Assessing user bias in affect detection within context-based Spoken Dialog Systems

Syaheerah Lebai Lutfi*†, Fernando Fernández-Martínez*, Andrés Casanova-García*,
Lorena López-Lebón* and Juan Manuel Montero*
*Speech Technology Group
Universidad Politécnica de Madrid, Madrid, Spain
Email: [syaheerah, ffm, lorena, acasanova, juancho]@die.upm.es
†School of Computer Sciences, University Science of Malaysia, Penang, Malaysia

***Abstract*—This paper presents an empirical evidence of user bias within a laboratory-oriented evaluation of a Spoken Dialog System. Specifically, we addressed user bias in their satisfaction judgements. We question the reliability of this data for modeling user emotion, focusing on *contentment* and *frustration* in a spoken dialog system. This bias is detected through machine learning experiments that were conducted on two datasets, users and annotators, which were then compared in order to assess the reliability of these datasets. The target used was the satisfaction rating and the predictors were conversational/dialog features. Our results indicated that standard classifiers were significantly more successful in discriminating frustration and contentment and the intensities of these emotions (reflected by user satisfaction ratings) from annotator data than from user data. Indirectly, the results showed that conversational features are reliable predictors of the two abovementioned emotions.**

I. Introduction

As machines and people begin to weave the fabric of society together, spoken conversational agents (SCA) are increasingly being developed to expedite tasks that were previously carried out using other modalities that were less seamless or required explicit communication [29], especially within a domestic environment.

Automatic affect detection of users during real-time conversation is the key challenge towards our greater aim of infusing affect into a natural-language mixed-initiative HiFi-control spoken dialog agent (henceforth 'HiFi agent'). The voice-only HiFi agent was previously developed by GTH (details in [15]).

It is strongly acknowledged that the artificiallity of a laboratory environment poses a huge challenge when collecting unmasked emotional data [27]. Real-world usage is very difficult to simulate in a laboratory setting due to lack of contextual information (e.g., users are given certain objectives or missions to fulfill when interacting with an SCA, representation of actual physical environment etc.). Evaluators tend to adapt to the less natural setting, adjust their tolerance levels and mask their feelings or opinions of the system that is being evaluated. Thus data is usually collected using a sample of users that are less representative. Evaluating an SCA by providing the users with their own tasks and goals that are fitted to the evaluation scopes and objectives also, in actual use context, could reveal problems that were previously not detected during laboratory evaluation. Though user biases in laboratory evaluations are known phenomena, they have not been empirically tested, at least not in the area of Affective Computing. This raises the question of whether the laboratory-collected emotion-inherent data used to train a system that is intended for use in a natural environment is reliable. In this paper, we attempt to address this question by providing empirical evidences of user bias.

Specifically, our main contribution is to derive empirically based conclusions on the relationship between user *satisfaction ratings* derived in a laboratory-led evaluation and their inferences of users' *frustration* and *contentment*. Thus we address the questions in relation to:

1) the *reliability* of satisfaction judgment data when modeling these two emotions in a spoken dialog system, and,
2) using dialog as source of cues for affect detection (at last for now, these two emotions), and hence the correlation between dialog features and satisfaction rating.

II. Context of application

Within spoken dialog systems, while there are considerable amount of studies that address agent believability and also user affective states that accompany other environments, especially learning [e.g., 17, 18, 31, 33], games/entertainment [e.g., 29, 37] and call centers/ information-services [e.g., 22, 24], very few aim at identifying emotions that influence interactions within a *domestic* environment. Studies in a closer domain such as those of Human-Robot Interaction for intelligent homes are typically concerned with the design space of service-robots towards improving their believability through life-like essences such as appearance (e.g., anthropomorphism) and some other physical acts (e.g., headpose, gaze, motor skills), accounting for intimacy and engagement with the robot in order to increase people's acceptance of the former as companions [8, 34]. However, the idea of demonstrating social intelligence of a domestic agent in such a way that it would be regarded as a companion, can be quite far-fetched, considering the scope and the technical facet of the application (that may be a speech-only application) and the relevant affective states targeted for adaptation [17] that may be rather limited. Thus in striving to be natural and adaptive towards the user, these systems are not expected to be a human clone - as in to possess human social qualities to the utmost degree [12], but suffice when we get it to "evoke humanness in us", as Cassell [6] puts it. Thus, a plausible and feasible goal would be to have a dialog system that is expressive enough [12]

