified as belonging to these categories:

– *social*: i.e. hi, ciao, hello, thanks, AIBO, etc.
– *attention*: i.e. AIBO, hei, you, aho, etc.
– *positive*: i.e. well, good, yes, love
– *formal*: i.e. please, could
– *negative*: no, stop, bad, stupid, hate
– *action verbs*: move, do, go, speed up, help, etc.
– *conversation verbs*: tell, know, etc.

Table VII shows the relation between the category of intention and the category of cue words with the level of agreement.

Step 5.

After the human annotation process, the dataset needs to be completed with the related acoustic features. We used Praat functions in order to perform a macro-prosodic or global analysis and to extract from the audio file of each move features related to:

TABLE VII.
RELATION BETWEEN INTENTIONS AND CUE WORDS.

Feedback Type	cue words categories	agreement	kappa
WantToDo(U,AIBO,a),	formal, action verbs	0.7	0,4
WantToKnow(U,AIBO,f)	formal, conversation verbs	0.75	0,5
Approve(U,AIBO)	positive	0.7	0,4
Disapprove(U,AIBO)	negative	0.75	0,5
GetAttention(U,AIBO)	attention	0.8	0,6
SocialCue(U,AIBO,c)	social	0.9	0,8

- The variation of the fundamental frequency (f0): pitch minimum, mean, maximum and standard deviation;
- The variation of energy (RMS): max and standard deviation.

We did not consider the speech rate, because we are going to interpret very short sentences and move on average in our domain.

As we have already pointed out we consider the *activation* and the *arousal*, which can be used to distinguish high activation emotion from low activation one, by analysing the value of the *f0*; the *average f0* values, its *range,* its *dynamics* and *intensity* are all higher for high activation emotion. In this way we can distinguish *surprise*, *joy*, *anger*, *fear* from *sadness* and *disgust*.

The *valence* dimension allows distinguishing positive from negative emotions. Negative emotions features are characterized by fast f0 lowering, rising in intensity, high prominence of the maximum energy values, longer pauses. On the basis of this features it is possible to distinguish fear and anger from joy.

After adding these features to the dataset, it was necessary to transform the numeric values relatives to the pitch and energy into discrete values, in order to handle these data in our model.

To this aim we used a three-value scale (low, medium, high). In order to assign each numerical value corresponding to the pitch and the energy values to one of these discrete values, we calculated the 33% and 66% percentage. We divided in this way the numeric interval of each of the extracted features into three parts. Then, values falling into the first numeric set were considered as low, those falling in the second one as normal and the rest high.

Step 6.

Then, we completed the annotation of the collected corpus with the following data: the name of the audio file, the gender of the subject that pronounced the sentence, the event that triggered the user sentence, the user and environment situation. These additional data were recorded during the experiment.

The following Table VIII shows an annotated element of the corpus corresponding to the question “where are you going?” expressing the intention to disapprove the robot action.

TABLE VIII.
AN EXAMPLE OF THE ANNOTATION FOR AN ELEMENT

Name	Value
File	Aibo10
Sentence	where are you going?
Gender	f
Move	ask
Env context	c 10
Event	AIBO makes noise
Feedback	disapprove(U,AIBO,move)
Valence	1
Arousal	3
P mean	High
P max	High
P min	High
P var	Normal
E max	High
E var	High

The variables *p_min*, *p_max*, *p_mean*, *p_var* identify the values of the pitch minimum, maximum, medium and its variance, respectively. The values of the maximum energy and its variance are identified with *e_var* and *e_max* respectively.

IV.BUILDING THE MODEL

In order to learn the dependencies among acoustic features, linguistic content of the user move, attitude and intention, we decided to use a probabilistic model, as it seems to be appropriate in the considered context. Indeed, understanding human attitude and intention from speech input involves capturing and making sense of imprecise, incomplete and sometimes conflicting data [31].

As far as concern the building of the model we could use a learning algorithm, able to learn both the structure and the parameters of the network, or we could design the structure according to the theories that guided our experiment steps and then learn the parameters accordingly.

In a first phase of the project we used the NPC learning algorithm of Hugin 6.5 (see www.hugin.com for more details) on the labelled database of cases in the corpus. In this way we could learn both the structure and the parameters of the network.

We randomly extracted 37 cases from the labelled database of selected cases, because 37 was the average of moves for each subject collected during the experiment.

