I *Feel* You: Towards Affect-Sensitive Domotic Spoken Conversational Agents

Syaheerah Lebai Lutfi*, Fernando Fernández-Martínez,
Andrés Casanova-García, Verónica López-Ludeña, and Juan Manuel Montero

Speech Technology Group, Universidad Politécnica de Madrid, Madrid, Spain
{syaheerah,ffm,acasanova,veronicalopez,juancho}@die.upm.es
http://www-gth.die.upm.es/

Abstract. We describe the work on infusion of emotion into limited-task autonomous *spoken conversational agents* (SCAs) situated in the domestic environment, using a **N**eed-inspired task-independent **Emo**tion model (NEMO). In order to demonstrate the generation of affect through the use of the model, we describe the work of integrating it with a natural-language mixed-initiative HiFi-control SCA. NEMO and the host system communicates externally, removing the need for the Dialog Manager to be modified as done in most existing dialog systems, in order to be adaptive. We also summarize the work on automatic affect prediction, namely frustration and contentment from dialog features, a non-conventional source, in the attempt of moving towards a more user-centric approach.

Keywords: Spoken Conversational Agents, affect prediction, domotic applications, Affective HiFi SCA, frustration, contentment, conversational features, satisfaction judgment.

1 Introduction

Emotion is quintessential for intelligence, to the point that psychologists and educators have re-defined intelligence to include emotion and social skill. With the mass appeal of computer-mediated agents, computers are no longer viewed as machines whose main purpose is to complete tasks, rather they are required to have the social abilities that humans naturally demonstrate in their daily interactions. Thus the developments of conversational agents typically move towards including socio-emotion content, which upgrades them to being socially intelligent.

* The work leading to these results has been supported by INAPRA (MICINN, DPI2010-21247-C02-02), TIMPANO (TIN2011-28169-C05-03) and ITALIHA (CAM-UPM) 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.

J. Bravo, R. Hervás, and M. Rodríguez (Eds.): IWAAL 2012, LNCS 7657, pp. 261–269, 2012.

This paper concerns the incorporation of a recently developed task-independent emotional model into a voice-only social domotic agent. Using this model, the generation of emotion is driven by *needs*, inspired by human's motivational system, hence called NEMO (Need-inspired EMOtion Model). The intention is to incorporate NEMO into existing SCAs in order to enable them to be affect-sensitive. This is accomplished by predicting user affective states and responding to them with appropriate affective responses, through an emotional text-to-speech system. The focus of the paper is not on NEMO in itself (which has been described elsewhere [13]), but how this model will be used in non-adaptive applications, in order to make them more adaptive to the users' emotion. Though NEMO is a generic and task-independent architecture, actual events and situations are required in a specific domain in order to run this model. Therefore to demonstrate affect sensing and generation through the use of this model, we describe the work of integrating this model with a natural-language mixed-initiative High-Fidelity-control spoken dialog (henceforth 'HiFi-NEMO') towards the goal of a socially intelligent HiFi agent. Specifically, this part is described in the first part of the paper. The second part focuses on building a real-time automatic detection of affect, as robust automatic detection is vital to any affect-sensitive system.

2 Infusing Emotions into the HiFi Agent

We attempt to infuse emotions into an existing HiFi agent using NEMO. The baseline (non-adaptive version) HiFi system is a proprietary system developed by Grupo Technología del Habla (GTH), (see details in [8]). As mentioned earlier, NEMO is a need-inspired system, whereby the agent elicits an emotion that is coherent with the situation *in view* of its different needs. One of the most influential needs for the agent's emotion modification, is the Success need. Therefore, this section focuses on the Success need. The Success level is influenced by various events that are related to different tasks. Previously this was done by updating a predefined percentage of an individual event (values differ according to events). For example, in the previous prototype of NEMO that was put to test with a domotic robot, Groucho, whose main tasks are to manage various domestic appliances (see [13,14][1]), an event detected by sensory inputs such as user touching or caressing the agent's face might have a fixed value of 0.5, indicating a medium success level or winning a game might have a fixed value of 0.7, a high success level. In integrating NEMO with the HiFi spoken conversational agent though (henceforth 'HiFi-NEMO'), we moved a step ahead by adopting a more user centric approach, using machine learning to *automatically predict* the said values by learning from past evaluation's data using a trained classifier, described later in Section 3. An interaction event is predicted as good or bad (and also of *how* good or bad) and the corresponding values will then be taken to compute the Success need.