that the human interlocutor respond to it by applying native speaker intuition. Though some users tend to treat machines similar to humans [30], they may not mind some 'hiccups' in the interaction as long as there is no major breakdown in communication, as asserted by Edlund et al. [12].

III. Affect detection using satisfaction ratings (target) and conversational features (predictors)

Real time automatic detection of emotion is vital for any affect-sensitive system. User *satisfaction* judgment could indicate contentment or frustration [2, 10, 23] and the relationship of similar emotions and satisfaction judgment have been empirically proven in [19, 21] and also in our work, which will be further described. Although user satisfaction has been used as a classic measure of user opinions of computer systems, including SCAs, studies concerning affective SCAs do not treat the user's *opinion* as a reflection of his or her *affect*. A different approach is usually adopted to investigate user emotions while interacting with a SCA, commonly involving manual labeling task; independent judges listening to the users' utterances and then labeling them with several emotion categories on a turn-to-turn basis. Human listeners do not usually achieve high agreements on these emotion classifications [3, 5], even when using trained judges [9]. Cowie et al. [7] pointed out that challenges in using emotion labels are not only limited to ensuring that the labels are correct, but also that the raters *agree* on those labels. It has also been reported that perceived and actual states can be rather divergent [35].

To model satisfaction we used satisfaction rating as the target and *conversational features* as predictors, obtained from a corpus collected in a past evaluation [14]. The users involved in the evaluation did not have previous experience in interacting with the HiFi agent, and their participation were not rewarded. What makes our approach different from others is that we used target and predictor variables whose potentials are often ignored to model affect. While many studies focus on numerous channels for affect detection, very few have explored dialog as a potential source [9]. User affect could be mined from dialog or conversational elements, which are always cheaper and are usually obtained with little or no computational overhead. By looking for emotional information beyond the mainstream visual (facial, gesture, posture) and vocal elements (acoustical or prosodical), such as those extracted from conversational elements, one could combine these two elements into a single decision framework to infer a more meaningful social phenomenon. Often many socially related traits, such as age, culture and personality are detectable from the way a speaker interacts, and are not directly picked up from the words that are spoken [20].

IV. Affective states accompanying interactions with domestic spoken dialog agents

Based on the observations of the interactions in the videos from past evaluations of the spoken dialog HiFi agent, we were able to identify a set of emotions that frequently occurred during user-HiFi agent interaction. Typical emotions involved were contentment, frustration, confusion and boredom. These emotions are within the same family of some of the basic emotions proposed by Ekman and Friesen [13] namely happiness, anger, surprise and sadness respectively, but in finer and less intense nuances. One other emotion of interest was self-frustration, in which users displayed discontentment towards themselves for erroneously addressing the system. We also added neutral to represent situations where there was no particular emotion of the aforementioned type present. This paper would however focus on discriminating affect between two classes: *contentment* and *frustration*, two types of emotions that are known to be prevalent within spoken HCI. These two categories of affect represent positive and negative user emotional state and their varying intensities (e.g., at the end of an interaction, a particular user might have felt intensely content with the system when the user gave a score of 5 or 'excellent' (on a 5-point scale), and rather frustrated when he or she gave a score of 3. This depends on the model that was chosen for modeling the two emotions- different models have different groupings of scores, elaborated later in Section V-B1. A score of 3, for example, may either represent a low-intensity frustration (category Three version 2) or slight contentment (category Three version 1)

