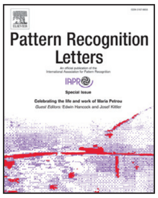
Pattern Recognition Letters 99 (2017) 87–95

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Recognizing users feedback from non-verbal communicative acts in conversational recommender systems

Berardina De Carolis[*], Marco de Gemmis, Pasquale Lops, Giuseppe Palestra

Department of Computer Science, University of Bari Aldo Moro, via Orabona 4, 70125, Bari, Italy

ARTICLE INFO

Article history:
Available online 13 June 2017

MSC:
41A05
41A10
65D05
65D17
65D17

Keywords:
Behavioral analysis
User profiling
Conversational recommender systems
Social robots

ABSTRACT

Conversational recommender systems produce personalized recommendations of potentially useful items by utilizing natural language dialogues for detecting user preferences, as well as for providing recommendations. In this work we investigate the role of affective factors such as attitudes, emotions, likes and dislikes in conversational recommender systems and how they can be used as implicit feedback to improve the information filtering process. We thus developed a multimodal framework for recognizing the attitude of the user during their conversation with DIVA, a Dress-shopping InteractiVe Assistant aimed at recommending fashion apparel. Wee took into account speech prosody, body poses and facial expressions for providing implicit feedback to the system and for refining the recommendation accordingly. The shopping assistant has been embodied in the Social Robot NAO and has been tested in the dress shopping scenario. Our experimental results show that the proposed method is a promising way to implicitly profile the user and improve the performance of recommendations when explicit feedback is not available, thus demonstrating its effectiveness and viability.

1. Introduction

Conversational systems interact with people through natural language dialogues to assist or entertain people. Dialogues can be based exclusively on language or they can be used in conjunction with additional modalities like facial expressions, gestures and so on. When the conversational approach is applied to recommender systems, the main goal of the dialogue is to acquire knowledge about preferences of users in order to support them in decision-making tasks. During the conversation, the system may acquire not only information provided explicitly by the user but also information about affective factors, such as attitudes, emotions, likes and dislikes, by observing user behavior. In this paper we try to address the following research question: can user *non-verbal behavior* be employed effectively as implicit feedback in a recommender system?

The main motivation for this investigation is that non-verbal communicative acts can be considered as important mind-markers, providing information on a person's mental state [21], namely beliefs, goals and emotions. Emotions have a considerable effect on our decision-making and our reactions to stimuli. For instance, upon entering a shopping environment, while looking at items the user might experience a vast set of emotions which can affect her physiological state and result in variations of the voice and facial and bodily expressions [5]. According to the framework proposed in [27], the affective state of the user should be detected during the entire interaction with the system in order to improve recommendations. Moreover, affective factors seem appropriate especially in fashion stores because this experience is strongly characterized by emotional responses to visual stimuli [15] and our emotions provide continuous feedback about our decision processes in the store. These considerations supported our decision to perform the proposed investigation in the domain of fashion recommendation. Our study aimed at analyzing the behavior of users in order to discover their emotional attitude towards apparels in a shop and to infer their preferences accordingly.

The main contributions of this paper are:

1. a study that collected a corpus of interactions between users and human shopping assistants. Each interaction move was annotated according to the liking attitude conveyed by users, as well as with information about purchased items. The resulting dataset is a valuable resource to infer knowledge about the affective state of the user;
2. a model for the recognition of user emotional feedback. The model has been represented as a Dynamic Bayesian Network, due to the ability of this formalism to represent the user mental state by taking into account uncertainty and graduality [18].

[*] Corresponding author.
E-mail address: berardina.decarolis@uniba.it (B. De Carolis).

http://dx.doi.org/10.1016/j.patrec.2017.06.011
0167-8655/© 2017 Elsevier B.V. All rights reserved.

The evidence variables of the model are provided by the signals in the face, body poses and voice prosody;

3. an experiment in which the human shopping assistant was substituted by the social robot and affective factors in the user's communicative behavior were automatically recognized. The results are quite encouraging since the decisions predicted by the model matched those taken by humans in 89% of cases;
4. to show whether and how the model can be exploited to provide recommendations, we applied the framework and the data collected from the previous study to provide implicit feedback to a content-based recommender system [11,14]. Even if collaborative filtering is the most common choice for fashion recommender systems [12], we decided to adopt a simple recommendation model based on item similarity in order to avoid the rating sparsity problem. We used the results of the model (user liking of an item) as an implicit rating. Based on ratings stored in the user profile, we computed the ranking of other fashion items. The results showed that implicit feedback provided by the proposed model can be used to provide accurate recommendations.

As a proof of concepts we developed DIVA (Dress-shopping InteractiVe Assistant), a conversational recommender system based on the proposed model, able to provide recommendations to the user in a fashion store. The agent, embodied in NAO,[1] is able to carry out simple dialogues typical of the domain and to deliver personalized recommendations. The paper is structured as follows. Section 2 provides an overview on related work. In Section 3 we summarize which are the relevant signals in multimodal communication during the shopping experience. In Section 4, we describe the model. Section 5 explains the implementation of the framework for real-time recognition of signals. The results of the evaluation of the model are shown in Section 6. Finally, conclusions and future work directions are presented in Section 7.

2. Related work

Shopping assistant agents have been used in e-commerce web sites with the principal aim of engaging the user and improving trust in the system. In [24], the authors suggest that avatars can enhance the effectiveness of web-based sales. In this case the agent should be designed in order to maximize attractiveness and expertise. [16] describe the design and implementation of a virtual sales clerk capable of providing customers in a 3D on-line shop with advice. To emulate the experience of human face-to-face conversation with their users, these agents use speech, gaze, gesture and facial expressions. Recent related work reported the use of robots to support users' shopping activity. Different studies compared the impact of recommendation and the advice provided by a social robot with respect to virtual agents and results show that robots have a more positive impact on people with respect to 2D/3D virtual agents displayed on a screen [23]. When used in real word environments, robots have a physical presence that seems to attract people and affect their decision-making more effectively than virtual agents [26]. Acceptance of robots with respect to their virtual counterparts is higher since they are perceived as more engaging, credible and informative than virtual agents shown on a screen. Given these findings social robots should be used in the near future in many daily-life application domains (entertainment, education, healthcare, information provision, and so on). At present little related work reports on the use of social robots in shopping malls or as shopping assistants. For example, [4] propose a humanoid mobile shopping robot that helps customers by providing product information, finding products, and guiding them through the shop without any behavioral adaptation based on the user/customer profile. [25] conducted field trials with four social mobile robots in a mall and investigated how they attracted more people to a shop with robots. These studies works evaluated mainly the impact of the developed systems on the users' shopping behavior and interactions between robots and people; they did not focus on the recommendation and personalization context. This issue has been discussed in [13] which uses humanoid robots to recommend specific items to customers according to their purchasing behavior as observed by sensors before arriving at a specific shelf.

