Towards an Expressive Embodied Conversational Agent Utilizing Multi-Ethnicity to Augment Solution Focused Therapy

Mark Allison
School of Computing and Information Sciences
Florida International University
Miami, Florida, USA.

Lynn Marie Kendrick
College of Education, Leadership and Counseling
St. Thomas University
Miami, Florida, USA.

Abstract

In this article, we present ongoing research, EMO, an affective embodied conversational agent platform, aimed at depicting multi-ethnic, multi-modal communication patterns in a credible manner. We employ the methodology of integrating counseling concepts early in the design to effectively target a specific domain. The system is geared to augment solution focused therapy. We present a prototype of the architecture as proof of concept and evaluate the platform for affect portrayal.

Introduction

Expressive embodied conversational agents (EECA) are means to credibly portray non-verbal behavior to increase engagement (Cowell and Stanney 2005). While affective agent design as a disciple is relatively new, tremendous advances have been made. These advances, however significant, still lack much of the credibility and engagement inherent in the ideal model; the human being. The aim of the research presented is the development of an agent platform to augment Solution Focused Therapy (SFT) by introducing non-cognitive emotion portrayal and regulation based on ethnicity and gender.

In this article, we describe, EMO, ongoing research towards an affective multimodal system aimed at creating believable agents (including facial expressions, body language and voice) capable of expressing affect in terms of short, medium and long term aspects of affect namely emotions, moods and personality respectively. We exploit avatar affective rendering with a dialog system so as to enable the avatar to converse with a human on an identified restricted domain, SFT. We do not propose EMO to be utilized in lieu of treatment but rather employed voluntarily as a augmentation to support therapist and client in the therapy process.

There exists the need and enormous potential of incorporating user-centric technologies to facilitate therapy in mental health situations (Lisetti 2008) (Coyle et al. 2007). While human computer interface has been well researched the domain of mental health offers distinctive challenges due to the somber ramifications of mental health concerns. Our approach recognizes the challenges and propose an architecture based on the guidelines for design and evaluation proposed in (Doherty, Coyle, and Matthews 2010) of a technology designed in collaboration with mental health professionals and adhering to the therapy model. Ethnic and gender factors contribute substantially to the therapy process (Erdur, Rude, and Barón 2000) are a core architectural concern and of the platform.

The contributions of this paper are:

- An architecture to support the portrayal of affect within a EECA, based on ethnicity and gender
- Behavioral inhibition to decouple experienced from displayed emotions based on table lookup
- A list transition system to constrain unnatural mood transitions
- An evaluation of the display subsystem.

This paper is organized as follows: We begin by presenting a background of EECA's and SFT then provide a motivating scenario to thread the article. Subsequently, we describe the architecture of the platform and evaluate the display component. Related works are discussed and we conclude with an overview and future directions of the research.

Background

In order to facilitate a knowledge base, we present background in the problem and solution domains.

Expressive Embodied Conversational Agents

Expressive Embodied Conversational Agents (EECA) are a subset of Human Computer Interfaces (HCI) which seeks the development of a more natural means to transfer information and the internal affective state. The transference of internal affective state is critical to effective communication. The inclusion of affect in agent design represents a paradigm shift recognizing the intricate binding of affect to social intelligence. The works of Reeves and Nass (Reeves and Nass 1996) and subsequently (Nass 2000), demonstrates that humans tend to anthropomorphize, transferring expectations and social rules towards a virtual agent. As such, inherent within EECAs, are an effective means by which a machine may communicate a concept, idea or strategy due to the ability to additionally communicate through non-verbal means.