[1] A couple of demos showing applications of different domains used to test this model can be acessed here:`http://www.syaheerah.com/?page_id=789`

2.1 Architecture of HiFi-NEMO

Most spoken dialog systems have an architecture that is similar to the HiFi SCA, as shown on the left side of Figure 1. The user utters a sentence and the Speech Recognizer captures the sounds from the user's speech, matches the recognized words against a given set of vocabulary. Then the matched words are passed to the Language Understanding module to extract the concepts (semantic information) of the sentence. A series of concepts are then passed to the Dialog Manager to activate dialog goals. The Dialog Manager decides both the actions to be taken and the feedback to the user for the current dialog turn, and passes the semantic information to the Natural Response Generator module to generate a suitable textual response to the user. The text-to-speech (TTS) module then synthesizes the message and speaks to the user. The original non-adaptive HiFi SCA version used a non-emotional commercial TTS. A detailed architecture of the non-adaptive system is given in [8]. In converting the HiFi SCA into an affect-adaptive system, its existing components were *not* modified. Instead the HiFi SCA communicates *externally* with NEMO. The interaction between the system's modules and NEMO is shown in Figure 1. The information flow is similar as described previously, but this time, the Dialog Manager additionally passes certain dialog features that are significant predictors of the user emotional state to the Affect Predictor. The Affect Predictor classifies the emotion state of the user following a Simple Logistics trained model. The classification result is then passed on to the need module to update the agent's Success need. Consider a user having a few bad dialog turns – perhaps the HiFi SCA failed to completely understand the user request and repeatedly asks the user to provide new information and extends the otherwise short dialog. In this case, the Dialog Manager sends certain relevant features (request turns, contextual information etc.) to the Affect Predictor. Based on these features, the Affect Predictor predicts that the user is frustrated. This information is then updated to the need module, which modifies the agent's Success need. The agent now perceives the user as being frustrated and therefore its Success need is low. The dynamicity of the need level also depends on the situations of the previous turns; consecutive or continuous prediction that the user is frustrated causes the agent's Success

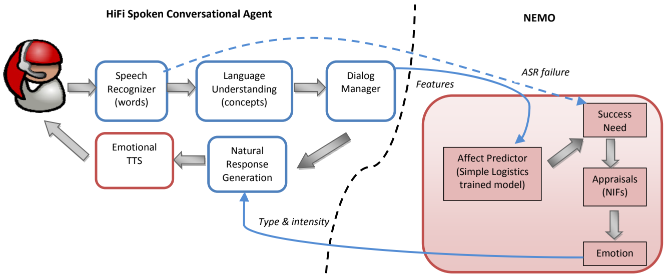
HiFi Spoken Conversational Agent
NEMO
Speech Recognizer (words)
Language Understanding (concepts)
Dialog Manager
Features
ASR failure
Success Need
Affect Predictor (Simple Logistics trained model)
Appraisals (NIFs)
Emotion
Emotional TTS
Natural Response Generation
Type & intensity

Fig. 1. The architecture of HiFi-NEMO

need satisfaction to decrease rapidly, and so when a good event (turn) appears right after, (and the user is now predicted to be in a positive emotion), the agent will not immediately change its state to a joyful one, but rather surprised or neutral, depending on the situation. Conversely, if the agent is in a joyful state for sometime, and continues with turns that are perceived as good (user predicted to be satisfied in consecutive turns), the drive to gratify its Success need will not be as significant as in the other case, and so its joyful state reaches its maximum and starts decaying into a a neutral state, though it continues to perceive the ongoing events as positive ones.