Recently, research on personalized interaction and recommender systems is considering the impact that affective factors (emotions, mood, personality) have on the quality of recommendations [28]. Initially, these data were considered interesting but, since emotions were difficult to acquire in real time, they were not used in practice. With recent advances in unobtrusive technologies for the detection of emotions these data are becoming interesting in the domain of personalized systems especially in those concerning music recommendations [9]. These findings motivated our approach to recognize multimodal signals in the users' communicative acts during a dialog with the social robot and interpreting them in terms of emotional attitude towards apparel in a shop and inferring users' immediate interests.

3. Analysis of Non-verbal Communicative Signals

Affective arousal modulates all human communicative signals and facial expressions and body gestures seem to be most important in the human judgment of behavioral cues [1]. Previous findings indicate that to interpret someone's behavioral cues, people rely on facial expressions and to a lesser degree on body gestures and vocal expressions. With the aim of developing a model for recognizing and fusing signals (in the users' behavior) denoting a liking attitude while looking at a dress or when interacting with a shopping assistant, we investigated the typical of this domain. The model has then been tested in the experimental session in which the human shopping assistant was substituted by a social robot and recognition of affective factors in the user communicative behavior was made automatically. To do so we collected and analyzed a dataset of realistic interaction examples in which emotion expressions were as similar as possible to those expressed in a real shopping experience by carrying out a study in which a human shopping assistant supported the interaction with an interactive wall display showing fashion apparel. The collected dataset has been annotated by three human raters and taking that as a starting point we developed a module for recognizing emotional feedback in the dress-shopping domain.

3.1. Collecting the dataset

The study has been conducted in a research laboratory in which an interactive wall display, like one that might be present in a real shop, was installed. We collected, 15 interactive experiences from subjects between 19 and 28 years old, Italian and equally distributed by gender. We assigned to each subject the same goal: finding and buying a dress for the wedding of your best friend. We also fixed a budget of 140 euro in order to motivate subjects to perform the task in a realistic way. Each of them had to buy a complete outfit (shoes, dress, jacket) within the assigned budget. Subjects interacted with a kind of Virtual Shopping system (see Fig. 1) that displayed complete outfits (like on the mannequin) on the wall by using Kinect-like gestures they could browse among items, see more details and buy items. A human shopping assistant was available (the same for all subjects) to provide information and suggestions about dresses. Before starting the study we

[1] https://www.ald.softbankrobotics.com/en/cool-robots/nao.

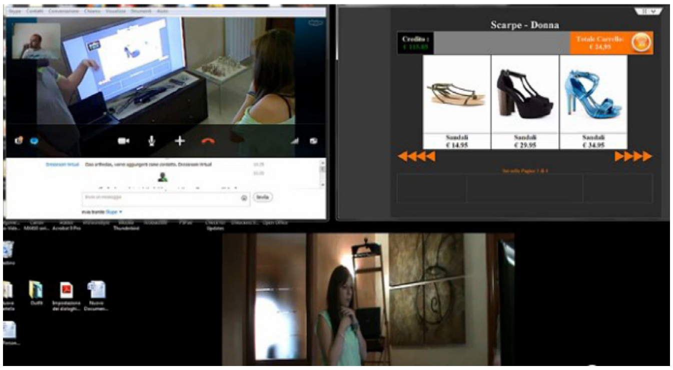


Fig. 1. An example of interaction during the study.

administered to each subject a simple questionnaire aimed at collecting some personal data (age, gender and education) and their level of interest in being trendy and well-dressed. The facilitator of the study briefly explained to each subject how to use the system and let her familiarize herself with it using items that were not included in the item-set of the study. After this assessment and briefing phase, subjects were asked to give their consent to video-record their behavior during the study. Then, they could start interacting with the system. After the study each subject was fully debriefed with an open question about the pleasantness and the length of the task. All the subjects found the study pleasant and of suitable length. We collected 345 subjects' moves in total. Each move recorder was then annotated by three human raters with respect to the liking and disliking attitude conveyed by the subject through face, voice and gestures. Each communicative act received a label for each communicative channel and a final label for the attitude (like, neutral and dislike) using a majority agreement criterion.

3.2. Signals in facial expressions

The annotation labels for the analysis of facial expressions were selected according to the dimensional approach [22]. Following Russell's approach, we use the valence (negative, neutral and positive) and the arousal (low, medium and high) dimensions. Recognizing the value of only these two dimensions is justified since the valence indicates a failure/success in the achievement of a users goal (i.e. buy a dress) and the arousal is related to the importance of the goal (i.e. how much I liked a dress). We thus annotated expressions related to pleasant and unpleasant feelings using the following labels: Positive Evaluation Face (PEF) corresponding to positive valence, Neutral Face (NF) and Negative Evaluation Face (NEF). At this stage of the project we did not consider the arousal dimension for facial expressions since for us it was important to recognize only the liking and disliking attitudes.

3.3. Signals in the speech prosody

In analyzing signals in the spoken utterance of the user during the conversation with the shopping assistant, we considered the prosody of the uttered sentence. Research in emotional speech has shown that acoustic and prosodic features can be extracted from the speech signal and used to develop models for recognizing emotions and attitudes [3]. In fact, the effects of emotion in speech tend to alter the pitch, timing, voice quality, and articulation of the speech signal and reliable acoustic features can be extracted from speech, which vary with the speaker's affective state. In this case we considered again the valence and arousal that could be detected in the speech. For our purpose we considered the following labels when annotating speech during the interaction: Positive Intonation (PI) denoting sentences with a positive valence, NegativeHigh Intonation (NHI) and NegativeLow Intonation (NLI) denoting respectively a negative valence with a high or low arousal, and Neutral Intonation (NI). This distinction between NegativeLow and NegativeHigh is important since it allows us to distinguish anger (usually caused by an exogenous agent) form sadness (usually related to an endogenous agent).