Solution Focused Therapy

We propose our approach as an augmentation of solution focused therapy to underrepresented demographics. Solution focused therapy (SFT) is a method used to help a client achieve a brief solution to a problem. Typically the issues are resolved in two or three sessions.These communication sessions focus on helping a client deal with internal conflict in a short time period. SFT has experienced increased popularity due to its fast solution process and getting the client to look for solutions that may have worked in the past (Cade and O'Hanlon 1993) (De Shazer 1985). In being a relatively quick approach, this method helps clients to focus on the positive and not the negative aspect of what is being experienced. It is through the process of the therapeutic session, that the therapist helps the client change their cognitive process as well as to engage the client through actions that lead to new perspectives and in effect will change their behavior. This process helps the client to see that problems have exceptions and the exceptions are the solutions. The client can simply change by the way that they are thinking about a problem or rather the way that they discuss the problem. We targeted this particular approach to therapy due to its duration and ability to utilize EMO conversation logs.

Motivation

To motivate the necessity for this research direction we present a fictitious scenario which would typify the use of an augmented therapy platform.

Henry is a Hispanic male, age 28, who had served in the US Army for four years. He spent the last two years in Afghanistan and he currently serves as a Logistical Supervisor/ Battalion Maintenance Sergeant. He was discharged and has had some major issues in regards to mortally wounding enemy combatants and have seen his comrades killed during the time that he served in the military. He fully understands that he had to do his job but also is saddened by his actions. In his cultural community Hispanic males rarely seek counseling, however he knows that he needs to see someone to resolve this issue. A strong religious background triggers internal conflict. After returning to the United States, he is constantly having dreams and issues with adjusting to a civilian lifestyle. He shares many stories of life in Afghanistan but wants to be able to return to normalcy and try to adjust to being back in the US.

He has sought counseling, however due to financial constraints and the desire to keep therapy sessions secretive due to fear of being ostracized, therapy sessions are minimal though the need is great.

EMO Architectural Components

The architecture of EMO is based on loosely coupled subsystems responsible for effective affect portrayal with focused dialog based on domain expert vignettes. The major components, shown in Figure 1, are the display, affect and dialog subsystems.

EMO Display and Voice

Following (Paleari and Lisetti 2006), the three dimensional animation of the multi-ethnic affective agents are accomplished through the Haptek SDK. Character movements are achieved through proprietary API commands which had to be translated from FACS, the Facial Action Coding System (Ekman 1999), to achieve the platform for our desired facial affect portrayal. FACS classifies the facial expression of emotion using Action Units,(AU). The 44 Facial AUs comprises specific musculature movements and may combine to produce a facial expression of emotion. Figure 2 shows how AUs 1,2 and 5 are combined to display novelty. We need to make it explicitly clear at this juncture that not all the AUs required could be mapped to the proprietary parameters of the Haptek system, however the decision was made to continue its use as the benefits of the blending musculature and fatty body deformation proprietary techniques produced the desired natural characters movements we required.

Affect Portrayal Considerations The face is the primary expressive region of the body (Argyle 2007). Since humans transfer social rules to virtual agents (Nass 2000), we can infer that the face will be integral in the communication of affect for our platform. To enhance engagement we extend the portrayal to full bodied avatars. Body language, posture and gestures, which is within the capabilities of EMO, contributes to the transmission of the affective message (Bianchi-Berthouze and Cairns 2006) and is a significant modality, however the extent to which the communication is facilitated should be weighed with extenuating factors when deciding which avatar is to be used. Improper usage of the body may detract from the facial message and may even be construed as deception (Ekman 1997). As an example, the expressiveness sought by the addition of a body may compromise the overall message as latency becomes a factor; animating the said body will require additional computational resources.

Full bodied agents are able to use hands to elaborate a spatial component where the verbal channel may not be as effective i.e. The cat was this big and was sleeping over there (J Cassell, 2001). Communication via deictic gestures, however are not a common occurrence in normal conversations within our domain therefore it is ignored. Some of the posture considerations implemented are crossing arms for resistance fist- clenching for aggressiveness and sulking to display sadness as shown in Figure 4,

Avoiding the Uncanny Valley Since the engagement of the user is the ultimate goal of the design of EMO, we strive towards the use of photorealistic embodiments however there is a caveat called the uncanny valley which disengages the user that must be avoided. The uncanny valley (Mori 1970) is the experience of unease one develops as the resemblance of a nonhuman entity achieves near human

