An Emotionally Aware Embodied Conversational Agent

Socially Interactive Agents Track

Samuel S. Sohn
Rutgers University
samsksohn@gmail.com

Xun Zhang
Rutgers University
xun.zhang@rutgers.edu

Fernando Geraci
Rutgers University
fergeraci@gmail.com

Mubbasir Kapadia
Rutgers University
mk1353@cs.rutgers.edu

KEYWORDS

Verbal Behavior Understanding; Nonverbal Behavior Understanding; Affect (Emotion, Mood, Personality); Novel Architectures

ACM Reference Format:
Samuel S. Sohn, Xun Zhang, Fernando Geraci, and Mubbasir Kapadia. 2018. An Emotionally Aware Embodied Conversational Agent. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

A conversation is a bilateral communication process that involves a deeper understanding of exchanged ideas than that which is needed for a unidirectional issuance of commands. Among contemporary conversational agents, the most prevalent are those which cater to the latter approach.

The key factors for maintaining the fluency and immersiveness of the communication with an intelligent conversational agent are understanding and behavior. Understanding is a product of the analysis of the input to the agent, which begins with multimodal information from image and audio streams. Recognition of emotions, text, and intent is performed on these streams in order to facilitate the analysis. Behavior describes the output of the agent's deliberation of the input signals; i.e., the means for the agent to respond to the user. This behavior is realized through speech, facial expressions, and body language in the form of audio and animations.

Existing conversational agents place the burden of the conversation on the user, requiring the user to both elicit interactions and formulate speech catering to the agent's ability to understand. Although technological advancement is making the interaction less cumbersome, a major shortcoming is the user-centric focus that has left the agent underdeveloped. Atomic user-query to agent-response interactions can improve, but without eliciting the user's engagement, conversations will not improve in quality. Therefore, we present a framework that focuses on developing the agent such that it engages the user through multiple modes.

2 RELATED WORK

Intelligent conversational agents have been a focus of research for decades now. Eliza [18] and Parry [3] are some of the first chat-bots ever created. Moving two decades forward, we can find ALICE [8], the Artificial Linguistic Internet Computer Entity. It is another type of chatbot inspired by Eliza using AIML files to store its knowledge. There have also been approaches to having a general-domain question-answering agent in a museum [11, 15]. Leuski et al. [11] and Robinson et al. [15] describe a general question-answering agent used in a such a scenario, trained using utterances. Kapadia presents a general framework for authoring, animating and, evaluating autonomous virtual human simulations [9].

A simple approach using hard-coded rules can benefit from an additional statistical model layer, as seen in [20]. It is also possible to learn generation rules from a minimal set of authored rules or labels using the vector space framework as described in [1]. Morris [13] provides an example of using conversational agents in games. Tarau and Figa [17] provide another example of an agent in an educational environment in which a Prolog database is used to store a variety of possible question-answer sets that the agent can pick from.

Emotions can be analyzed from different aspects. The Facial Action Coding System (FACS) [4] established a formal way to describe facial expressions. Emerich et al. proposed a emotion recognition method by analyzing speech and facial expressions [5], whose results showed the correlation of facial expressions and speeches. Potard et al. proposed an evaluation baseline for synthesized emotional speech and facial expression [14], and a recent study [2] shows that correctly modeling emotions and large-scale personalities for virtual agents can greatly boost the user experience and engagement.

Finally, believable agent embodiment plays a key role in engaging human-computer interactions. There are frameworks such as SmartBody which address complex problems, such as lips synchronization [19] and motion blending techniques [7], and ADAPT [10], a complete testbed of purposeful human characters in a rich virtual environment. The concept of meaningful embodied agents can be expressed in terms of their capacity to perform behaviors, as presented in parameterized behavior trees [6]. Stratou and Morency proposed a context-aware nonverbal behavior analysis framework that has been applied for medical use[16].

3 MULTIMODAL SENSING

Figure 1 shows the overall multimodal emotion and user intent sensing structure in our framework. This structure can detect and represent emotions from various sources, recognize speech and analyze user intents. The output of the sensing layer is passed to

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

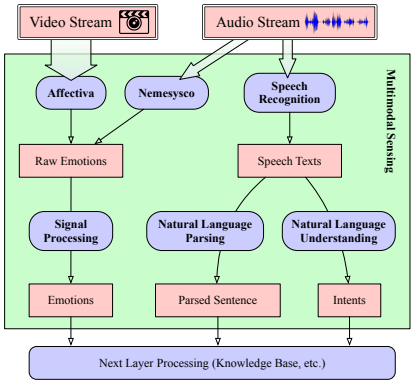

Figure 1: The pipeline of the multimodal sensing framework

the deliberation layer to store knowledge gained from the user and generate the appropriate responses to the user's intents.

