NADiA - Neural Network Driven Virtual Human Conversation Agents

Jason Wu
School of Interactive Computing,
Georgia Institute of Technology
jasonwu@gatech.edu

Sayan Ghosh
Institute for Creative Technologies,
University of Southern California
sghosh@ict.usc.edu

Mathieu Chollet
Institute for Creative Technologies,
University of Southern California
mchollet@ict.usc.edu

Steven Ly
Institute for Creative Technologies,
University of Southern California
sly@ict.usc.edu

Sharon Mozgai
Institute for Creative Technologies,
University of Southern California
smozgai@post.harvard.edu

Stefan Scherer
Institute for Creative Technologies,
University of Southern California
scherer@ict.usc.edu

ABSTRACT

Advances in artificial intelligence and in particular machine learning and neural networks have given rise to a new generation of virtual assistants and chatbots. Within this work, we present *NADiA* - Neurally Animated Dialog Agent - that leverages both the user's verbal input as well as their facial expressions to respond in a meaningful way. NADiA combines a neural language model that generates appropriate responses to user prompts, a convolutional neural network for facial expression analysis, and virtual human technology that is deployed on a mobile phone. Here, we evaluate NADiA's anthropomorphic characteristics and its ability to understand the human interlocutor using both subjective as well as objective measures. We find that NADiA significantly outperforms state of the art chatbot technology and produces comparable behavior to human generated reference outputs.

KEYWORDS

Virtual Agent; Chatbot; Neural Language Model; Convolutional Neural Network; Animation

ACM Reference Format:
Jason Wu, Sayan Ghosh, Mathieu Chollet, Steven Ly, Sharon Mozgai, and Stefan Scherer. 2018. NADiA - Neural Network Driven Virtual Human Conversation Agents. In *IVA '18: International Conference on Intelligent Virtual Agents (IVA '18), November 5–8, 2018, Sydney, NSW, Australia.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3267851.3267860

1 INTRODUCTION

Conversational technologies integrated in products such as Apple's Siri, Google Home, and Amazon's Alexa have made their way into people's everyday lives. Due to advances in artificial intelligence, natural language processing, and the increased availability of cloud computing platforms, technologies, conversational agents can be effectively used for complex tasks such as cognitive behavior therapy, entertainment, and medicine.

In contrast to conversational agents that rely mostly on text or language based technologies, human face-to-face communication relies on additional communicative modalities or channels, including facial expressions, paralinguistic aspects of the voice (e.g., prosody or voice quality), as well as gestures. To accommodate this, researchers have recently focused on multimodal conversational interfaces. These interfaces known as Embodied Conversational Agents (ECA), or virtual agents [6], typically consist of anthropomorphic representations of a human and use natural communicative modalities, such as natural language and nonverbal behavior (e.g., gestures, facial expressions, postures), to interact with users.

The conversational agent architecture proposed in this work, NADiA, relies on neural network research shown to provide state-of-the-art behavior understanding, recognition, and generation. In addition, a key motivation of using neural networks for NADiA is the limited computational power needed for deployment, allowing the system to deliver high precision and state of the art performance in low resource environments such as mobile phones or embedded robotic systems. In our work, NADiA was tested and developed on a Samsung Galaxy 7 mobile phone [1]. We take full advantage of mobile phone hardware to deliver a multi-modal conversational interaction by leveraging the microphone for the automatic speech recognition and camera for the facial expression analysis and facial expression mimicry.

We combine and evaluate three main technologies, (1) a neural language model to generate meaningful responses to human user prompts, (2) a CNN (Convolutional Neural Network) to recognize and react to user's affective state, and (3) virtual agent software to deliver the responses in an anthropomorphic fashion, in order to improve the experience of the human user. We conduct a series of subjective and objective experiments to provide evidence that NADiA can improve conversational agent interactions. Specifically, we investigate three main research questions:

RQ1 - Is it possible to train a neural network based unscripted virtual character that is able to sustain brief smalltalk interactions and create the appearance that the virtual character understands its interlocutor?

[1]The stimuli for the human perception tests were rendered from the mobile phone

IVA '18, November 5–8, 2018, Sydney, NSW, Australia

ACM ISBN 978-1-4503-6013-5/18/11…$15.00
https://doi.org/10.1145/3267851.3267860

RQ2 - How anthropomorphic is the appearance of the virtual characters to human judges, and what is the influence of the model that generates the responses?