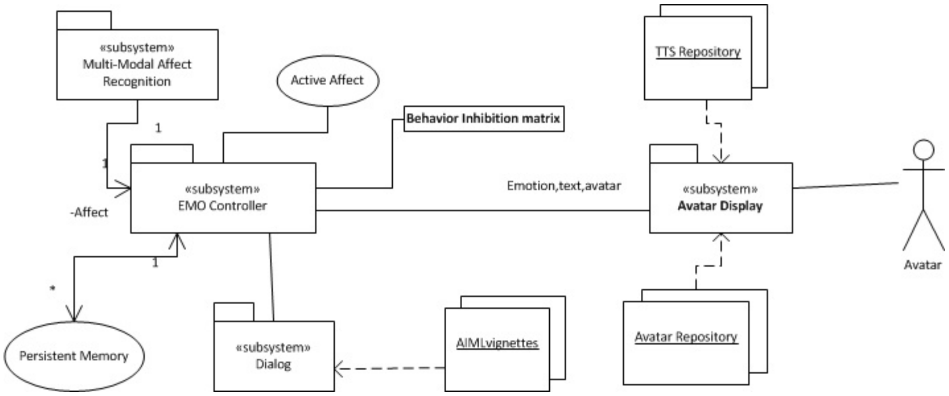
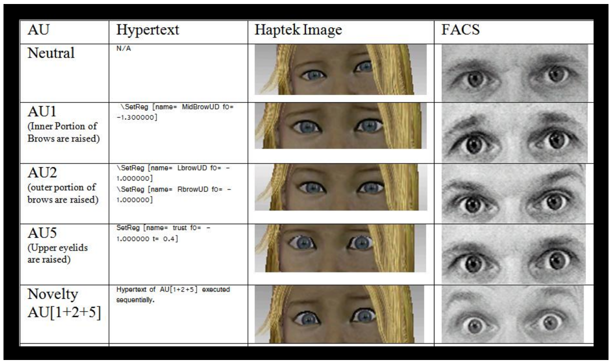


Figure 1: High Level EMO Architecture



AU	Hypertext	Haptek Image	FACS
Neutral	N/A		
AU1 (Inner Portion of Brows are raised)	\SetReg [name= MidBrowUD f0= -1.300000]		
AU2 (outer portion of brows are raised)	\SetReg [name= LbrowUD f0= -1.000000] \SetReg [name= RbrowUD f0= -1.000000]		
AU5 (Upper eyelids are raised)	SetReg [name= trust f0= -1.000000 t= 0.4]		
Novelty AU[1+2+5]	Hypertext of AU[1+2+5] executed sequentially.		

Figure 2: Facial Action Units Portrayed on an Actor and Avatar

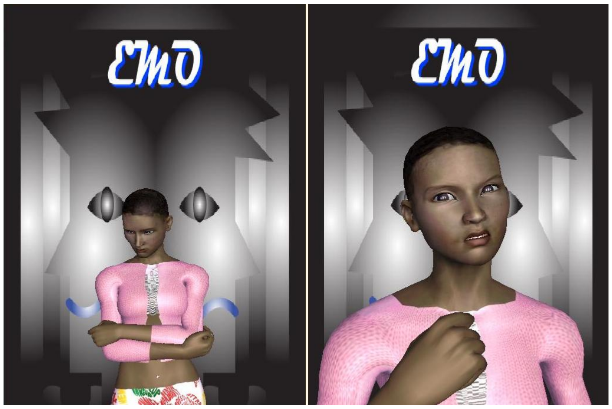


Figure 3: Using Posture and Proximity to Denote Emotion. Left sadness, Right Anger

appearance. MacDorman (MacDorman 2005) attributes the feeling of unease as an effect of our innate fear of death as the visible synthetic aspects of the embodiment may trigger an association to a moving corpse. The phenomenon is however only confirmed when the image has abnormal features and is possible to be avoided through careful modeling and testing.