3.4. Signals in gestures

Our gestural habits can convey several kinds of information regarding our attitude. Given the various possibilities for gestures to annotate, we selected those denoting an attitude of confidence, closure or doubt [8] towards observed items. We analyzed the literature on the topic and we classified gestures according to arms and hands positions. Arms are quite reliable indicators of mood and feeling, especially when interpreted with other signals. They act as defensive barriers when across the body, and conversely indicate feelings of openness and security when in open positions, especially combined with open palms. Hands are also involved in gesture recognition. They are very expressive parts of the body and they are used a lot in signaling consciously - as with emphasizing gestures - or unconsciously - as in a wide range of unintentional movements which indicate otherwise hidden feelings and thoughts. For instance, picking one's nose denotes a social disconnection, inattentiveness or stress, neck scratching denote doubt or disbelief, putting a hand on the mouth or chin or cheek denote evaluation. A subtler version of the evaluative position is with the hand supporting the head. There are also iconic gestures that may indicate liking or disliking, as for instance the thumbs-up/down gesture, that should be taken into account in this context. Even if the position of the legs cannot be considered as a part of gesture, in evaluating the users attitude we take into account whether their legs are crossed or not. We considered the following attitudes determinable from gestures relevant in our study: PositiveEvaluation (PE), NegativeEvaluation (NE), Doubt (D).

3.5. The annotation task

We used the labels we defined to annotate each communicative act in the collected dataset. Given the set of collected dialogues between the shopping assistant and individual users, we split each dialogue into segments. Segments are typically composed of a multimodal dialog move performed by the user and the preceding one of the shopping assistant. In cases where users produce multiple moves without any shopping assistant intervention, segments consist of a shopping assistant's move followed by consecutive user moves. Three independent annotators were involved in this task. Two of them were psychologists expert in communication theory and one of them was an expert in human-computer interaction. Before the annotation task, each of them was introduced and trained to the annotation task and to the meaning of each label. Annotators worked on one dialogue at a time, and within a dialogue they were presented with one segment at a time, following the order in which the moves appear in the dialogue. For each segment they selected one label for each modality and one for the attitude. Labels used to denote the attitude in the multimodal communicative act during the dialog are: Like, Neutral, and Not-Like. In order to test the validity of the annotation and 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 Fleiss 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

Table 1
Inter-annotators agreement.

Signal Label	Kappa	Agreement
Facial Expressions		
PEF (positive)	.87	almost perfect
NF (neutral)	0.16	slight
NEF (negative)	0.56	moderate
Speech		
NI (neutral)	.17	slight
NHI (negative)	0.8	substantial
NLI (negative)	0.8	substantial
PI (positive)	0.5	moderate
Signals in Gestures		
PE (positive)	.6	moderate
NE (negative)	0.8	substantial
D (doubt)	0.4	fair
Global		
Like	0.6	moderate
NotLike	0.8	substantial
Neutral	.17	slight

SA1: Good afternoon, I'm Gilda and I'm here to give you information and suggestions about dresses, shoes, accessories and complete outfits.

U1: Hi, I'm Imma <waving and smiling> (S: PI, G: PEG, F: PEF, Attitude: Like)

SA2: If you need help just tell me.

U2: OK! <thumb-up and smiling> (S: PI, G: PEG, F: PEF, Attitude: Like)

U3: Look at dresses (S: -, G: - , F: NF, Attitude: Neutral)

U4: Look at details of dress D2 <smiling> (S: -, G: -, F: PEF, Attitude: Like)

U5: Put dress in the shopping cart <smiling> (S: -, G: - , F: PEF - Attitude: Like)

U6: Look at shoes (S: -, G: -, F: -, Attitude: Neutral)

U7: Look at details of shoes S5 (S: PI, G: -, F: PEF, Attitude: Like)

U8: Look at price of shoes S5 (S: -, G: -, F: NEF, Attitude: Not-Like)

U9: Damn they are too expensive (S:NHI -, G: -, F: NEF, Attitude:Not-Like)

SA3: Do you think they are too expenses? We can see if there is a similar pair of shoes but cheaper.

U10: Well... they were very nice <meanwhile she touches her nose> (S: NI, F: NEF, G: NEG, Attitude: Not-Like)

U11: Look at details of shoes S7 (S: -, G: DG, F: -, Attitude: D)

U12: Look at details of shoes S9 <she smiles>These are not bad! (S: PI, G: -, F: PEF, Attitude: Like)

U13: Put shoes in the shopping cart (S: -, G: -, F: NF, Attitude: Neutral)

U14: Ok I have finished, I'm happy with this outfit <thumbs-up and smiling> (S: NI, G: PEG, F: PEF, Attitude: Like)

Fig. 2. An example of annotated dialog (translated from Italian) between a subject and a human shopping-assistant. Non-verbal communication is reported as a text.

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 1 summarizes the inter-annotator agreement. The annotators agreed in assigning a label to most of the signals. Only the level of agreement about the 'neutral' communicative acts was slight and lower than we expected.

Fig. 2 illustrates an example of annotated dialog between a subject, whose moves are denoted with U_i, and the human shopping assistant, denoted with SA_j. In the annotation S reports the label about the Speech prosody, G the one for gestures and F the one for facial expressions.

4. Dynamic modeling of the user attitude

We used the dataset we collected to structure the user attitude model. Due to the uncertainty and possible incompleteness of data typical of user modeling tasks, probabilistic approaches are very often used for this task especially when affect is involved [3].

In our framework the user model is represented as a Dynamic Belief Network (DBN) that takes into account as evidence variables the prosody analysis, body pose and the facial expression valence. DBNs are an extension of Bayesian Networks (BNs), to enable modeling temporal environments [17]. A DBN is made up of interconnected time slices of usually structurally identical static BNs; previous time slices are used to predict or estimate node states in the current time slice. The conditional probability distributions of nodes with parents in different time slices are defined by a state evolution model, which describes how the system evolves over time. Temporal DBNs typically follow the form of a first order Markov process, which assumes that future states are independent of the models prior history, given its current state. A significant advantage of DBNs over static networks is their ability to explicitly represent changes in a models belief state over time. By structuring contemporaneous state variables into discrete time slices, and progressively appending new slices to the network, DBNs maintain a growing history of the process being modeled. The method allows us to deal with uncertainty in the relationships among the variables involved in user's liking attitude estimation and to represent situations that gradually evolve from one dialog step to the next.

Fig. 3 shows the DBN that is used to infer how the attitude of the user evolves during the shopping experience according to signals expressed in the verbal and non-verbal part of the communication. The shopping attitude (Attitude) is the variable we want to monitor which depends on observable ones. The probability of each state (notlike, neutral, like) of this variable depends on the recognized signals in the user move deriving from the acoustic, facial and gestural analysis as well as on previous attitudes of the user (Attitude_Prev). Root nodes in the model correspond to a variable that takes the values of the confidence of the classified instance.

