NEMOHIFI: An Affective HiFi Agent

Syaheerah Lebai Lutfi*
School of Computer Sciences,
University Sains Malaysia
11800 Penang
Malaysia
syaheerah@cs.usm.my

Fernando Fernández-Martínez
Departamento de Teoría de la Señal y Comunicaciones,
Universidad Carlos III de Madrid
28911 Madrid, Spain
ffm@tsc.uc3m.es

Jaime Lorenzo-Trueba, Roberto Barra-Chicote, Juan Manuel Montero
Grupo Technología del Habla
Universidad Politécnica de Madrid
28040 Madrid, Spain
jaime.lorenzo,barra,juancho@die.upm.es

ABSTRACT

This demo concerns a recently developed prototype of an emotionally-sensitive autonomous HiFi Spoken Conversational Agent, called NEMOHIFI. The baseline agent was developed by the Speech Technology Group (GTH) and has recently been integrated with an emotional engine called NEMO (**N**eed-inspired **Emo**tional Model) to enable it to adapt to users' emotion and respond to the users using appropriate expressive speech. NEMOHIFI controls and manages the HiFi audio system, and for end users, its functions equate a remote control, except that instead of clicking, the user interacts with the agent using voice. A pairwise comparison between the baseline (non-adaptive) and NEMOHIFI is also presented.

Categories and Subject Descriptors

H.4 [**Human Centered Computing**]: Human Computer Interaction, Emotion; D.2.8 [**Computing Methodologies**]: Artificial Intelligence—*Machine translation, Natural language generation, Speech recognition*

Keywords

Affective HiFi, Spoken Conversational Agent, Affective Agent, NEMO, NEMOHIFI, Speech Technology

1. INTRODUCTION

This paper presents a summary of the the NEMOHIFI[1] a Spoken Conversational Agent (SCA) that manages the HiFi audio system. For end users, its functions equate a remote control (select a CD, track or radio channel, record music, change channels etc.), except that instead of clicking, the user interacts with the agent using voice. The baseline (non-adaptive version) HiFi system is a proprietary system developed by GTH [1].

1.1 The Emotional Engine, 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 neutral-voiced commercial TTS. A detailed architecture of the baseline system is given in [1] and the evaluation results reported in [2].

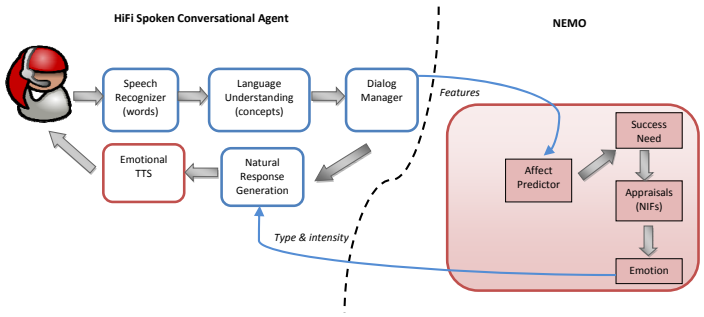

Figure 1: The architecture of HiFi-NEMO

*Corresponding Author

[1]The work leading to these results has received funding from the European Union under grant agreement nr. 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (Comunidad Autonoma de Madrid, S2009/TIC-1542) projects. Jaime Lorenzo's work has been partially funded by Universidad Politecnica de Madrid under grant SBUPM-QTKTZHB. Authors especially thank Fernando González, Javier Ferreiros, Julián David Echeverry and Beatriz Martínez for helping with the demo recording.

ICMI'13, Dec 09-13 2013, Sydney, NSW, Australia
ACM ACM 978-1-4503-2129-7/13/12.
http://dx.doi.org/10.1145/2522848.2531755 .

In converting the HiFi SCA into an affect-adaptive system, its existing components were *not* modified, except for the Natural Response Generation textual content (NRG). 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 need, particularly its 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 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.