Gender considerations The architecture modifies the intensity of affect portrayal based on the gender of the avatar. Gender has its role in differentiating the portrayal of affect among humans, considering males and females are exposed to the same stimuli. The sex's level of emotion suppression is determinant on the particular emotion being expressed (McRae and Gross 2008). Males tend to suppress the emotion of sadness to a greater extent than females whereas the opposite is true for anger and pride. Females may portray emotions to a greater extent than do males assuming aspects of socialization brought about by such factors as culture and societal roles.

Ethnicity To add focus to our research, we examine the differences in the expression regulation of emotions within ethnic groups of European, Asian, African and Hispanic descent. These groups were identified as they represent distinct physical and linguistic characteristics. Visual features play an important role as humans tend to categorize other individuals as being in-group or out-group, based not on extensive analysis, but easily ascertainable visual characteristics such as ethnicity; these social rules are transferred to the Human Computer Interaction and ethnically similar virtual agents are perceived as more trustworthy and socially attractive (Nass 2000).
Prior research suggests that the ethnic factor is significant in social interactions. The significance of ethnicity in science should be embraced and not treated as a nuisance factor; ethnicity has a significant role in our social interactions (Sue 1999). The depiction of ethnicity among our embodiments is accomplished through three dimensional modeling from two dimensional photography of frontal and profile images of a self identified ethnic representative model using Haptek's People Putty (see Figure 4)

Speech Synthesis and Vocal Intonation Lip synchronization between the Text To Speech (TTS) system and the avatar is handled internally by Haptek. EMO employs multiple TTS voices each mapped to a specific avatars This model not only allows for proper gender vocal representation but

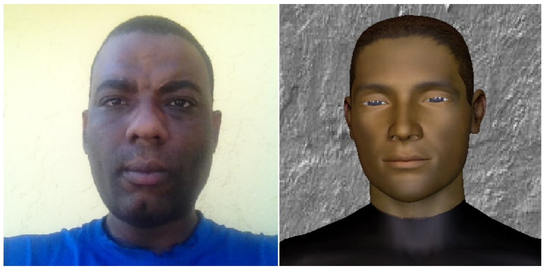


Figure 4: Animatable 3D Ethnic Avatar Rendered from 2D Image

allows for multiple languages. Currently EMO is only able to converse in English and Latin American Spanish. The Language and gender selection is automated when the user chooses an avatar.

Affect Representation

Our Affect representation lays the framework to model empathy by constraing mood and emotion transitions using a list transition system. Our array of ECAs portrays emotions and display acts generated by the systems interpretation of the users affective state through conversation, and dependent on the avatars temperament which in our case are predefined personality traits based on the five factor model of personality, namely Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism (or OCEAN). We will describe in this section how we represent the affective phenomenas, emotions, moods and personality and their interrelationships as it pertains to our model. Affect recognition is not the focus of the research at this point and is substituted by a manual driver. The following labels are octants of temperament space (PAD)(Mehrabian 1996b):

Mood	Pleasure	Arousal	Dominance
Exuberant	+	+	+
Bored	-	-	-
Dependent	+	+	-
Disdainful	-	-	+
Relaxed	+	-	+
Anxious	-	+	-
Docile	+	-	-
Hostile	-	+	+