It should be noted however that the Dialog Manager receives dialog features only when the Speech Recognizer is successful in *detecting* the user's words. When a speech recognition failure occurs, the failure event updates the agent's Success need directly, bypassing the Dialog Manager (as indicated by the dotted arrow in Figure 1). A failure message is passed to the need module which decreases the agent's Success satisfaction level.

Next, the agent's Success need information updates the rest of the modules in NEMO and to generate an emotion that is coherent with the agent's assessment of its current Success need. Finally the chosen emotion matches against the natural response generation for a suitable response content and is synthesized into a speech response of a specific intensity of the chosen emotion by an Emotional TTS. This TTS is built by [2] and is used in replacement of the original neutral one. It is capable of generating speech in various colourings of the Big Six emotional categories, proposed by [6]:neutral, joy, sadness, fear, anger, surprise and also a combination of these. This is done by interpolating the prosodical variation of one emotion into another.

3 Automatic Detection of Affect

Real time automatic detection of emotion is vital to any affect-sensitive system. In this section we describe the method used to automatically predict the agent's Success need value of HiFi-NEMO, that will subsequently update the cognitive appraisals' module in order to generate a suitable affective response. As mentioned previously, in HiFi-NEMO's context, the success rate of an interaction modulates the agent's need, particularly its Success need. An interaction is deemed successful when the *user* is content with the agent's performance. Thus a fundamental challenge in converting a non-affective HiFi agent (or *any* systems in general) into an affective one is robust automatic detection of user affect. A highly satisfied user also satisfies the agent's success need, hence user-agent satisfactions have a positive linear relationship. User affect can be reflected in the user *satisfaction* judgment [1,5,12] and the relationship of affect and satisfaction judgment have been empirically proven in [10,11] and also in our work, which will be further described. To model user affect, we used *satisfaction rating* as the target and *conversational features* as predictors, obtained from a corpus collected in a past evaluation [7]. 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 [4]. User affect could be mined from conversational elements, which are always cheaper and are usually obtained with little or no computational overhead. However, since the focus of this paper lies on modeling affect in the HiFi agent, we limit this section into summarizing the method and the outcome of the experiments carried out in automatic affect detection using the data from two studies that will be described shortly.

We also 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 contented with the system when the user gave a score of 5 or 'excellent' (in a 5-point scale), and rather frustrated when he or she gave a score of 3)[2]

3.1 User and Annotator Studies

To model affect by predicting user satisfaction, we used the HiFi-AV2 corpus [9], collected during a *user* study. HiFi-AV2 consists of audiovisually recorded information of real, non-acted interactions (N=190 interaction sessions) between user and non-adaptive version of the HiFi agent. In this study, users interacted with the HiFi agent hands on and 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, very good). Later, we used a reduced version of the same corpus to obtain satisfaction and affect-labelled data from several independent *annotators*. The corpus was reduced to 10 speakers that were chosen randomly (N=100 sesions) to downsize manual labelling efforts. This study was similar to the first one, except that the annotators *also* perceived user emotion in each interaction - 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 (the nuances within the six basic emotions proposed by [6]) 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). Additionally we also now have *affect-labelled* data by annotators.

3.2 Affect Classification from Conversational Features

In order to obtain a model of user affect, we conducted two experiments; Experiment I was evaluated on the satisfaction-labelled data from both user and

[2] Depending on the model that was chosen - different models have different groupings of scores, elaborated later in Section 3.3. A score of 3, for example, may either represent a low-intensity frustration (category Three version 2) or slight contentment (category Three version 1).

annotator subject group. Experiment II involved the emotion-labelled data by annotators. 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) [15] was used for these purposes. One or more classification algorithms were chosen from different categories including rule-based classifiers (ZeroR as baserate and OneR), functions (SimpleLogistic, SMO), meta classification schemes (Multischeme, MultiBoost, AdaBoost) and trees (J48). A 10-fold cross validation technique was used for all the classification tasks.