DBNs, as employed in this paper, are said to be 'strictly repetitive models'. This means that the structure and parameters of individual time slices are identical and temporal links between two consecutive time slices are always the same. In our simulations, every time slice corresponds to a user move and temporal links are established only between the dynamic variable that we are monitoring in two consecutive time slices. At the beginning of interaction, the model is initialized; at every dialog step (T_i), knowledge about the evidence produced by the multimodal analysis is entered and propagated in the network: the model then revises the probabilities of the Attitude node.

Fig. 4 shows the monitoring of the overall attitude during the sample dialog using the developed framework. In the lower part of the window the general shopping attitude is shown. In particular the probability of interpreting user's communicative behavior as NotLike, Neutral and Like is shown. In the upper part the trend of the attitude is shown for each users move in the dialog.

5. Real-time recognition of signals in DIVA

DIVA is a conversational recommender system based on the previous DBN for modeling the users shopping attitude and embodied in a social robot. The DBN is installed on a workstation and connected to the robot via Wi-Fi. In order to validate the performance of the model in real time applications we developed three modules: one for the users' attitude recognition from face, from voice, and the other from body poses and gestures.

5.1. Facial expression analysis

Our facial expression analysis module is able to analyze and recognize facial expression of users. This module processes the sequence of images captured by the robot's camera and then applies a three-step procedure that consists of: Face Detection, Facial Features Extraction, and Classification. The facial analysis module processes the sequence of images captured by the robot camera

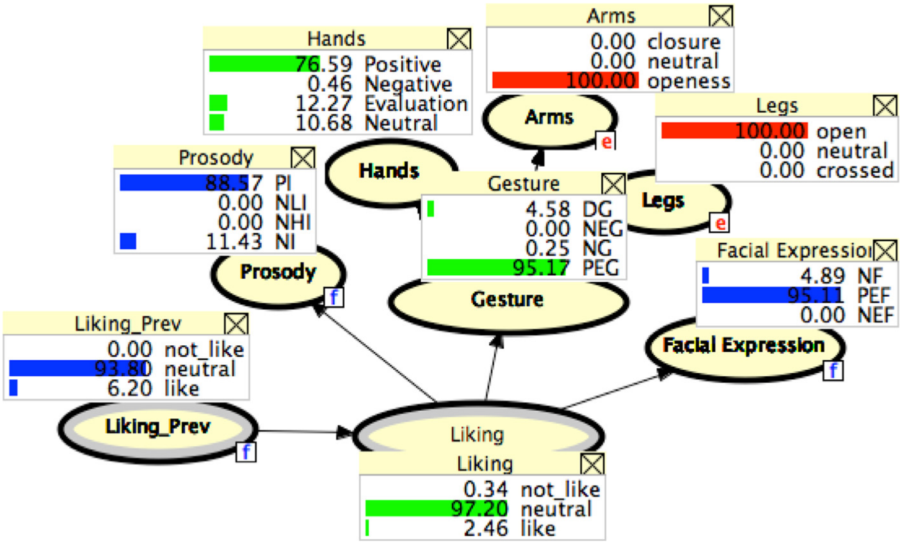


Fig. 3. DBN for modeling the user's shopping attitude.

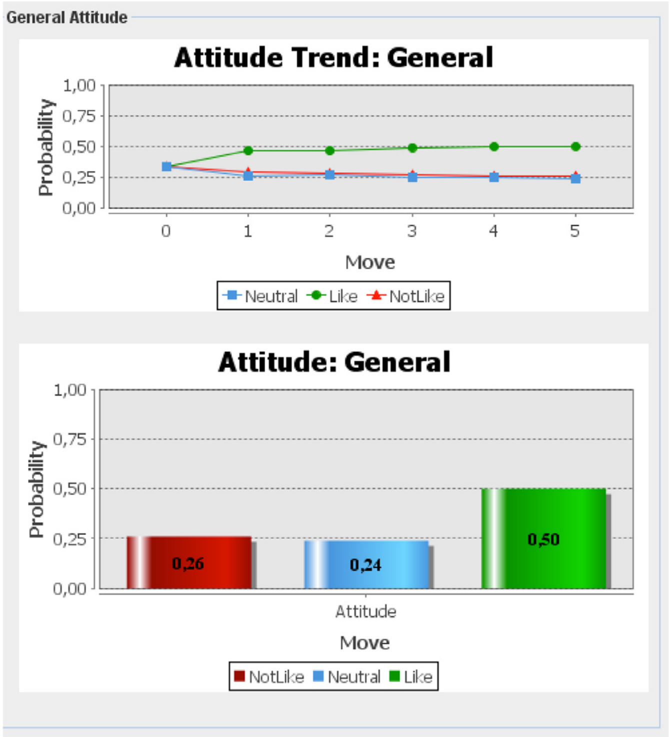


Fig. 4. An example of the framework monitoring window.

and detects the face in the images. Then, the facial analysis module sends the detected face to the modules for facial features extraction and classification. Face Detection is a two-class problem aimed to determining the presence or not of a face in an image. Our system uses the well-known Viola and Jones face detector that minimizes computational costs and achieves high detection performance. The facial expression recognition module is able to identify Ekmans six basic emotions [6]: anger, disgust, fear, happiness, sadness, surprise. Then the detected face is processed for facial features extraction. The Staked Active Shape Model (STASM) approach is used for locating 77 facial key points. Given a set of facial fiducial points, 32 geometrical features (linear, polygonal, elliptical and angular) are calculated. Linear features are defined by the Euclidean distance between two key points. The system uses 15 linear features to describe all parts of a face [20]: i) 6 for the upper face: 3 for the left eyebrow, 2 for the left eye and 1 for the cheeks; ii) 9 for the lower face: 1 for the nose and 8 for the mouth. Polygonal features are determined by the area of irregular polygons constructed on 3 or more key facial points. In this case, 3 polygonal features are defined: i) 1 for the left eye; ii) 1 between the corners of the left eye and the left corner of the mouth; iii) 1 for the mouth. Furthermore, 7 elliptical features are computed using the major and minor axes ratio: i) 1 for the left eyebrow; ii) 3 for the left eye: eye, upper and lower eyelids; iii) 3 for the mouth: mouth, upper and lower lips. Finally, 7 angular features are defined by the angular coefficient of the straight line through two facial key points: i) 1 for the left eyebrow; ii) for the mouth corners. The last step performed by the facial expression recognition system is classification. The facial expression classification module analyzes the geometrical facial features in order to predict the captured facial expression. The facial expression recognition system adopted is fully automatic. It achieves a recognition rate of 95.46% using a Random Forest Classifier and six-class expression classification [20]. In the current prototype of the system, facial expressions are analyzed and we use the valence of the recognized emotion and from this we get the probability of PEF (PositiveEvaluationFace), NEF (NegativeEvaluationFace) and NF (Neutral Face) that are set as evidence in the model.