Table 1: Octants of Temperament Space

Following the affect taxonomy adopted by (Lisetti and Gmytrasiewicz 2002), our systems personality represents long term affect, moods represent medium term affect and emotions are short term. The representation of emotions is based on a variant model. The variant model approach looks at emotions in terms of its placement in a three dimensional affect space. The view of a three dimensional affective space representation namely pleasure, arousal, and dominance (PAD) (Mehrabian 1996b). Emotions may fall anywhere within this PAD space dependent on their fundamental intensity along these three axes (See Table 1) . Emotions are therefore represented in our model as a point *(x,y,z)* in the 3-Dimensional PAD space. PAD space with each axis ranging from -1 to 1 may now be viewed as a sphere. Moods, considered medium term affect, in a PAD space representation will become the octants of the PAD sphere. Each octant corresponds to one of the eight moods namely hostility, exuberance, boredom, docile, relaxed, disdainful, anxious and dependent. The current mood is transitioned by constantly being impacted upon by new emotions. In our model, we ascertain the current mood to be a function of the previous mood, current emotion, and personality settings. In early experiments, this model experienced unnatural mood transition therefore we curtailed mood transitions employing a list transition system to determine state transitions. Currently, within our implementation, the system receives affective cues from lexical analysis of the text inputted by the user. Our affect recognition will be revisited to include additional modalities as the research progresses. We utilize (Mehrabian 1996a) to map each emotion experienced to a PAD value. This value impacts an active emotion point AEP, which floats within the 8 octants and determines EMO's moods under the constraint of the LTS. The mood itself will trigger a change in the dialog system to allow EMO to converse with the user in a different manner while in different moods. The affinity towards a particular octant and therefore mood is dependent on the personality setting of the the avatar. Personality is based on extroversion and neurotisism components of the OCEAN model. As an example, a negative extroversion coefficient would correspond to the propensity to behaviors common to individuals predisposed to introversion. The active affect is however not necessarily the displayed affect. This is the arena of the emotion regulation component. After all we wouldn't necessarily want a hostile avatar conversing with a patient. Gender, Ethnicity and Active Emotion are used to query our three dimensional matrix Θ ,which is generated at runtime based on empirical data from the study as described in (Gross and John 2003), resulting in the coefficient of inhibition, σ. The final input to the expression generation function is the intensity of the the emotion being experienced, E_0.

The expression generation function is responsible for the resulting intensity of that emotion to be displayed by the Virtual Agent. Our expression generation function is formalized as :

$E = E_0 \rho \sigma$

Whereby is the intensity of the original emotion being experienced, E is the resultant intensity of the emotion display derived from, σ , the *coefficient of inhibition* and ρ , the personality modifier described above.

Dialog

We have built our dialog system based on ALICE (Wallace 1995). ALICE is a pattern matching dialog system which expands substantially the well-known early dialog system ELIZA. ALICE comprises of two major components, an AIML (Artificial Intelligence Markup Language) knowledgebase and an interpreter. The knowledge base of ALICE comprises of dialog responses categorized and tracked. essentially the dialog database from which our avatar derives its speech. We have rewritten much of the AIML to incorporate dialog conducive to solution focused therapy. AIML uses multiple instantiations of a category as its basic unit of knowledge. Each category comprises of a pattern which corresponds to the users text input, and a template which becomes the users response. The AIML pattern tag either matches the users input literally, or may utilize a wildcard approach which will accept any instantiation of a particular syntactical structure. The template tag is multifaceted in that it may be as simple as a context appropriate text response, or it may refer to yet another category or a session memory item. Session memory items include knowledge the ECA has gained during a current session, including facts like the gender and name of the user.

The ALICE interpreter whose primary function is effective graph construction and traversal of the dataset. Since the inherent AIML pattern recognition is less rigorous than more formal regular expression parsing typical of natural language processing, the interpreter compensates with preprocessing activities such as spelling and vernacular checking. The problem of Natural Language Processing, converting human language to a formal representation is inherently complex in that ambiguities arise without the use of concepts central to an unrestricted domain. ALICE's approach is not to tackle this issue by attempting strong AI but to use careful dialog scripting of the AIML set, based on specific virtual vignettes, drawing from human domain experts to appear intelligent.

We address the specifics about the context of SFT by carefully editing the AIML scripts. This interdisciplinary stage of our research will involve working in a close interaction with psychologists, social workers and their clients to apply a participatory approach to the design of the interaction through virtual vignettes. For our implementation, in the domain of solution focussed therapy, the ALICE system has limitations which we had to address. In order to effectively communicate with the user using the appropriate affective cues, EMO needs to respond according to its moods. Accordingly, EMO utilizes multiple ALICE dialog systems which are mapped to moods, however the dialog focus and direction is singular.