V. Automatic detection of affect

A. User and annotator studies

To model affect by predicting user satisfaction, we used the audio-visual HiFi-AV2 corpus [see 16], collected during a *user* study which consists of audiovisually recorded information of real interactions between user and non-adaptive version of HiFi agent (thus the emotions conveyed during these interactions were *non-acted*). In this study, each user interacted with the HiFi agent hands on in 10 sessions (N=190 interaction sessions) which were guided by pre-defined basic, advanced and free scenarios. In basic set of scenarios, the users were strictly guided and only had to address a single task - e.g."You should try to stop the CD from playing". In the advanced set users were less guided, and given a more complex combinations of tasks - e.g."You should attempt to play a track from the CD at a higher volume", and in the free set users were not constrained, given no restriction but were told that the tasks should focus on the three main devices contained within the HiFi system - the CD player, tape player or radio channel. At the end of each interaction, they rated the HiFi agent by providing a score between 1-5 Likert point (1 being very poor to 5, excellent). It should be noted that this study was conducted with the intention of only measuring the agent's performance, without forseeing the integration of any social intelligence (e.g.,emotions).

Later, we used a reduced version of the same corpus to obtain satisfaction and affect-labelled data from several independent *annotators* (this paper focuses only on the satisfaction labelled data, however). The corpus was reduced by randomly selecting interaction samples from 10 users (N=100 sesions). In this study, the annotators were asked to rate *user emotion*.

They also had to rate the agent by giving a satisfaction rating, similar to the users - the annotators were given a set of full recordings (from the start until the end of an interaction) and they were free to label as many defined emotions (as stated in Section IV) detected throughout the whole interaction. It is important to note that the annotators were asked to rate the agent based on the perspective of the *user*, and were naïve on real-users ratings - in other words the annotators put themselves in the users' shoes and rated the system as how the users should have rated the system. Thus we could view both datasets as that of users' actual ratings and targeted ratings (by annotator). Ultimately, three satisfaction-labelled datasets were obtained: A full set of 190 interaction samples (UserFULL), the selected 10 users of 100 samples (UserSEL) and the same selected samples labelled by annotators (AnnotSEL).

B. Experiments

In order to obtain a model of user affect, we conducted several experiments on data using only *conversational features* (see Table I) and conversational features *plus module-related features* (see Table II). In both types of experiments, we applied standard classification techniques in which several classifier schemes were utilized with the intention of comparing the performance of the various classification techniques, apart from determining which technique(s) yield the best performance. The Waikato Environment and Knowledge Analysis (WEKA) [36] was used for these purposes. One or more classification algorithms were chosen from different categories including rule-based classifiers (ZeroR as the baseline - at 50% chance, and OneR), functions (SimpleLogistic, SMO), meta classification schemes (Multischeme, MultiBoost, AdaBoost) and trees (J48). A 10-fold cross validation technique was used for the classification task.

1) Clustering: All satisfaction-labelled datasets were first resampled in order to obtain a better distribution; samples with similar outcomes were grouped together, and this was repeated five times to satisfy all combinations of classification problems as shown in Table III. This way we were also able to determine which clusters obtained optimized classifications.

VI. Results and Discussions

A. Classification evaluated on UserFULL dataset

The results from the experiment with the UserFULL dataset revealed no statistically significant result - at best, only 5 percent improvement from the baseline to OneR, revealing that the satisfaction score could not be predicted from the dialog features. This begs the question of whether the users were rating the system randomly or were just being positively biased. Upon closer inspection of the data, we found that there were too few cases for point 1 (very poor) and point 2 (poor) categories, and majority cases turn out to have 4 (good) or 5 (excellent) points. This ceiling effect in reporting the satisfaction score suggested that users might have been acquiescent when assessing the HiFi agent. In the light of this discovery, we studied the correlation between the satisfaction score and the *actual recognition accuracy*, to confirm that