3.1 Emotion Recognition

Due to existing limitations in emotion sensing, sensor fusion is needed in order to obtain accurate emotion signals in varying conditions. Namely, while the user is speaking, buccal facial expressions significantly decrease signal-to-noise ratio for visual emotion recognition (ER), and otherwise, auditory ER (via Nemesysco™) is less effective than visual ER (via Affectiva™).

Different solutions for ER provide recognition results in varying dimensions with no distinction made between the probability of an emotion and the intensity of the emotion. Therefore, we propose a standard emotion space to unify present and future ER solutions.

3.2 ECI Emotion Model

The ECI (Emotion, Certainty, Intensity) emotion model is a representation of the Lövheim cube of emotion that was inspired by the HSV color model [12]. The Lövheim model is a unified representation of Ekman's six basic emotions: anger, disgust, fear, happiness, sadness, and surprise [4]. In the Lövheim model, emotions are a blend of signals from three neurotransmitters: serotonin, dopamine, and noradrenaline. These signals are represented as orthogonal axes in three-dimensional space and the emotions are maximal at corners of a cube in the first octant with a vertex at the origin, making this model an analogue to the RGB color model. Therefore, the equations for RGB to HSV conversions are used for the conversion of the Lövheim model into the ECI model. In the ECI model, Emotion (analogous to Hue) is a degree (i.e., non-discrete scalar) value representing the emotion. Certainty (analogous to Saturation) is a measure of how distinct an emotion is from general excitement, which is located along the space diagonal from the origin. Intensity (analogous to Value) measures the intensity of the emotion or general excitement.

The ECI model is a compromise between the valued legitimacy of the Lövheim model and the benefits in intuitiveness of the cylindrical coordinate system. As such, it will be robust to advances in ER technology.

3.3 Text and Intent Recognition

Audio signals are also used for automatic speech recognition (ASR), and the resulting text is used for intent recognition through natural language understanding (NLU). The focus of this research is not on improving technologies for ASR and NLU, but on leveraging their interfaces into a framework for an intelligent virtual agent. Therefore, platform-specific commercial solutions and Rasa NLU are used for ASR and NLU respectively.

Intent recognition is used to categorize sentences and extract entities (i.e., words that are useful to the categories). For example, "I am feeling happy." would be categorized as "state_mood" and "happy" would be the entity. This level of granularity is ideal for a manual approach to content authoring; i.e., intent-responses mappings, which act as the agent's deliberation.

4 KNOWLEDGE REPRESENTATION

Based on the prior intent recognition, an intuitive representation of the knowledge gained from the user is in the form of category-entity pairs (some singleton and others not). For example, the previous "state_mood" example would be represented as a singleton and a category such as "state_interest" would not.

5 MULTIMODAL BEHAVIOR SYNTHESIS

After perception (i.e., multimodal sensing) and deliberation (i.e., intent-response mappings), a behavior mechanism realizes responses through synchronized audio and animations. To facilitate synchronization across different modes of output (e.g., audio, lip animations, and body language animations), the behavior mechanism is built entirely from behavior trees [6]. Behavior trees simplify the parallel execution of these output modes and allow for their modularization and parameterization such that 1.) they can be considered as a single, rich output mode and that 2.) they can be varied (by the parameters) with ease.

The aforementioned synchronization of behavior trees is achieved by solving a particular scheduling problem. The input consists of chronological orderings of behavior trees across all involved output modes, cross-modal synchronizations (e.g., that two behavior trees must end at the same time), and output mode priorities that preserves the highest priority mode and incrementally synchronizes the others. To illustrate this naively, suppose that for two output modes, audio (high priority) and lip animations, there are an audio clip and the corresponding lip animations synchronized such that both modes start and end at the same times. As audio is preserved, the lip animations will be fit to match the audio time frame. If the lip animations exceed the audio duration, the animations must be trimmed or sped up; otherwise, they must be padded or slowed down.

6 CONCLUSIONS

The implementation culminates in an agent who is able to interact with a human person, bringing human-computer interaction to a new level of complexity by being emotionally aware and expressive through rich behaviors. The integration of modules for emotion recognition and natural language understanding allows for upgrading as these technologies advance. This intelligent virtual agent framework is capable of both withstanding the test of time and enabling a compelling virtual conversational agent.