As we will show later on, this subset was used in the testing phase, while the rest of cases (555) were used for learning the network.

The model resulting after some optimization steps is shown in Figure 3.

Variables in this network are mainly related to:

- the recognized **move category**: this information is extracted by the linguistic parser and belongs to one of the categories listed in Table II;
- the presence and category of **cue words** that may change the linguistic interpretation of the move; this variable may then assume values in the set described in Table VII;
- the **environment context** situation that may have triggered the feedback move (i.e. the room temperature is 32° C);
- the **robot action**;
- the **valence** associated to the move that can assume values in the set very positive (5), positive (4), neutral (3), negative (2) and very negative (1);
- the **arousal** associated to the move that can assume values in the set high (3), medium (2) and low (1);
- the extracted **acoustic features**: pitch features and energy variation and maximum were considered relevant to the recognition;
- the **feedback intention** beyond the speech act: this is the variable that we want to monitor using the model and can assume values in the set described in Table VI.

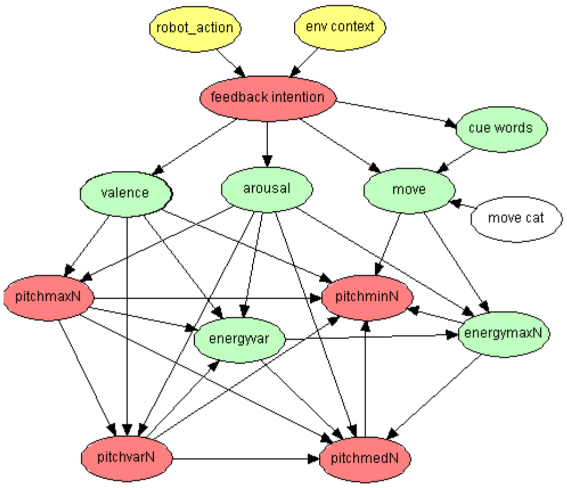


Fig. 3. The structure of the feedback recognition model.

In learning the model, we have set some constraints in the higher level of the structure. In particular we set the dependence of the *feedback intention* node with the four main variables resulting from the annotation process: the valence, the arousal, the move category and the cue word. These three variables, then, are related to the acoustic parameters whose relations are represented in the lower section of the network. This part has been completely learnt both in the structure and the parameters.

Further more, as we wanted to use the model in a more dynamic way, during the interaction, we had to link the *feedback* to the *robot_action* and to the *env_context* variable. This two variables could help in identifying the cause of the user feedback.

For examples, the user could disagree about the temperature situation in the room by saying: “it is hot in here” with a negative attitude. In this case the intention to disapprove does not regard directly the robot behaviour but the comfort situation of the environment. While, in case the user reproaches the robot when it moves or makes some noise, the disapproval regards its action.

V.EVALUATING THE MODEL

The evaluation of the model was performed on a subset of data extracted randomly from the selected corpora. As we said in the previous Section, we selected randomly 37 entries from our annotated corpus before the learning process (37 was the average of moves for each subject).

Then as input evidences of our model we used the following values: first **only** the **speech features**, then only the **linguistic content** of the **move** and then **both** of them. Finally we compared the predicted results with the human annotation in all the three cases.

The global accuracy of the model, expressed as the total of correct predictions for every category on the total of the moves in the dataset, is of about 60% in the case of speech features only or linguistic content only, while it increases to the 75% when considering both features.

Since we use a probabilistic model we considered as correct a prediction of the value of the feedback intention variable the one of its six states with the highest probability.

Table X reports detailed results about the prediction of each feedback category in the three cases.

We used *F1-score* that combines *precision* and *recall* to evaluate the performance of the model. For a feedback intention category fc_i , we define *precision* as the ratio of the number of moves classified correctly as fc_i and the number of moves in the corpus annotated as fc_i . We define *recall* as the ratio of the number of moves classified correctly as fc_i and the total number of moves in the corpus belonging to the class fc_i . The closer the precision, the recall and F1-score are to 1, the more accurate is the prediction.

As a general comment we can say that the model performs a better prediction when using speech + move data vs. the other two conditions, especially in those cases when the linguistic content contrasts the voice tone in identifying the type of feedback intention. This is evident in the case of the “Approve” and “Disapprove” class where the recognition from the linguistic content only performs worse.