TABLE I
Conversational features

Features	Description
Turns Taken	Number of turns needed to complete a scenario.
Contextual Turns	Number of turns taken where contextual information handling strategies are applied successfully.
System Requests	Number of turns taken where the system requests missing information from the user.
Executed Action	Number of turns required to accomplish a particular goal (execute a specific action).
Help Request	User interrupts the interaction to request for some help.
Cancellation Request	User promptly quits current interaction and starts a new one.
Silence Timeouts	Timeout occurs after silent phase of a given duration.
Recognition Timeout	Timeout occurs when recognition timer expires. E.g.: When user speaks lengthy sentence, and violates the time limit.
System Failures	Occurs when the system failed to receive IR commands
Repeat Speech Recognition	User repeats an utterance and system captures newly recognized words in the repeated utterance.
Repeat Speech Understanding	User repeats an utterance that has the same semantic content.
Speech Recognition Rejection	Occurs when words in an utterance obtain lower confidence score than certain threshold.
Non-Language Understanding Rejection	Occurs when the concepts in an utterance obtain a lower confidence score than a certain threshold, albeit good overall recognition score.
Out-of-domain words	occurs when words uttered are meaningless in view of dialog goal (i.e., the system is not able to determine any word that influences the execution of an action).
Dialog Time	Time required (in seconds) to complete a dialog.

TABLE II
Module-based features

Features	Description
Recognition	Num. of words per scenario, average num. of words per sentence, % good words, num. of sentence per scenario, % good sentence, average confidence of sentence per scenario, num. of good scenario (based on recognition).
Understanding	Num. of concepts per scenario, average concepts per sentence, % good concepts, average confidence of concepts per scenario, num. of good scenarios (based on concepts).
Dialog Manager	Complexity of interaction per scenario (complexity = num.of goals/num. of executed goals), % of incomplete goals, % completed goals.
Dialog Act	Greeting, Request, Imperative, Offense, Pardon, Grateful, Farewell, Correction, Consultation, Confirmation.

TABLE III
DATASETS RE-CLUSTERED ACCORDING TO SIMILARITY OF SCORE POINTS INTO ALL POSSIBLE COMBINATIONS OF CLASSES

Category	Label				
	very poor	poor	satisfactory	good	excellent
5-Five (original class)	1	2	3	4	5
4-Four	-	1,2	3	4	5
3V1-Three (version 1)	-	1,2	3	4,5	-
3V2-Three (version 2)	-	1,2,3	-	4	5
2V1-Two (version 1)	-	1,2,3	-	4,5	-
2V2-Two (version 2)	-	1,2	-	3,4,5	-

the scores were biased. Weak correlation between the users' satisfaction score and the actual recognition accuracy (r=.15) explained that users rated the system more favourably and were less critical towards the agent. In converse, the *annotators* depended on this criterion significantly (r=.36, p<.01) to do the same.

B. Classification evaluated on UserSEL and AnnotSEL datasets

TABLE IV
COMPARISONS OF SIGNIFICANT IMPROVEMENTS IN CLASSIFICATION ACCURACIES IN DETECTING SATISFACTION SCORE FROM CONVERSATIONAL FEATURES

Category	Classifiers							
	Base rate		SiLog		SMO		Ord	
	U	A	U	A	U	A	U	A
5	38.0	36.0	-	49.3	-	44.6	-	51.3
4	38.0	36.0	-	53.1	-	43.4	-	52.0
3V1	71.0	47.0	-	64.0	-	61.1	-	62.5
3V2	38.0	53.0	-	-	50.7	-	-	-
2V1	71.0	**53.0**	-	**75.0**	-	**74.4**		69.4
2V2	93.0	83.0	-	-	-	-	-	-

SiLog= Functions.SimpleLogistics, SMO= Functions.SMO, Ord= Meta.Ordinal.
U= UserSEL, A= AnnoSEL.
Results were truncated to display only the *best* statistically significant classification improvements (at p<.05)