REFERENCES

[1] Rafael E Banchs and Haizhou Li. 2012. IRIS: a chat-oriented dialogue system based on the vector space model. In *Proceedings of ACL 2012 System Demonstrations*. 37–42.

[2] Andry Chowanda, Peter Blanchfield, Martin Flintham, and Michel Valstar. 2016. Computational Models of Emotion, Personality, and Social Relationships for Interactions in Games. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2016)*. 1343–1344.

[3] Colby. 1972. PARRY. (1972). https://www.chatbots.org/chatbot/parry/

[4] Paul Ekman and Wallace V. Friesen. 1978. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press.

[5] Simina Emerich, Eugen Lupu, and Anca Apatean. 2009. Emotions Recognition by Speech and Facial Expressions Analysis. In *2009 17th European Signal Processing Conference*. 1617–1621.

[6] Alexander Shoulson et. all. [n. d.]. Parameterizing Behavior Trees. In *Motion in Games, Edinburgh, UK, November, 2011*. https://doi.org/10.1007/978-3-642-25090-3_13

[7] Andrew W. Feng, Yazhou Huang, Marcelo Kallmann, and Ari Shapiro. 2012. An Analysis of Motion Blending Techniques. In *The Fifth International Conference on Motion in Games*.

[8] A.L.I.C.E AI Foundation. 2002. ALICE. (2002). http://www.alicebot.org/

[9] Mubbasir Kapadia. 2011. *Authoring and Evaluating Autonomous Virtual Human Simulations*. Ph.D. Dissertation. Los Angeles, CA. Advisor(s) Faloutsos, Petros.

[10] Mubbasir Kapadia, Nathan Marshak, and Norman I. Badler. 2014. ADAPT: The Agent Development and Prototyping Testbed. *IEEE Transactions on Visualization and Computer Graphics* 99, PrePrints (2014), 1. https://doi.org/10.1109/TVCG.2014.251

[11] A. Leuski, R. Patel, D. Traum, and B. Kennedy. 2006. Building effective question answering characters. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. 18 – 27.

[12] H. Levkowitz and G.T. Herman. 1993. GLHS: A Generalized Lightness, Hue, and Saturation Color Model. *CVGIP: Graphical Models and Image Processing* 55, 4 (1993), 271 – 285. https://doi.org/10.1006/cgip.1993.1019

[13] T.W. Morris. 2002. Conversational Agents for Game-Like Virtual Environments. In *AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment*. 82–86. http://www.qrg.northwestern.edu/Resources/aigames.org/papers2002/TMorris02.pdf

[14] Blaise Potard, Matthew P. Aylett, and David A. Baude. 2016. *Cross Modal Evaluation of High Quality Emotional Speech Synthesis with the Virtual Human Toolkit*. Springer International Publishing, Cham, 190–197. https://doi.org/10.1007/978-3-319-47665-0_17

[15] S. Robinson, D. Traum, M. Ittycheriah, and J. Henderer. 2008. What would you ask a conversational agent? Observations of human-agent dialogues in a museum setting. In *LREC 2008 Proceedings*.

[16] G. Stratou and L. P. Morency. 2017. MultiSense #8212;Context-Aware Nonverbal Behavior Analysis Framework: A Psychological Distress Use Case. *IEEE Transactions on Affective Computing* 8, 2 (April 2017), 190–203. https://doi.org/10.1109/TAFFC.2016.2614300

[17] P. Tarau and E. Figa. 2004. Knowledge-Based Conversational Agents and Virtual Storytelling. In *ACM Symposium on Applied Computing*. 39–44. http://www.cse.unt.edu/~tarau/research/2003/vschat.pdf

[18] Weizenbaum. 1966. Eliza. (1966). http://www.med-ai.com/models/eliza.html

[19] Yuyu Xu, Andrew W. Feng, Stacy C. Marsella, and Ari Shapiro. 2013. A Practical and Configurable Lip Sync Method for Games. In *ACM SIGGRAPH Motion in Games*. Dublin, Ireland. http://ict.usc.edu/pubs/A%20Practical%20and%20Configurable%20Lip%20Sync%20Method%20for%20Games.pdf

[20] Steve Young, Milica Gasic, Simon Keizer, FranÃğois Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management.. In *Comput. Speech Lang.* 24(2):150âĂŞ174.