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, known as the GTH-EMO TTS, built by [3]. GTH-EMO TTS is used in replacement of the original neutral one. Relevant emotions for this demo would be colourings of *anger*, *sadness* and *neutral*.

1.2 Automatic Affect Prediction

The user state is modelled using an existing dataset, that was later re-labelled by 17 independent annotators. A set of conversational features are used as the predictors and user satisfaction rating as the target. Machine learning experiments were conducted on two datasets, users and annotators, which were then compared in order to assess the reliability of these datasets. Our results indicated that standard classifiers were significantly more successful in discriminating emotions (especially *contentment* and *frustration*) and their intensities (reflected by user satisfaction ratings) from annotator data than from user data. The speech recognition accuracy is 75% on average. See [5] for details.

1.3 User study

A user study was conducted with 24 subjects, in which both versions of the agent (non-adaptive and emotionally-adaptive) was compared. Subjects are Spanish native speakers that were mostly University students (between 18-35 years). Evaluation results showed that NEMOHIFI was significantly preferred over the baseline agent - 25.0% preferred the baseline agent and 75.0% opted for NEMOHIFI, $\chi^2(1)=36.0$, p=.000 [6]. The results provide substantial evidences with respect to the benefits of adding emotion in a spoken conversational agent, especially in mitigating users' frustrations and ultimately improving their satisfactions. Pairwise t-test results for subjective qualities between both versions are shown in Table 1, where $*$ denotes substantially statistically significant results. For more details, see [6][2]

Table 1: *t*-test results comparing the mean subjective ratings between baseline and NEMOHIFI agents

Metric	Mean BASEHIFI	Mean NEMOHIFI	BASEHIFI–NEMOHIFI	*t*	Sig.
PERFORMANCE	1.13	1.41	-.28	-1.43	.16
RESPONSE	1.46	1.29	.17	1.05	.30
VOICE	1.67	2.14	-.47	-3.31	.001*
ATTITUDE	.90	2.11	-1.21	-7.76	.000*
NATURALNESS	0.42	1.54	-1.32	-7.41	.000*
GSS	1.00	1.69	-1.21	-4.89	.000*

Figure 2: Screenshot from the demo.

References

[1] Fernández-Martínez, F.; Ferreiros, J.; Lucas-Cuesta, J. M.; Echeverry, J. D.; San-Segundo, R. and Córdoba, R., Flexible, Robust and Dynamic Dialogue Modeling with a Speech Dialogue Interface for Controlling a Hi-Fi Audio System. *Proceedings of the IEEE Workshop on Database and Expert Systems Applications (DEXA 2010)*, Springer, 2010

[2] Fernández-Martínez, F.; Ferreiros, J.; Lucas-Cuesta, J. M.; Montero, J. M.; San-Segundo, R. and Córdoba, R. Towards building intelligent speech interfaces through the use of more flexible, robust and natural dialogue management solutions. *Interacting with Computers*, 2012, 24, 82-498.

[3] Lorenzo-Trueba, J.; Watts, O.; Barra-Chicote, R.; Yamagishi, J.; King, S. and Montero, J. M., Simple4All proposals for the Albayzin Evaluations in Speech Synthesis. *VII Jornadas de Tecnología del Habla (Iberspeech2012)*, 2012

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

[5] Lutfi, S.; Fernández-Martínez, F.; Lucas-Cuesta, J.; López-Lebón, L., A Satisfaction-based Model for Affect Recognition from Conversational Features in Spoken Dialog Systems, *Speech Communication*, 2013. Accepted for publication.

[6] Lutfi, S.; *User-centric need-driven affective model for Spoken Conversational Agents: Design and Evaluation.* Unpublished Thesis. E.T.S.I.T, University Politécnica de Madrid, Spain

[2]The demo of the interaction between a user and NEMOHIFI (Figure 2) could be viewed here: `http://www.syaheerah.com/?page_id=789`