We will again refer to our scenario to clarify the the augmentation process.

At Henry's initial SFT session the therapist will determine the scope of the therapy according to questions such as:

- What do you think your problem is now?
- How will you know when the problem is solved?
- How will you know when you wont have to come here any more? What will the signs be?

Subsequently a homework assignment is given at the first session. Henry is asked to observe his internal emotional state and report back to the therapist how he percieves the current situation and how he would like to proceed . At this point EMO may be used to capture thoughts and conversation through its persistent memory. This would give the therapist a broader look at the problem and solution sets. Conversing with the avatar may act as a debriefing machanism. In talking and reflecting about feelings, the therapy process is supported and enhanced .

Through crafted vignettes infused within the AIML conversation set, EMO may strive towards ascertaining the answer to the Miracle Question, as in *Suppose one night while you were sleeping there was a miracle and this problem is solved. How would you know? What would be different?* and the Scaling Question, *On a scale of one to ten, how do you feel about being able to sleep through the night?* . These questions are core components to SFT.

Evaluation

Methodology

We evaluate our display model of affect modulation within a set of avatars with ethnic traits. The implemented model was refined through beta testing with a small sampling. We then studied the effect of pairing with ethnically similar videotaped actors with avatars to gauge the plausibility of 6 distinct displays of emotions happiness, sadness, anger, surprise, disgust and fear. The sampling size of our study was 67 taken from undergraduate students. The experiment included 5 second clips posted to an online survey site. The participants ages ranged between 18 and 36 with an average age of 23.

Results

The result of the study is shown in Figure 5. We ascertained whether the studies participants were able to correctly identify the emotion being displayed, identified a different emotion or refrained altogether.

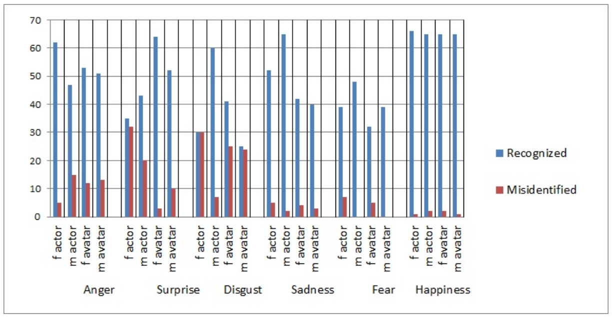


Figure 5: Survey results of avatar credibility study

Discussion

The results shows that the avatars were well recognized. All avatars were correctly identified over 50% of the time. In the case of surprise the avatars were recognized at a higher rate than the humans. This may be attributable to actors being less likely to mimic surprise after being prepared for the study.

Related Work

There are numerous approaches to a generalized system to model affect (Gebhard 2005), (Gratch and Marsella 2004). Our work is designed towards a specific domain. The works of Lisettti et al. (Lisetti and Nasoz 2002) and Iacobelli and Cassell (Iacobelli and Cassell 2007) addresses the issue of ethnicity in the design of conversational agents, our work extends these concepts by realizing an affective engine and consider ethnic emotion regulation.

Conclusion

We have presented our preliminary work towards an architecture capable effectively modeling gender and ethnicity multi-modal communication in the domain of SFT. Our avatar display survey results are promising , however we continuously revisit and reevaluate based on domain concerns. Future research directions will seek to refine the affect and dialog subsystems and evaluate their plausibility within the realm of therapy.

References

Argyle, M. 2007. *Social interaction*. Aldine.

Bianchi-Berthouze, N., and Cairns, P. 2006. On posture as a modality for expressing and recognizing emotions.

Cade, B., and O'Hanlon, W. 1993. *A brief guide to brief therapy.* WW Norton & Co.