For instance, considering the case in which the robot moves towards a place despite the user will, the sentence:

“where are you going?”

is parsed as

“Ask(U,AIBO,where(is_going,AIBO))”

with a cue word belonging to the *action_verb* category.

The model, in this case, classifies the feedback intention as:

“Want_to_Know(U,where(is_going,AIBO))”

taking the state of the variable with the highest probability (47% on 6 possible choices), while the probability of the disapprove intention is around 16%.

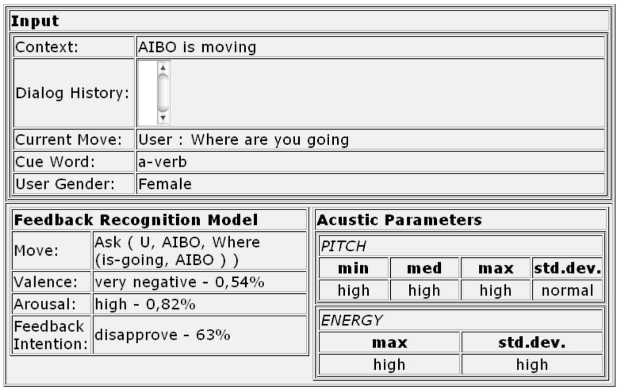
However, if we take into account evidences about acoustic features, then the most probable feedback intention becomes the disapproving one (63%) and the most probable valence of the voice tone becomes negative (54%) with an high arousal (82%).

To monitor the behaviour of the robot we have developed a

TABLE IX.
CONFUSION MATRIX FOR INTENTION RECOGNITION - SPEECH ONLY/TEXT ONLY/ SPEECH+TEXT .

		WantToDo	*WantToKnow*	*KnowAbout*	*Approve*	*Disapprove*	*GetAttention*	*SocialCue*	*F1-score*
WantToDo	Speech	55	31	0	0	20	2	0	0.59
	Text	69	37	0	0	0	2	0	0.62
	S+T	74	21	0	5	8	0	0	0.77
WantToKnow	Speech	10	26	0	0	11	0	0	0.41
	Text	17	30	0	0	0	0	0	0.41
	S+T	9	32	0	0	6	0	0	0.59
KnowAbout	Speech	0	0	4	1	2	0	0	0.44
	Text	0	0	5	0	0	2	0	0.5
	S+T	0	0	4	1	1	1	0	0.61
Approve	Speech	0	0	3	71	0	19	18	0.72
	Text	11	12	5	51	0	17	15	0.61
	S+T	0	0	1	85	0	10	15	0.83
Disapprove	Speech	13	17	3	0	119	17	0	0.68
	Text	17	20	3	0	89	21	19	0.66
	S+T	2	9	1	0	136	9	12	0.81
GetAttention	Speech	1	5	1	0	15	29	4	0.43
	Text	0	0	0	0	12	38	5	0.52
	S+T	0	0	0	0	9	41	5	0.66
SocialCue	Speech	0	0	0	13	15	15	52	0.61
	Text	0	0	0	5	0	11	79	0.74
	S+T	0	0	0	3	7	9	76	0.75

tool. Figure 4-A shows how this tool allows monitoring the result of the propagation of the evidences in the model in the described example.



Input

Context:	AIBO is moving
Dialog History:	
Current Move:	User : Where are you going
Cue Word:	a-verb
User Gender:	Female

Feedback Recognition Model

Move:	Ask (U, AIBO, Where (is-going, AIBO))
Valence:	very negative - 0,54%
Arousal:	high - 0,82%
Feedback Intention:	disapprove - 63%

Acustic Parameters

PITCH

min	med	max	std.dev.
high	high	high	normal

ENERGY

max	std.dev.
high	high

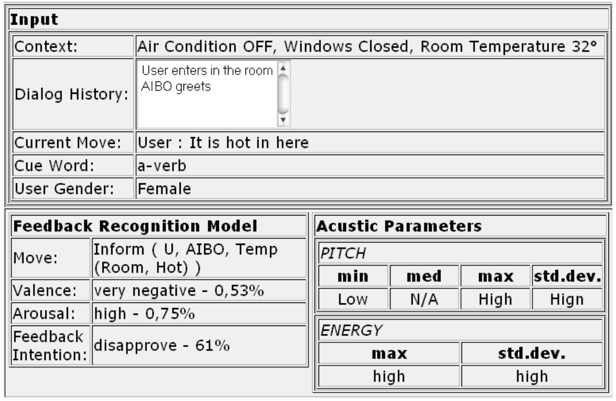
Fig.4a The result of the model prediction after the propagation of evidence values.