3.3 Data Redistribution

All satisfaction-labelled datasets were first resampled in order to obtain a more uniformed 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 1. This way we were also able to determine which clusters obtained optimized classifications.

4 Summary of Results and Discussions

4.1 Experiment I (Model 1)

Table 2 presents the statistically significant improvements of classification results over baserate in percentage accuracy for Experiment I. Results revealed that there was a significant effect of the subject type: $F(1{,}40)$=83.07, p<.001, partial η^2=.68. Classifiers evaluated on user data mostly revealed worse results than baserate with exception to SMO, whilst at least three classifiers that were evaluated on annotator data show significant improvement (at p<.001) over baserate in each category, with exception to category 3V2 and 2V2. This indicates that most classifers were able to predict satisfaction from dialog features based on the *annotator* data, suggesting that annotators were more impartial when judging the HiFi agent. Classification evaluated on annotator data however yielded interesting result and is more suitable to be used for user affect modeling. Thus we now focus on the results from annotator data. The chart in Figure 2 illustrates the interactions between the factors that obtained the best classification improvements for *annotator* dataset, that are statistically significant.

Table 1. Datasets re-clustered according to similarity of score points into all possible combinations of classes

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

4.2 Experiment II (Model 2)

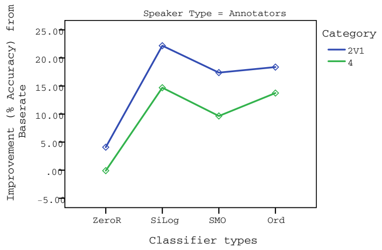

Fig. 2. Summarized interaction chart for annotators dataset

Experiment II involved classification evaluated on the emotion-labelled data based on inter-annotator agreement. The computation that derive such agreement was adopted from [3]. Results revealed that the SMO scheme yielded the best statistically significant (at p<.01) improvement of classification over baserate, with improvement of 13.2%, followed by the Ordinal and Simple Logistics schemes, with 9.8% and 9.2% improvements respectively, both statistically significant at p<.05.

Both satisfaction and emotion-agreement data are significantly correlated (Pearson r=.29, at p<.01) but provide complementary information to model user affect. The data used for Model 1 involved users providing a satisfaction score at the end of each interaction; a global average score that represented the user's opinion of the agent - that could also be considered as the best representation of score for each turn, should we have to rate the system on a turn basis. Thus this information is suitable to be used to predict user affect on a *turn* basis. On the other hand, Model 2 involved annotators perceiving user emotion at different locations of a particular interaction, when the emotions displayed were most obvious. However the information on the exact locations within the interaction was not noted. Hence this kind of prediction is suitable to be performed at the end of the whole interaction, in order for the agent to be alerted that the user has been frustrated at some point during the interaction.

Table 2. Comparisons of significant improvements in classification accuracies in detecting satisfaction score from conversational features (for both *user* and *annotator* datasets)

Category	Classifiers							
	Base rate		SiLog		SMO		Ord	
	U	A	U	A	U	A	U	A
Five	38.0	36.0	-	49.3	-	44.6	-	51.3
Four	38.0	36.0	-	53.1	-	43.4	-	52.0
Three (version 1)	71.0	47.0	-	64.0	-	61.1	-	62.5
Three (version 2)	38.0	53.0	-	-	50.7	-	-	-
Two (version 1)	71.0	53.0	-	75.0	-	74.4		69.4
Two (version 2)	93.0	83.0	-	-	-	-	-	-

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

5 Conclusions, Current and Future Directions

We propose an approach of incorporating emotions into spoken dialog systems using NEMO. We demonstrated this by describing the integration on a proprietary baseline system, a non-adaptive HiFi agent.