Table IV shows classification improvements (in % accuracy) over baserate using only *conversational features* for datasets labelled by both users (UserSEL) and annotators (AnnotSEL). As indicated in the table, at least three classifiers that were evaluated on the *annotator* data (AnnotSEL) showed significant improvement over the baserate in each category, with the exception to categories 2 and 3 (both V2). On the other hand, the classifiers evaluated on the user data (UserSEL) mostly revealed worse results than baserate with exception of SMO, which improved significantly over baserate for category 3V2. This indicates that most classifiers were able to predict the satisfaction judgment from dialog features based on annotator data, suggesting that the annotators were more impartial when judging the HiFi agent. An Analysis of Variance (ANOVA) was performed in order to evaluate the performance of the classifiers across categories. The ANOVA results indicated that there was a significant main effect of the categories (various groupings) on the improvement of classification accuracy over the baserate, $F(5,40)$=7.52, p<.001, partial η^2=.48.

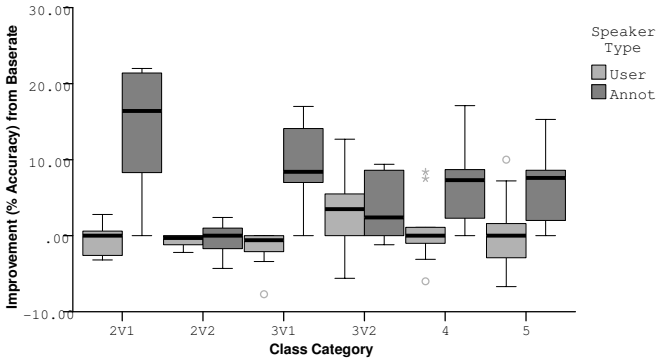

Fig. 1. Improvement accuracy in percentage by categories

Bonferonni *post hoc* pointed that the classifiers performed best when discriminating two classes (2V1), in which points 1, 2 and 3 are collectively tagged as *poor* and point 4 and 5 as *good*, and was significantly higher than only category 2V2 (M_{cat2V1}=6.58, SD=9.7) - see Figure VI-B. However, when point 3 was tagged as *good* in the other version of the two-class problem (2V2), the result was contrary - the classifers' performances were significantly worse than the rest of the categories (M_{cat2V2}=-.69, SD=1.63), suggesting that point 3 is a better representation of *poor* rather than *good*. In other words, when participants gave a satisfaction score of point 3, they probably were indifferent with the system, rather than mildly contented. Category 4 showed the next best improvement rate (M_{cat4}= 3.72, SD=6.27).

Next, Table V shows classification improvements (in % accuracy) over baserate using conversational features plus *module-based features* that were listed in Table II. The table is limited into showing the best classification improvements for category 2V1, since this category yielded the best statistically significant improvement for the first experiment presented in Table IV.

TABLE V
SIGNIFICANT IMPROVEMENTS IN CLASSIFICATION ACCURACIES IN DETECTING SATISFACTION SCORE FROM CONVERSATIONAL+ MODULE-BASED FEATURES

Conv. + [..]	Classifiers							
	Base rate		SiLog		SMO		Ord	
	U	A	U	A	U	A	U	A
Recog	71.0	53.0	-	74.4	-	75.3	-	71.0
Und	71.0	**53.0**	-	**77.1**	-	**78.7**	-	73.4
DM	71.0	53.0	-	73.6	-	75.5	-	65.5
DA	71.0	53.0	-	65.1	-	-	-	64.2
Recog+Und+DM	71.0	53.0	-	74.5	78.6	73.5	-	68.3
All	71.0	53.0	-	73.6	-	74.4	-	80.7

Conv.=Conversational features, Recog= Recognition, Und= Understanding, DM= Dialog Manager, DA=Dialog Act.
U= UserSEL, A= AnnoSEL.
Results were truncated to display only the *best* statistically significant classification improvements (at p<.05)

The results in Table V shows that the inclusion of the features extracted from the understanding module has slightly improved the recognition rate above and beyond conversational

features, however not significant(2.1% for Simple Logistics and 4.3% for SMO).