Figure 4-B shows another example of prediction. In this case the user enters in the room and says with a negative attitude:

"it is hot in here"

in this case the model recognizes another disapproving feedback about the environment situation even if the move is classified linguistically as an inform speech act.

This is inferred by considering the very negative valence and high arousal deriving from the prosodic features and the information coming from the environment context.



Input

Context:	Air Condition OFF, Windows Closed, Room Temperature 32°
Dialog History:	User enters in the room AIBO greets
Current Move:	User : It is hot in here
Cue Word:	a-verb
User Gender:	Female

Feedback Recognition Model

Move:	Inform (U, AIBO, Temp (Room, Hot))
Valence:	very negative - 0,53%
Arousal:	high - 0,75%
Feedback Intention:	disapprove - 61%

Acustic Parameters

PITCH

min	med	max	std.dev.
Low	N/A	High	Hign

ENERGY

max	std.dev.
high	high

Fig. 4b. Another example of prediction.

VI.CONCLUSIONS AND FUTURE WORK

As part of ambient intelligent research, there is a need to recognize the user's intention during the interaction in order to adapt the environment behaviour accordingly. The interaction may happen in a seamless way, for instance combining sensors data, or through an embodied agent that acts as a "mediator" between the user and the environment.

In this paper we presented our results in recognising user's feedback intention when this mediator is represented by a social robot that is considered responsible for the success of interaction between the user and the environment.

Since spoken input is considered one of the more natural ways to interact with robots, we focused our research on the analysis of the spoken sentence using two information sources: from the linguistic content and intonation of the voice. In particular, we performed one set of experiments based on the Wizard of Oz method whose results were annotated and analyzed in terms of linguistic communication content, presence of cue words, valence and arousal of the voice tone and intention.

This collected corpus was used to learn the structure of the Bayesian network model to be used for recognising the probability for a user to have a particular intention toward the robot and/or the environment.

In order to validate this model we used as testing dataset a subset of moves randomly extracted from the corpus.

The performed experiment proved that using both knowledge sources for recognising the user's intention improves the prediction accuracy of the model.

The same approach could be used in other domains such as e-learning, e-health, etc. where user's emotion and intention recognition plays a key role for adjusting the behaviour of the system. In e-learning, for instance, the student engagement is constantly monitored [30]. Understanding user's feedback is a crucial aspect of effective learning environments, therefore monitoring continuously and unobtrusively student's frustration, boredom, enthusiasm, is important for tuning the presentation of learning material, determining the success of the learning process [44].

In our future work we plan to integrate this model in the architecture of NICA (as the name of the project Natural Interaction with a Caring Agent). NICA is a robot aiming at assisting elderly people in their daily life tasks. NICA has to be able to provide what we collected as two types of results, one showing a service-oriented assistance but also social and emotional feedback. For this reason it is important to extract from the user spoken sentence information about his/her attitude towards the robot and the environment.

In this application domain, we need to investigate on how previous intentions influence the current one so as to express this relation as a function to build a temporal link in a Dynamic Belief Network (DBN, [34]). Moreover, we want to use the recognised feedback intention as a triggering condition for planning the most appropriate robot's behaviour.

Results obtained so far from the overall response of users to the experiments seem to confirm that the idea of using a social robot as a mediator is a good way to overcome barriers that people may find in using smart environments.

ACKNOWLEDGMENTS

We wish to thank Vincenzo Silvestri and Irene Mazzotta for their help during the experiments.

REFERENCES

[1] G. Riva, F. Vatalaro, F. Davide, and M. Alcaniz, (Editors). “Ambient Intelligence: the Evolution Of Technology, Communication And Cognition Towards The Future Of Human-Computer Interaction” (Emerging Communication), 2005.

[2] D. Randall, “Living Inside a Smart House: A Case Study,” in Harper, R., (editor), Inside the Smart Home, Springer-Verlag, London. 227-246. 2003.