5.2. Speech prosody

Our approach to classify emotions from speech is based on the dimensional model. Therefore, the analysis of the user's spoken utterance is made by two classifiers: one for recognizing the valence and the other for the arousal dimension. We thus developed a web-service called VOCE 2.0 (VOice Classifier of Emotions ver. 2.0) that classifies the valence and arousal of the voice prosody with an approach very similar to the one described in [29]. The major steps in speech emotion recognition are audio segmentation, feature extraction and the actual classification of the feature vectors into valence and arousal values [7]. VOCE can be used in two ways: off-line for creating and analyzing the emotional speech corpus and, being a web-service, on-line for tracking the affect in voice in real-time. While the off-line version allows us to build the classifier, the on-line emotion recognition just outputs the recognized emotion's valence and arousal and maps the combination of these values into one of the basic emotions by providing the emotion label. Let's see now how the classifier has been trained and how we use it in real-time. A more detailed description of VOCE can be found in [2].

5.2.1. Dataset

Although our approach is based on the dimensional model, since we could not find any corpus, among the few available for Italian, in which emotions were annotated according to their valence and arousal we used the emotion dataset [10]. Among the three available ones the emotion dataset has been used for the EVALITA ERT challenge and therefore we could compare our results with other research works in this domain. The emotional speech characteristics were extracted from the Italian subset of emotion and contains 220 audio files corresponding to sentences for the 6 basic emotions (joy, anger, surprise, sadness, disgust, fear) and the neutral one recorded by professional actors. In order to use the dimensional approach on this dataset we mapped each emotion to the corresponding valence (negative, neutral and positive) and arousal (low, medium, high) using the approach explained in [22]. For instance, 'anger' is mapped into negative valence and high arousal, while 'sadness' is mapped into negative valence and low arousal, 'happy' is mapped into positive valence and medium arousal.

5.2.2. Feature extraction and classification

In developing VOCE we exploited different combinations of features and several classification algorithms. For this task we used Praat.[2] In particular, besides pitch and energy related features, we extracted features related to the spectrum, harmonicity and the Mel-Frequency Cepstral Coefficients (MFCCs) that is used to describe a spectrum frame, its first and second derivative in time are used to reflect dynamic changes. Table 2 shows the features extracted from each audio file using Praat functions. In order to find the best set of features we tested three conditions with several classification algorithms: i) Support Vector Machines (SVM), which offer robust classification to a very large number of variables and small samples. ii) Decision trees, that work with simple classification rules that are easy to understand. The rules represent the information in a tree based on a set of features. iii) Artificial Neural Network (ANN), and in particular the Multilayer Perceptron algorithm. iv) k-Nearest Neighbors (kNN) is one of the simplest of classification algorithms available for supervised learning. It classifies unlabeled examples based on their similarity with examples in the training set.

The three set of features were: i) ALL: all the attributes in Table 2; ii) MFCC: MFCC features only; iii) No MFCC: all the features except MFCC. From the analysis of the performance of the most commonly used algorithms for classification starting from numeric features the most accurate ones were MLP (Multi-Layer Perceptron) and SMO (Sequential Minimal Optimization) algorithms for training a support vector classifier in Weka.[3]

The accuracy was validated using a 10 Fold Cross Validation technique. A k-fold cross-validation with $k = 10$ was used to make validations over the classifiers. This technique allows the model to be evaluated when facing an unknown dataset. The results of the classification of valence, arousal and derived emotion labels are shown in Table 3. The results show that, for both algorithms, using the complete set of features improves accuracy, however using only MFCC related features gives us an accuracy comparable with the one obtained using the set of all features. The worst setting is when MFCC is not considered. As far as the choice of algorithm is concerned, even if MLP had slightly better accuracy, the time to create the model and classify a vocal input is higher (100:1). Since VOCE has to be employed in real-time classification tasks we selected SMO. Since the arousal dimension is related to the importance of the goal and the valence dimension is related to the achievement vs. the threatening of the goal, our speech classifier performs well in recognizing negative states like those related to anger, and allows us to distinguish positive from negative attitudes. However, as expected, some emotions are easier to recognize than others. For example, humans are much better at recognizing anger than happiness; therefore, our results can be considered acceptable under this view. Comparing our results with other works based on the same dataset [19] we can say that our approach has comparable accuracy over the set of basic emotions for the Italian language.

Table 2
The set of features extracted from speech for emotion recognition.

Feature	Description
Pitch	
PitchMin	Minimum value
PitchMed	Average value
PitchMax	Maximum value
PitchMinMaxDiffLog	Logarithmic differentiation
PitchMinLog	Minimum Logarithmic
PitchMedLog	Average logarithmic
PitchMaxLog	Maximum logarithmic
PitchDevSta	Standard Deviation
PitchSlope	Slope
Energy	
EnergyMin	Minimum value
EnergyMed	Average value
EnergyMax	Maximum value
EnergyMinMaxDiff	Logarithmic differentiation
EnergyDevSta	Standard Deviation
Spectrum	
SpectrumCentralMoment	Central moment
SpectrumDevSta	Standard Deviation
SpectrumGravityCentre	Central tendency
SpectrumKurtosis	Degree of centralization
SpectrumSkewness	Degree of asymmetry
Harmonicity	
HarmonicityMin	Minimum value
HarmonicityMed	Average value
HarmonicityMax	Maximum value
HarmonicityDevSta	Standard Deviation
MFCC	
MFCCnMin	Minimum of nth MFCC
MFCCnMed	Average of nth MFCC
MFCCnMax	Maximum of nth MFCC
MFCCnDevSta	Standard Deviation of nth MFCC

Table 3
Accuracy of the two classifiers for valence arousal and derived emotion labels.

Features	MLP	SMO
ALL		
Valence	7045	7136
Arousal	809	7727
Emotion	7136	6863
MFCC		
Valence	6909	6454
Arousal	80	75
Emotion	7045	6818
No_MFCC		
Valence	6409	5545
Arousal	7409	6909
Emotion	5318	5318

[2] http://www.praat.com.

[3] http://www.cs.waikato.ac.nz/ml/weka/.