What is more interesting is that the results above confirmed that the users have been undoubtedly biased or acquiescent. Acquiescence bias holds that respondents to a questionnaire have a tendency to show agreeable behaviour or positive connotations [28] out of politeness [30, 32] - due to the belief that the researcher has a positive judgment of his or her own product and differing with this judgment would be impolite to the researcher, or simply because it takes less effort to just favor the system regardless of its performance than carefully weighing each optional level of good and bad scores. It is noted that user bias is quite common especially in laboratory settings compared to the field environment users [11] who do not have any 'moral' or imposed obligations to give positive judgments.

Criticism of laboratory-led SCA evaluation also concerns the use of predefined scenarios, in which users were denied the freedom of selecting the tasks on their own as they would have done in a non-constricted environment [4] and that they stress on task-completion [1]. These reasons might have caused them to ignore certain aspects of the interaction, such as ease of interaction (or 'comfort factor', termed by [26]) and report a biased satisfaction rating. In our case this could be true - the fact that users were actually requested to address a certain number of goals in a predefined scenario (a 'mission-based' situation) might have caused them to ignore the ease of interaction. When an individual is imposed by certain criteria (e.g: "You should put on the HiFi system") he or she tends to focus only on meeting the criteria for ultimate success, regardless of the consequences. Thus users might only be concerned about *whether* they have achieved a particular goal, but not with *how* it is being achieved. As long as their goals were met, users were satisfied, leading them to rate the agent's overall performance highly. In contrast, annotators were not provided with any predefined scenarios, and therefore gave more impartial ratings.

VII. Conclusions

Our main contribution in this paper is to detect user bias empirically within a laboratory-led evaluation. Whilst we demonstrated that in general, conversational features could predict frustration and contentment (and the intensities of these emotions) from satisfaction ratings, predicting them using data obtained directly from *users* were not possible. We found that users were inclined to inflate the agent's performance by evaluating the system favourably regardless of its actual performance, and thus 'masked' their satisfactions. It is a known fact that it is almost impossible to totally simulate a real world environment in a laboratory, and therefore laboratory data on emotions often cannot be generalized throughout the population. While we are not claiming external validity, we argue that the data could be reused in order to produce a valid finding. We did this by asking annotators to rate satisfactions as imaginary users. Classifications were evaluated on both these *actual* and *target* datasets. The results revealed that satisfaction from the latter were significantly predictable, but not from the former, suggesting that when not constrained in a laboratory setting, users (in this case, *annotators*) were more impartial. Thus, by comparing users' and annotators' datasets, we were able to detect positive bias. In future evaluations (using the same types of scenarios), we would use the annotators' data as a baseline for detecting user bias.

Our second contribution is to show empirically that conversational features, a non-conventional source, could be used as a single source to model user affect reliably by predicting satisfaction ratings in HCI within a limited-task domestic domain. The conversational features were used as affect predictors and the satisfaction judgments were the target. For this task we used an annotation method that is less sophisticated (such as the use of untrained judges to rate satisfaction instead of rating emotions) and smaller array of features for classification tasks. Nevertheless, emotion classification improvements achieved statistically significant results over baserate.

VIII. Current and future directions

We have implemented the emotion classifier into the emotion model, and the latter is incorporated into the HiFi agent in order to make it more social and affective during real-time interaction. We have also developed a suitable response generation model according to the various intensities of the predicted user frustration and contentment. Future work involves evaluating the HiFi agent. A series of cross evaluations will be conducted between users and adaptable/non-adaptable versions of the agent to compare the findings. The evaluation also includes analyzing the impact of the abovementioned generation model on user experience, other than user affect, by modifying those responses based on different agent personalities. For example, a novice user may prefer a dominant agent that is more verbose, explicit and directive, whilst a user that is more familiar with the system may favour a submissive system that is more apologetic and user-led, as suggested in [25, 30]. Thus, the agent may need to respond differently to a *frustrated novice* user than to a *frustrated expert* one.