RQ3 - What additional benefit is gained from the facial expression mimicry of the virtual character that is enabled through an end-to-end convolutional neural network approach?

The remainder of the paper is organized as follows: First, Section 2 introduces some of the related work with respect to natural language generation for conversational agents as well as the impact of behavior mimicry of agents on their human interlocutors. Then, Section 3 introduces and describes the main components of the NADiA architecture that is deployed on a mobile device. Section 4 details the experimental setup of our subjective and objective evaluations, and Section 5 details the main findings of our experiments. Section 6 discusses the main findings with respect to our research questions **RQ1-3**, and lastly Section 7 concludes the paper.

2 RELATED WORK

2.1 Natural Language Generation for Conversational Agents

Early research in creating believable conversational agents were largely motivated by the *Imitation Game* (also known as the Turing Test), proposed by Alan Turing in 1950. To maintain grammatical correctness and produce logical responses, early conversational agents, known as chatterbots, relied mostly on programmer-defined pattern matching, conversational networks, and activation networks, which sometimes were able to obtain limited success in restricted Turing Test evaluations [22]. ALICE (Artificial Linguistic Internet Computer Entity) introduced a new markup language for programmer-defined conversation knowledge and responses called AIML (Artifical Intelligence Markup Language) supporting XML definitions of categories, patterns, and templates [28]. Current commercially-available chatbot software such as *Cleverbot* utilize similar approaches to rule-based conversation generation, augmented with mechanisms for learning new response patterns from conversations.

More recently, the focus of conversational research has shifted away from solely generating convincing dialog and towards the creation of functional natural language interfaces and conversational modeling.

Recent work has shown that neural language models and recurrent sequence-to-sequence models are able to encode limited conversational context a viable approach to language modeling with a large, unstructured corpus [27].

To facilitate more realistic responses, other cues such as affect and contextual understanding can be used during response generation. Conversational models that encode affective signals such as user satisfaction and emotional state have shown to be effective for more believable conversational agents and low-perplexity language models [12, 14].

2.2 Virtual Agent Models and Facial Mimicry

Virtual agents that can take on many roles have been used in many application domains and for many purposes, such as promoting healthy exercise in older adults [4], rehearsing job interviews with a virtual recruiter [8] or screening for depressive disorders [10]. Endowing virtual agents with certain behavioral capabilities can allow them to be more engaging and likable. One of these abilities that was investigated in the context of virtual agents is mimicry which can increase liking and rapport between individuals.

Experimental studies on the perception of emotion-mimicking virtual agents found that participants reported greater amounts of positivity, warmth, and realism in the presence of smiles while feeling more at ease and well-understood [3, 19, 25].

While prior work showed the positive effects of emotional mirroring virtual agents, these existing models of virtual agents' facial mimicry relied on facial recognition and facial animation technologies with a large delay between the processing of the visual feed, facial behavior recognition, and production of facial mimicry [3, 19]. By leveraging neural networks, we aim to prioritize fast inference speed to support real-time or near real-time feedback.

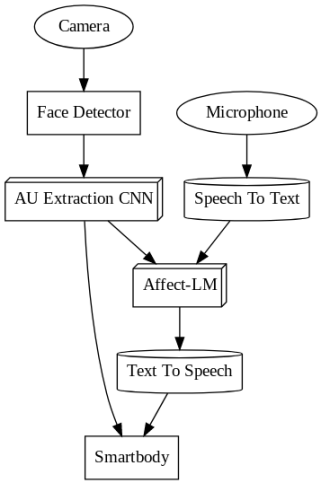

Figure 1: NADiA Architecture Diagram

3 NADIA ARCHITECTURE

The NADiA architecture consists of three main parts: (1) NADiA's ability to generate natural language is based on a novel neural language model named Affect-LM [14]. (2) NADiA is further capable of mimicking the human user's facial expressions. This capability is enabled through a convolutional neural network that detects the user's face and analyzes his/her facial expressions and renders the same expression on NADiA's virtual face. (3) The appearance of NADiA is enabled through the use of the Smartbody architecture [24]. Overall NADiA is deployed on a mobile phone and is able to respond to human user prompts in near realtime. An overview of the architecture is provided in Figure 1 the following section describes each component in detail.