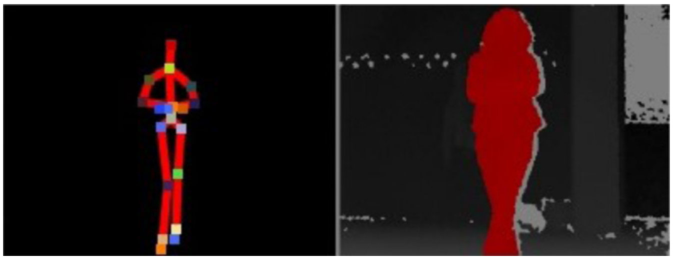


Fig. 5. Kinect DTW recognition of crossed arms and legs.

5.3. Body poses and gestures

Recognizing attitude from body poses and gestures is performed using Microsoft Kinect 1. We consider only those nodes in the skeleton that the Kinect SDK is able to detect, thus considering only a subset of poses and gestures. These are detected with Microsoft Kinect and recognized using KinectDTW (Dynamic Time Warping) that we previously trained for recognizing poses and gestures concerning arms, hands and legs (see Fig. 5).[4] Even if in this context the dynamic recognition of body movements is not necessary and, therefore, body poses could be recognized simply by matching the user skeleton points into predefined patterns. We used DTW in any event since the module was used in another project in which we also recognized dynamic gestures. Recognized hands and arms poses and the legs aperture are then used as evidence in the model and propagated in the net. Fig. 5 illustrates the KinectDTW recognition of crossed arms and legs.

6. Evaluation of the framework

The goal of the main experimental session, described in Section 6.1, was to understand whether our framework was able to recognize the attitude of the user in an accurate way. Then, we performed another preliminary experiment, described in Section 6.2, to evaluate if the framework could be used for implicit rating of items within an emotion-aware recommended system.

6.1. Experimental procedure

The framework was tested using NAO as a Virtual Shopping Assistant and with values for the evidence variables coming from the module for real-time recognition of signals. We performed a Wizard of Oz study [30] in which the human virtual assistant was replaced by NAO displayed on the wall at the side of the shopping system (see Fig. 6). The Nao robot was tele-operated and the experimenter could see the output of the system before control the robot. Participants involved in this study, 48 in total, were selected among the people that asked to participate and were selected so as to have similar features with those involved in the previous experiment phase.

The experiment design was the same as in the previous one, but, this time, the facial expressions were detected using the algorithm described in the previous section, while the subject poses and gestures were detected with Microsoft Kinect. Moreover, the subject was wearing a wireless microphone whose output was sent to VOCE 2.0. Then, after a brief explanation on how to use the system, subjects were asked to give consent to video-record their behavior during the experiment and could start interacting with the system. After the experiment each subject was fully debriefed with an open question about the pleasantness and the length of the task. A corpus of about 720 subjects' moves was collected. Each move was annotated by three human raters with respect to the perceived attitude conveyed by the turn and received a final label for the attitude using a majority agreement criterion. Beside the human annotation, each move was also automatically annotated in terms of attitude using our framework. According to the result of the annotation experiments (both manual and automatic) we conducted a preliminary evaluation for tuning the structure and the probabilities of our multimodal framework.

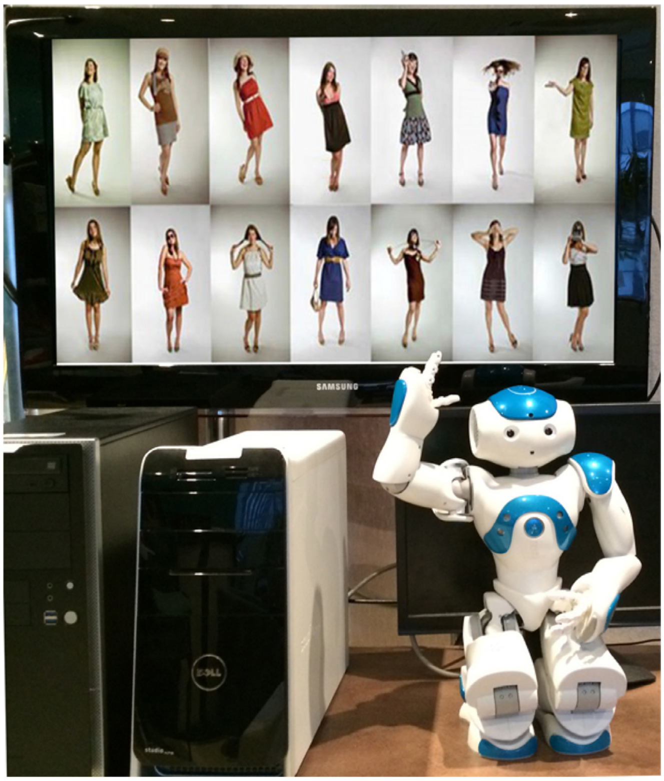


Fig. 6. An example of the prototype used for the experiment.

Results of this experiment show that the model agrees with the human annotation at 82% (via exact matching) in recognizing the overall attitude. Limitations are due to the bad recognition of the facial expression, voice and gestures. The cases in which there was a discordance were mainly due to neutral attitudes or due to cases in which the user was touching her face.

6.2. Evaluating the impact of the framework in the context of recommender systems

Supported by this result, we applied the described framework for computing an implicit rating to items seen by users during the interaction with DIVA. The implicit rating is used by a simple content-based recommender system to suggest dresses (see Fig. 7).

First of all we selected only items of clothing that were described according to the following features: Color, Fit (i.e. tight, comfortable, ...), Type (i.e. daily, elegant, ...), Length (Long, Short). Then, we applied a simple content-based algorithm based on cosine similarity.

The system filters items that have similar features. Since the purpose of the experiment was not the measurement of the performance of the recommender system, but the evaluation of the efficacy of the implicit feedback compared to the explicit rate of items we simplified the recommendation task. To evaluate the approach we compared the prediction in terms of deviation between a predicted rating and the users true rating in two different conditions: baseline and feedback based on attitude recognition.

[4] http://kinectdtw.codeplex.com/.