Acknowledgement

The work leading to these results has been supported by INAPRA (MICINN, DPI2010-21247-C02-02), TIMPANO (TIN2011-28169-C05-03) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Authors also thank all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics. The first author thanks University Science of Malaysia and the Malaysian Ministry of Higher Education for the PhD funding. Authors also thank all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics.

References

[1] H. Ai, A. Raux, D. Bohus, M. Eskenazi, and D. Litman. Comparing spoken dialog corpora collected with recruited subjects versus real users. In *8th SIGdial Workshop on Discourse and Dialogue*, 2007.

[2] J. E. Bailey and S. W. Pearson. Development of a tool for measuring and analyzing computer user satisfaction. *Management Science*, 24:530–545, 1983.

[3] Z. Callejas and R. López-Cózar. Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Commun.*, 50(5):416 – 433, 2008.

[4] Z. Callejas and R. López-Cózar. Relations between de-facto criteria in the evaluation of a spoken dialogue system. *Speech Commun.*, 50:646–665, 2008.

[5] Z. Callejas and R. López-Cózar. On the use of kappa coefficients to measure the reliability of the annotation of non-acted emotions. In *Proceedings of the 4th IEEE tutorial and research workshop on Perception and Interactive Technologies for Speech-Based Systems: Perception in Multimodal Dialogue Systems*, PIT '08, pages 221–232, Berlin, Heidelberg, 2008. Springer-Verlag.

[6] J. Cassell. *Sistine gap: Essays in the history and philosophy of artificial life*, chapter Body Language: Lessons from the Near-Human. University of Chicago Press, 2007.

[7] R. Cowie, E. Douglas-Cowie, J.-C. Martin, and L. Devillers. *A blueprint for an affectively competent agent: Cross-fertilization between Emotion Psychology, Affective Neuroscience and Affective Computing*, chapter The esential role of human databases for learning in and validation of affectively competent agents, pages 151–165. Oxford University Press, 2010.

[8] K. Dautenhahn. Socially intelligent agents in human primate culture. In S. Payr and R. Trappl, editors, *Agent Culture: Human-Agent Interaction in a Multicultural World*, pages 45–71. Lawrence Erlbaum Associates, Mahwah, New Jersey, 2004.

[9] S. K. D'Mello, S. D. Craig, A. Witherspoon, B. McDaniel, and A. Graesser. Automatic detection of learner's affect from conversational cues. *User Model User-Adap. Inter*, 18:45–80, 2008.

[10] W. J. Doll and G. Torkzadeh. The measurement of end-user computing satisfaction. *MIS Quarterly*, 12:259–274, 1988.

[11] L. Dybkjær, N. O. Bernsen, and W. Minker. Evaluation and usability of multimodal spoken language dialogue systems. *Speech Commun.*, 43:33–54, 2004.

[12] J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson. Towards human-like spoken dialogue systems. *Speech Commun.*, 50(8-9):630–645, 2008.

[13] P. Ekman and W. Friesen. *The Facial Action Coding System: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

[14] F. Fernández-Martínez, J. Bläzquez, J. Ferreiros, R. Barra-Chicote, J. Macias-Guarasa, and J. M. Lucas-Cuesta. Evaluation of a spoken dialog system for controlling a hifi audio system. In *Proceedings of the IEEE Workshop on Spoken Language Technology*, Goa, India, 2008.

[15] F. Fernández-Martínez, J. Ferreiros, J. M. Lucas-Cuesta, J. D. Echeverry, R. San-Segundo, and R. Córdoba. Flexible, robust and dynamic dialogue modeling with a speech dialogue interface for controlling a hi-fi audio system. In *Proceedings of the IEEE Workshop on Database and Expert Systems Applications (DEXA 2010)*, Bilbao, Spain, September 2010. Springer.

[16] F. Fernández-Martínez, J. M. Lucas-Cuesta, R. B. Chicote, J. Ferreiros, and J. Macías-Guarasa. HIFI-AV: An audio-visual corpus for spoken language human-machine dialogue research in Spanish. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA).