Cowell, A. J., and Stanney, K. M. 2005. Manipulation of non-verbal interaction style and demographic embodiment to increase anthropomorphic computer character credibility. *Int. J. Hum.-Comput. Stud.* 62(2):281–306.

Coyle, D.; Doherty, G.; Matthews, M.; and Sharry, J. 2007. Computers in talk-based mental health interventions. *Interacting with Computers* 19(4):545–562.

De Shazer, S. 1985. *Keys to solution in brief therapy*. Ww Norton New York.

Doherty, G.; Coyle, D.; and Matthews, M. 2010. Design and evaluation guidelines for mental health technologies. *Interacting with computers* 22(4):243–252.

Ekman, P. 1997. *LYING AND DECEPTION*. Routledge. chapter Memory for Everyday and Emotional Events, 333–351.

Ekman, P. 1999. Facial expressions. *Handbook of cognition and emotion* 301–320.

Erdur, O.; Rude, S.; and Barón, A. 2000. Working alliance and treatment outcome in ethnically similar and dissimilar client–therapist pairings. Research Reports of The Research Consortium of Counseling & Psychological Services in Higher Ed 3, The University of Texas at Austin.

Gebhard, P. 2005. Alma: a layered model of affect. In *AAMAS '05: Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems*, 29–36. New York, NY, USA: ACM.

Gratch, J., and Marsella, S. 2004. A domain-independent framework for modeling emotion. *Journal of Cognitive Systems Research* 5:269–306.

Gross, J., and John, O. 2003. Individual differences in two emotion regulation processes: implications for affect, relationships, and well-being. *Journal of Personality and Social Psychology Bulletin* 85(2):348–362.

Iacobelli, F., and Cassell, J. 2007. *Ethnic Identity and Engagement in Embodied Conversational Agents*. Springer Berlin / Heidelberg. chapter Intelligent Virtual Agents, 57–63.

Lisetti, C. L., and Gmytrasiewicz, P. 2002. Can a rational agent afford to be affectless? a formal approach. *Applied Artificial Intelligence* 16(7-8):577–609.

Lisetti, C. L., and Nasoz, F. 2002. Maui: a multimodal affective user interface. In *MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia*, 161–170. New York, NY, USA: ACM.

Lisetti, C. 2008. Embodied conversational agents for psychotherapy. In *CHI 2008 Workshop on Technology in Mental Health, Florence, Italy*.

MacDorman, K. 2005. Androids as an experimental apparatus: Why is there an uncanny valley and can we exploit it. In *CogSci-2005 workshop: toward social mechanisms of android science*, 106–118.

McRae, K. O. K. M. I. G. J., and Gross, J. 2008. Gender differences in emotion regulation: An fmri study of cognitive reappraisal. *Group processes and Intergroup Relations* 11(2):143–162.

Mehrabian, A. 1996a. Analysis of the big-five personality factors in terms of the pad temperament model. *Australian Journal of Psychology* 48(2):86–92.

Mehrabian, A. 1996b. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology: Developmental, Learning, Personality, Social* (14):261–292.

Mori, M. 1970. *The Uncanny Valley.*, volume 7. 33–35. Translated by Karl F. MacDorman and Takashi Minato.

Nass, C. 2000. *Truth is Beauty: Researching Embodied Conversational Agents*. MIT Press. chapter Embodied Conversational Agents, 374–395.

Paleari, M., and Lisetti, C. 2006. Psychologically grounded avatars expressions. In *First Workshop on Emotion and Computing at KI*. Citeseer.

Reeves, B., and Nass, C. 1996. *The media equation: How people treat computers, television, and new media like real people and places.* Chicago, IL, US: Center for the Study of Language and Information; New York, NY, US: Cambridge University Press.

Sue, S. 1999. Science, ethnicity, and bias: Where have we gone wrong? *American Psychology* 54:1070–1077.

Wallace, R. 1995. Alice-artificial linguistic internet computer entity-the alice ai. foundation. *http\\ alicebot. org*.