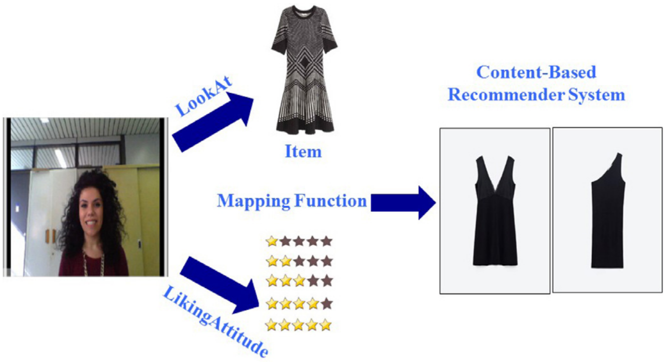
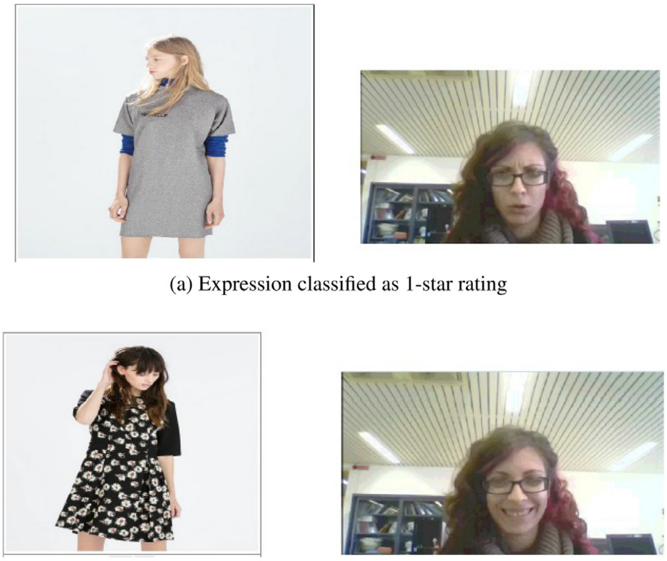


Fig. 7. Schema.

Table 4
Mapping liking attitude probability into a 5-star rating scale.

Liking Attitude	Stars
≥ 85%	5 *****
between 70% and 84%	4 ****
between 50% and 69%	3 ***
between 25% and 49%	2 **
24% ≤	1 *



(a) Expression classified as 1-star rating

(b) Expression classified as 5-star rating

Fig. 8. Example of a classified expression.

We conducted the experiment in a research laboratory in which we installed a display like the one that could be present in a real shop, and, this time, the recommendations provided by NAO were generated according to this item-to-item similarity. In the baseline version the initial profiling technique was made according to the explicit rating of 8 items. In the second condition, the initial user profile was based on the prediction made by our framework. In this case, we had to provide a rating about seen items to be used in DIVA for generating recommendations. For this reason we decided to map the liking-attitude probability inferred by the model to a traditional 5-star rating scale. We then developed a simple mapping function that starting form the liking attitude probability inferred by the model mapped this value to a traditional 5-star rating scale as shown in Table 4. Fig. 8 shows two examples of non-verbal communicative acts mapped into the star rating system.

Table 5
Summary of the experiment results.

	Explicit Feedback	Implicit Attitude Recognition
N	240	240
AVG (5-star rating)	2.59	2.56
Stddev	1.29	0.97

Table 6
The MAE and RMSE calculated for the two conditions.

	MAE	RMSE
Baseline	1.39	1.67
Implicit feedback	11.21	1.6

60 participants were involved in this experiment, drawn from a larger population to have similar features to those involved in the previous experiment phase. They were divided into two groups of 30 subjects. Participants of both groups were asked to perform the same task as in the first experiment. The facilitator of the experiment briefly explained to each subject how to use the system and let her familiarize herself with it using items that were not included in the item-set of the experiment. After this assessment and briefing phase, subjects were asked to give their consent to video-record their behavior during the experiment. Then, they could start interacting with the system. After the experiment each subject was fully debriefed with an open question about the pleasantness and the length of the task. In both cases we showed to each subject 8 initial items, differing in at least one of the previously mentioned features. In the baseline condition NAO asked each subject to explicitly rate each item, in the implicit feedback condition the rating was assigned using our framework (see Fig. 8a and b for two examples of facial expressions recorded during the experiment). In both cases, according to these ratings NAO started to recommend items similar to those evaluated positively by the user (with a rating from 3 to 5 stars). The recommender further refined its suggestions according to the rating provided explicitly in the baseline condition or inferred implicitly by monitoring the liking attitude during the interaction. We collected 240 evaluations for each condition, Table 5 summarizes results.

In order to evaluate and compare the performance of the recommender in the two conditions we calculated several measures: hit-ratio, the Mean Absolute Error (MAE), and the Root Mean Squared Error (RMSE). First of all we calculated the hit-ratio as a way of measuring how many "hits" the user would have in the n-sized list of recommended items. A "hit" can be defined as something that the user has clicked on, purchased, or saved/favorite (depending on the context). In our experiment we evaluated which items the subject brought and whether these were selected among those that were recommended according to the proposed implicit rating approach. If we consider the sum of the hits of every subject in 72% of cases the user bought one of the suggested dresses. Then we calculated the MAE and RMSE, which are the measures that allow us to evaluate the deviation between predicted ratings and actual ratings. Table 6 illustrates the results.

Even if the implicit feedback has slightly worst performance, the difference between the two conditions is small and they seem comparable. This observation is then supported by the paired t-test results ($\alpha = 0.05$) which indicates that there are no significant differences between the two groups of subjects ($p = 0.75$). The framework that used implicit instead of explicit ratings worked properly producing indistinguishable results.

7. Conclusions and future work directions

The proposed work is part of ongoing research aimed at developing DIVA, an emotion-aware virtual shopping assistant embodied in the social robot NAO. In order to endow the agent with the capability of evaluating a user's interest and building a user profile, we developed a framework for monitoring users attitude by analyzing multimodal communicative acts during the interaction. We thus developed a dynamic model for monitoring user liking attitude during the shopping experience. Then we integrated this module in a content-based recommender system and, in order to evaluate its effectiveness in a real shopping scenario, we tested its performance to set up the user profile in the cold-start phase. During the first study and the experiments the participants show interest in new technologies and in new interaction modalities (interactive wall display and robot). In the experiments the robot interacts with the participants by speech synthesis, movements and eye color. The presence of the robot seems to lend the most success to the framework, the participants find the robot amazing and enjoyable. The performance of the recommender was evaluated by comparing it with a baseline condition in which the user rating was provided explicitly. In terms of hit-ratio, our results show that on average most of the users look at suggested items and 72% of items purchased were selected from among the recommended dresses. In terms of MAE and RMSE results show that, even if the implicit feedback had slightly worse performance, the difference between the two conditions is small and the their performance is comparable.

These findings validate our research hypothesis that implicit feedback can be determined from automatic analysis of users multimodal communication and can improve the performance of a recommender system in cases where explicit ratings are not available by allowing the set up of a user profile. It can also be used to integrate and complement the explicit feedback, thus improving the quality of recommendations.