[17] K. Forbes-Riley and D. Litman. Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system. *Comput. Speech Lang.*, 25(1):105 – 126, 2011. Affective Speech in Real-Life Interactions.

[18] K. Forbes-Riley and D. Litman. Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor. *Speech Commun.*, 53(9-10):1115–1136, 2011.

[19] K. Gelbrich. Beyond just being dissatisfied: How angry and helpless customers react to failures when using self-service technologies. *Schmalenbach Business Review*, 61:40–59, 2009.

[20] J. Grothendieck, A. Gorin, and N. Borges. Social correlates of turn-taking behavior. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, ICASSP '09, pages 4745–4748, Washington, DC, USA, 2009. IEEE Computer Society.

[21] S. Kernbach and N. S. Schutte. The impact of service provider emotional intelligence on customer satisfaction. *Journal of Services Marketing*, 19/7:438–444, 2005.

[22] P. Laukka, D. Neiberg, M. Forsell, I. Karlsson, and K. Elenius. Expression of affect in spontaneous speech: Acoustic correlates and automatic detection of irritation and resignation. *Comput. Speech Lang.*, 25(1):84 – 104, 2011.

[23] E. A. Locke. *The nature and causes of job satisfaction*. Consulting Psychologists Press, Palo Alto, C.A, 1976.

[24] R. López-Cózar, J. Silovsky, and D. Griol. New technique for recognition of user emotional states in spoken dialog systems. In *Proceedings of the 11th Anual Meeting of the Special Interest Group on Discourse and Dialog (SIGDIAL)*, pages 281–288, 2010.

[25] F. Mairesse, M. A. Walker, M. R. Mehl, and R. K. Moore. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500, 2007.

[26] S. Möller. *Quality of Telephone-Based Spoken Dialogue Systems*. Springer, New York, 2005.

[27] R. W. Picard. Affective Computing for HCI. In *Proceedings of HCI International (the 8th International Conference on Human-Computer Interaction) on Human-Computer Interaction: Ergonomics and User Interfaces*, volume 1, pages 829–833, Hillsdale, NJ, USA, 1999. L. Erlbaum Associates Inc.

[28] P. M. Podsakoff, S. B. MacKenzie, and N. P. Podsakoff. Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88:879–903, 2003.

[29] P. Rani, N. Sarkar, and J. Adams. Anxiety-based affective communication for implicit human-machine interaction. *Advanced Engineering Informatics*, 21(3):323–334, 2007.

[30] B. Reeves and C. Nass. *The Media Equation: How people treat computers, television and new media like real people and places*. CSLI Publications, Standford, 1996.

[31] J. Robison, J. Rowe, S. Mcquiggan, and J. Lester. Predicting user psychological characteristics from interactions with empathetic virtual agents. In *Proceedings of the 9th International Conference on Intelligent Virtual Agents*, IVA '09, pages 330–336, Berlin, Heidelberg, 2009. Springer-Verlag.

[32] W. E. Saris, J. E. Krosnick, and E. M. Shaeffer. Comparing questions with agree/disagree response options to questions with construct-specific response options. 2005.

[33] L. Shen, M. Wang, and R. Shen. Affective e-learning: Using "emotional" data to improve learning in pervasive learning environment. *Educational Technology and Society*, 12(2):176–189, 2009.

[34] J.-Y. Sung, L. Guo, R. E. Grinter, and H. I. Christensen. My Roomba is Rambo: Intimate Home Appliances. In *UbiComp*. Springer-Verlag, 2007.

[35] A. Tcherkassof, T. Bollon, M. Dubois, P. Pansu, and J.-M. Adam. Facial expressions of emotions: a methodological contribution to the study of spontaneous and dynamic emotional faces. *Journal of Social Psychology*, 37:1325–1345, 2007.

[36] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan-Kaufmann, San Francisco, 2005.

[37] S. Yildirim, S. Narayanan, and A. Potamianos. Detecting emotional state of a child in a conversational computer game. *Comput. Speech Lang.*, 25(1):29 – 44, 2011.