[3] B. De Carolis, G. Cozzolongo, S. Pizzutilo, V.L. Plantamura, “Agent-Based Home Simulation and Control”. In Proceedings of ISMIS 2005, LNCS. Springer: 404-412. 2005.

[4] B. De Carolis, G. Cozzolongo, S. Pizzutilo. “A Butler Agent for Personalized House Control”. In Proceeding of ISMIS 2006, LNCS. Springer. 2006.

[5] A. Gárate, N. Herrasti, and A. López, “GENIO: an ambient intelligence application in home automation and entertainment environment”. In Proceedings of the 2005 Joint Conference on Smart Objects and Ambient intelligence: innovative Context-Aware Services: Usages and Technologies (Grenoble, France, October 12 - 14, 2005). sOc-EUSAI '05, vol. 121. ACM Press, New York, NY, 241-245. 2005.

[6] G. von Wichert, and G. Lawitzky. “Man-Machine Interaction for Robot Applications in Everyday Environments”. In Proc of the IEEE Int. Workshop on Robot and Human Interaction 2001 (RO-MAN 2001), Paris, Bordeaux, Frankreich, 18.-21. Sept. 2001.

[7] S. Buisine, and J.C. Martin, “Experimental Evaluation of Bi-Directional Multimodal Interaction with Conversational Agents”. Proceedings of Interact’03. 2003.

[8] K. Dautenhahn “Robots as social actors: Aurora and the case of autism”. In Proc. CT99, The Third International Cognitive Technology Conference, August, San Francisco, pages 359-374, 1999.

[9] T. W. Bickmore and R. Picard. “Establishing and maintaining long-term human-computer relationships”. Transactions on Computer-Human Interaction, 2004.

[10] J. Osada, S. Ohnaka, M. Sato: “The scenario and design process of childcare robot, PaPeRo”. Advances in Computer Entertainment Technology 2006.

[11] C. Bartneck, and J. Reichenbach, “Subtle emotional expressions of synthetic characters”. The International Journal of Human-Computer Studies (IJHCS), 62(2), pp. 179-192. 2005.

[12] Sony. (1999). AIBO. from http://www.AIBO.com.

[13] C. Breazeal. “Toward sociable robots”. Robotics and Autonomous Systems, 42:167–175, 2003.

[14] L. Cañamero and A. Blanchard and J. Nadel J., “Attachment Bonds for Human-Like Robots”, International Journal of Humanoïd Robotics. 2006.

[15] H. Jiang, and J.M. Vidal “From Rational to Emotional Agents”. In Proceedings of the AAAI Workshop on Cognitive Modeling and Agent-based Social Simulation, 2006.

[16] B. De Carolis, G. Cozzolongo, “Social Robots for Improving Interaction in Smart Environments”. In Proocedings of the Workshop Emotion in HCI. In conjunction with the 20th British HCI Group Annual Conference AISB07. 2007.

[17] A. Drygajlo, P.J. Prodanov, G. Ramel G., M. Meisser, and R. Siegwart, “On developing a voice-enabled interface for interactive tour-guide robots”. Journal of Advanced Robotics, vol.17, nr. 7,p.p. 599-616, 2003.

[18] S. Nakamura, Toward Heart-to-Heart Speech Interface with Robots. ATR UptoDate. Summer 2003.

[19] R. Banse, and K.R. Scherer, ”Acoustic profiles in vocal emotion expression”. Journal of Personality and Social Psychology. 1996.

[20] C.M. Lee, S.S. Narayanan, R. Pieraccini, “Combining acoustic and language information for emotion recognition”. Proceedings of ICSLP, 2002.

[21] J. Liscombe, J. Venditti, and J.Hirschberg, “Classifying subject ratings of emotional speech using acoustic features,” in Proc. of EuroSpeech, 2003.

[22] S. Mozziconacci and D.J. Hermes, "Role of intonation patterns in conveying emotion in speech", in Proceedings, International Conference of Phonetic Sciences, San Francisco, August 1999.

[23] P.Y. Oudeyer. “The production and recognition of emotions in speech: Features and Algorithms”. International Journal of Human Computer Studies, 59(1-2):157-183. 2002.

[24] J.L. Austin,. How to do things with words. Oxford: Oxford University Press. 1962.