However, this is an ongoing project that needs additional investigation, especially with respect to the factors that introduce noise to the behavior analysis since, for instance, the accuracy of the facial expression recognition software was dependent on the lighting conditions.

For future work, we plan to enrich the model with knowledge related to time spent looking at items as well as which details of the item (price, size, colors, etc.) determined a particular behavior. We intend to employ eye and head tracking using Kinect as for the body poses and gestures. Moreover, we will integrate this model in a framework for emotion-aware conversational recommender systems, which also takes into account the age and gender of the users.

References

[1] N. Ambady, R. Rosenthal, Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis., Psychol. Bull. 111 (2) (1992) 256.

[2] B. De Carolis, S. Ferilli, G. Palestra, D. Redavid, Emotion-recognition from speech-based interaction in AAL environment., in: AI* AAL@ AI* IA, 2016, pp. 92–104.

[3] B. De Carolis, N. Novielli, Recognizing signals of social attitude in interacting with ambient conversational systems, J. Multimodal User Interfaces 8 (1) (2014) 43–60.

[4] N. Doering, S. Poeschl, H.-M. Gross, A. Bley, C. Martin, H.-J. Boehme, User-centered design and evaluation of a mobile shopping robot, Int. J. Social Rob. 7 (2) (2015) 203–225.

[5] R.J. Donovan, J.R. Rossiter, G. Marcoolyn, A. Nesdale, Store atmosphere and purchasing behavior, J. Retailing 70 (3) (1994) 283–294.

[6] P. Ekman, W.V. Friesen, The repertoire of nonverbal behavior: categories, origins, usage, and coding, Semiotica 1 (1) (1969) 49–98.

[7] F. Eyben, B. Schuller, Opensmile:): the munich open-source large-scale multimedia feature extractor, SIGMultimedia Rec. 6 (4) (2015) 4–13, doi:10.1145/2729095.2729097.

[8] R.S. Feldman, Applications of nonverbal behavioral theories and research, Psychology Press, 2014.

[9] B. Ferwerda, M. Schedl, Enhancing music recommender systems with personality information and emotional states: a proposal., UMAP Workshops, 2014.

[10] V. Galata, Production and perception of vocal emotions: a cross-linguistic and cross-cultural study, PhD Thesis, University of Calabria, Italy, (unpublished), 2010 Ph.D. thesis.

[11] M. de Gemmis, P. Lops, C. Musto, F. Narducci, G. Semeraro, Semantics-aware content-based recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), Recommender Systems Handbook, Springer, 2015, pp. 119–159, doi:10.1007/978-1-4899-7637-6_4.

[12] Y. Hu, X. Yi, L.S. Davis, Collaborative fashion recommendation: a functional tensor factorization approach, in: Proceedings of the 23rd ACM International Conference on Multimedia, in: MM '15, ACM, New York, NY, USA, 2015, pp. 129–138, doi:10.1145/2733373.2806239.

[13] K. Kamei, K. Shinozawa, T. Ikeda, A. Utsumi, T. Miyashita, N. Hagita, Recommendation from robots in a real-world retail shop, in: International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction, ACM, 2010, p. 19.

[14] P. Lops, M. De Gemmis, G. Semeraro, Content-based recommender systems: state of the art and trends, in: Recommender Systems Handbook, Springer, 2011, pp. 73–105.

[15] K.A. Machleit, S.P. Mantel, Emotional response and shopping satisfaction: moderating effects of shopper attributions, J. Bus. Res. 54 (2) (2001) 97–106.

[16] C. Mumme, N. Pinkwart, F. Loll, Design and implementation of a virtual salesclerk, in: International Workshop on Intelligent Virtual Agents, Springer, 2009, pp. 379–385.

[17] A.E. Nicholson, J.M. Brady, Dynamic belief networks for discrete monitoring, IEEE Trans. Syst. Man Cybern. 24 (11) (1994) 1593–1610.

[18] T.D. Nielsen, F.V. Jensen, Bayesian networks and decision graphs, Springer Science & Business Media, 2009.

[19] A. Origlia, V. Galatà, B. Ludusan, Automatic classification of emotions via global and local prosodic features on a multilingual emotional database, in: Proc. of Speech Prosody, Citeseer, 2010.

[20] G. Palestra, A. Pettinicchio, M. Del Coco, P. Carcagnì, M. Leo, C. Distante, Improved performance in facial expression recognition using 32 geometric features, in: International Conference on Image Analysis and Processing, Springer, 2015, pp. 518–528.

[21] I. Poggi, Mind markers, The Semantics and Pragmatics of Everyday Gestures, Berlin Verlag Arno Spitz, 2001.

[22] J. Posner, J.A. Russell, B.S. Peterson, The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology, Dev. Psychopathology 17 (03) (2005) 715–734.

[23] A. Powers, S. Kiesler, S. Fussell, C. Torrey, Comparing a computer agent with a humanoid robot, in: Human-Robot Interaction (HRI), 2007 2nd ACM/IEEE International Conference on, IEEE, 2007, pp. 145–152.

[24] L. Qiu, I. Benbasat, Evaluating anthropomorphic product recommendation agents: a social relationship perspective to designing information systems, J. Manage. Inf. Syst. 25 (4) (2009) 145–182.

[25] S. Satake, K. Hayashi, K. Nakatani, T. Kanda, Field trial of an information-providing robot in a shopping mall, in: Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on, IEEE, 2015, pp. 1832–1839.

[26] K. Shinozawa, F. Naya, J. Yamato, K. Kogure, Differences in effect of robot and screen agent recommendations on human decision-making, Int. J. Hum. Comput. Stud. 62 (2) (2005) 267–279.

[27] M. Tkalčič, U. Burnik, A. Odić, A. Košir, J. Tasič, Emotion-aware recommender systems–a framework and a case study, in: ICT Innovations 2012, Springer, 2013, pp. 141–150.

[28] M. Tkalcic, B.D. Carolis, M.D. Gemmis, A. Odic, A. Kosir, Emotions and personality in personalized services: models, evaluation and applications, in: Human-Computer Interaction.

[29] T. Vogt, E. André, N. Bee, Emovoicea framework for online recognition of emotions from voice, in: International Tutorial and Research Workshop on Perception and Interactive Technologies for Speech-Based Systems, Springer, 2008, pp. 188–199.

[30] S. Whittaker, M.A. Walker, J.D. Moore, Fish or fowl: a wizard of oz evaluation of dialogue strategies in the restaurant domain., LREC, 2002.