Our main contribution is to show that the Dialog Manager of the baseline system were neither modified nor hardwired with affect-related rules as done in most existing dialog systems, in order to be emotionally rich. Instead the dialog manager communicates with the emotion system and manages the dialog using the emotionally-relevant features provided by the emotion classifier. Additionally, the emotion classifier is based on a learning-by-example method (of past data), not an imperative, hand-crafted one. These minimize costs in two ways; first, not only the requirement for domain-specific expert knowledge can be reduced, but the adaptation is also more user centric. Second, this model could also be re-used in new but similar domains, with minimum labour.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, however within a limited-task domestic domain. The conversational features were the 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 both satisfaction judgments and emotions) and smaller array of features for classification tasks. Nevertheless, emotion classification improvements achieved statistically significant results over baserate.

We have implemented the emotion classifier into the emotion model, and the latter is incorporated into the HiFi agent to make it more affective. We have also developed a suitable response generation model according to the various intensities of the predicted user frustration and contentment. Future work involves a series of cross evaluations that will be conducted between users and adaptable/non-adaptable versions of the agent to compare the findings.

References

1. Bailey, J.E., Pearson, S.W.: Development of a tool for measuring and analyzing computer user satisfaction. Management Science 24, 530–545 (1983)
2. Barra-Chicote, R., Yamagishi, J., King, S., Montero, J.M., Macias-Guarasa, J.: Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. Speech Commun. 52(5), 394–404 (2010)
3. Callejas, Z., López-Cózar, R.: On the Use of Kappa Coefficients to Measure the Reliability of the Annotation of Non-acted Emotions. In: André, E., Dybkjær, L., Minker, W., Neumann, H., Pieraccini, R., Weber, M. (eds.) PIT 2008. LNCS (LNAI), vol. 5078, pp. 221–232. Springer, Heidelberg (2008)
4. D'Mello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B., Graesser, A.: Automatic detection of learner's affect from conversational cues. User Model User-Adap. Inter. 18, 45–80 (2008)
5. Doll, W.J., Torkzadeh, G.: The measurement of end-user computing satisfaction. MIS Quarterly 12, 259–274 (1988)
6. Ekman, P., Friesen, W.: The Facial Action Coding System: A technique for the measurement of facial movement. Consulting Psychologists Press (1978)
7. Fernández-Martínez, F., Bläzquez, J., Ferreiros, J., Barra-Chicote, R., Macias-Guarasa, J., Lucas-Cuesta, J.M.: 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)

8. Fernández-Martínez, F., Ferreiros, J., Lucas-Cuesta, J.M., Echeverry, J.D., San-Segundo, R., Córdoba, R.: 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). Springer, Bilbao (2010)
9. Fernández-Martínez, F., Lucas-Cuesta, J.M., Chicote, R.B., Ferreiros, J., Macías-Guarasa, J.: 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 2010), European Language Resources Association (ELRA), Valletta, Malta (May 2010)
10. Gelbrich, K.: Beyond just being dissatisfied: How angry and helpless customers react to failures when using self-service technologies. Schmalenbach Business Review 61, 40–59 (2009)
11. Kernbach, S., Schutte, N.S.: The impact of service provider emotional intelligence on customer satisfaction. Journal of Services Marketing 19(7), 438–444 (2005)
12. Locke, E.A.: The nature and causes of job satisfaction. Consulting Psychologists Press, Palo Alto (1976)
13. Lutfi, S., Barra-Chicote, R., Lucas-Cuesta, J., Montero, J.: Nemo: Need-inspired emotional expressions within a task-independent framework. In: Proc.of Brain Inspired Cognitive Systems (BICS), Madrid, Spain (July 2010)
14. Sanz-Moreno, C., Lutfi, S., Barra-Chicote, R., Lucas-Cuesta, J., Montero, J.: Desarrollo de un asistente domótico emocional inteligente. In: XIX Jornadas Telecom I+D, Madrid, Spain (November 2009)
15. Witten, I.H., Frank, E.: Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, San Francisco (2005)