[25] J. Searle, “Speech Acts: An Essay in the Philosophy of Language”, Cambridge, Eng.: Cambridge University Press. 1969.

[26] W.E. Bosma and E. André, “Exploiting Emotions to Disambiguate Dialogue Acts”, in Proc. 2004 Conference on Intelligent User Interfaces, January 13 2004, N.J. Nunes and C. Rich (eds), Funchal, Portugal, pp. 85-92, 2004.

[27] C. Breazeal and L. Aryananda, "Recognizing affective intent in robot directed speech", Autonomous Robots, 12:1, pp. 83-104. 2002.

[28] A. Fernald, “Four-month-old infants prefer to listen to motherese”. Infant Behavior & Development, 8, 181-195. 1985

[29] J. Serpell, “The Domestic Dog: Its evolution, behavior, and interactions with people”. Cambridge University Press, New York, New York. 1995.

[30] D. Litman, K. Forbes, S. Silliman, “Towards emotion prediction in spoken tutoring dialogues”. Proceedings of HLT/NAACL, 2003.

[31] F.V. Jensen. “Bayesian Networks and Decision Graphs. Statistics for engineering and information science”. Springer, New York, Berlin, Heidelberg, 2001.

[32] A. Rao, M. Georgeff: BDI agents: From theory to practice. In Proceeding of the 1st International Conference on Multi-Agent Systems (San Francisco, CA, June 1995), V. Lesser, Ed., AAAI Press, pp. 312–319. 1995

[33] PRAAT: www.fon.hum.uva.nl/praat.

[34] I. Poggi and E. Magno Caldognetto, “Il parlato emotivo. Aspetti cognitivi, linguistici e fonetici”. Atti del Convegno Il parlato italiano. Napoli, 12-15 febbraio 2003. Napoli, D’Auria. CdRom, 2004.

[35] de Rosis, F, Batliner, A, Novielli, N, and Steidl, S (2007). ’You are Sooo Cool, Valentina!’ Recognizing Social Attitude in Speech-Based Dialogues with an ECA. In: Affective Computing and Intelligent Interaction, edited by Ana Paiva, Rui Prada, Rosalind W. Picard. Springer Berlin / Heidelberg, pages 179-190.

[36] Batliner, A., Hacker, C., Steidl, S., N¨oth, E., D’Arcy, S., Russell, M., and Wong, M. (2004). “You stupid tin box” - children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. In Proc. LREC 2004, Lisbon.

[37] D. Kulic and E. A. Croft. “Estimating Intent for Human Robot Interaction”. Proc of the Int. Conf. on Advanced Robotics. Coimbra, Portugal, June 29 - July 3, 2003.

[38] F. Enos and J. Hirschberg, “A framework for eliciting emotional speech: Capitalizing on the actor's process”. LREC 2006 Workshop on Corpora for Research on Emotion and Affect, Genova, 2006.

[39] S. Whittaker, M. Walker, M. and J. Moore, “Fish or Fowl: A Wizard of Oz evaluation of dialogue strategies in the restaurant domain”. Language Resources and Evaluation Conference. 2002.

[40] S. Oviatt, and B. Adams, “Designing and Evaluating Conversational Interfaces With Animated Characters”. In J Cassell, J Sullivan, S Prevost and E Churchill: Embodied Conversational Agents. The MIT Press, 2000.

[41] I. Bretan, A.L. Ereback, C. MacDermid, and A. Waern, “Simulation-Based Dialogue Design for Speech-Controlled Telephone Services”. Proceedings of CHI’95. 1995.

[42] G. Cozzolongo, B. De Carolis, S. Pizzutilo, “Social robots as mediators between users and smart environments”. In Proceedings of the 12th international Conference on intelligent User interfaces (Honolulu, Hawaii, USA, January 28 - 31, 2007). IUI '07. ACM Press, New York, NY, 353-356. 2007

[43] J. C. Carletta, “Assessing agreement on classification tasks: the kappa statistic”. Computational Linguistics, 22(2), 249-254. 1996.

[44] D’Mello S.K., Craig S.D., Gholson B., Franklin S., Picard R.W. & Graesser A.C.: Integrating Affect Sensors in an Intelligent Tutoring System. In Proceedings of the Workshop on Affective Interactions: The Computer in the Affective Loop Workshop, International Conference on Intelligent User Interfaces, 2005, 7-13.