3.1 Affect-LM Text Generation Model

To generate NADiA's responses to human user's prompts, we leverage a novel language model *Affect-LM* [14]. Affect-LM is capable of generating affective conversational text by inferring the affective context from the conversation history.

Affect-LM was trained on a large conversational corpus, namely the Fisher dataset [9]. This dataset consists of speech from dyadic telephonic conversations of 10 minutes each, along with their associated transcripts. We leverage this corpus both for training of neural language model Affect-LM, as well as the evaluation of NADiA.

3.2 Facial Mimicry CNN

The facial expression CNN extracts the activations of 18 Action Units (AUs) as defined by the Facial Action Coding System (FACS) from the front-facing smartphone camera. These AUs can be used to infer the user's affective state, serve as an input parameter to Affect-LM, and provide facial mimicry.

A histogram-of-oriented-gradients (HOG) object detector is used to provide cropped facial images to the CNN [18].

The facial expression CNN was trained using two freely available datasets of video and multi-media content, the UvA-NEMO Smile Database and the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [11] [20], which were both labeled using an open source facial behavior analysis tookit [1].

The CNN architecture consisted of three layers, each consisting of a two-dimensional convolutional layer, and a stage of max pooling with a ReLU (Rectified Linear Unit) activation layer. Similar to [15], a multi-label cross entropy loss was used for training, along with an RMSProp optimizer [26] and for 200 epochs. This configuration is similar to several state-of-the-art neural network architectures for object recognition and emotion/AU detection [16].

3.3 Smartbody Mobile Integration

Smartbody is an open-source project written in portable C++ and is usable on many different platforms, including Android. The behavior generation commands for NADiA are Smartbody scripts that communicate via Behavior Markup Language (BML), a language for describing verbal and non-verbal character animation behaviors.

Both neural network architectures communicate with the Smartbody library to animate the character's speech and expression. NADiA runs on a Samsung Galaxy S7 device with a Qualcomm Snapdragon 820 64-bit quad-core CPU and 4GB of RAM. The facial mimicry inference takes a total of around 200-300 milliseconds.

4 EXPERIMENTAL SETUP

In this section, we describe the experimental setup of our study. Our study was constructed using a 3x3 study design that comprises two independent variables (IVs) that are evaluated as our main effects.

The first IV is the source of the dialog (i.e., *Affect-LM*, *Cleverbot*, and *Fisher Reference*). *Affect-LM* represents our proposed neural language model, *Cleverbot* is a commonly used online accessible chatbot[2], and *Fisher Reference* refers to the actual conversation as it was recorded for the Fisher dataset. The second IV represents the type of appearance used for stimuli presentation (i.e., *Audio Only*, *No Mirroring*, and *Mirroring*). In the *Audio Only* condition the dialogs are presented using a static visual representation of NADiA (i.e., a screenshot of the virtual character). For both *No Mirroring* and *Mirroring* conditions NADiA is presented as a video stimulus and NADiA's lips are moving while she is speaking. For the *Mirroring* condition NADiA additionally attempts to mirror the facial expressions of the human interlocutor using the CNN described in Section 3.2.

We evaluate our research questions **RQ1-3**, introduced in Section 1, using both subjective evaluations enabled through an extensive study on the crowd-sourcing platform Amazon Mechanical Turk as well as the BLEU score, which has previously been used to evaluate similarity between human and artificial machine translation.

4.1 Subjective Evaluation - Perception Study

Our evaluation aimed at assessing our virtual agent model on several aspects. The following categories, hence, serve as our dependent variable (DVs): (a) how realistic or believable the virtual agent's behavior seems to subjects, both in terms of the generated text and facial expressions; (b) how enjoyable subjects think it would be to interact with the virtual agent; and (c) how generally easy to use the virtual agent seems to be. To that end, we re-used the evaluation scheme proposed by Lisetti *et al.* [19], which is based on a combination of scales from Heerink's model [17] and Bartneck's "Godspeed questionnaire" [2]. Specifically, these scales were originally designed for evaluating robots and artificial agents on a number of dimensions, including anthropomorphism, perceived enjoyment and perceived ease of use. In addition to these questions, we seek to evaluate how well the system can understand the human interlocutor and added one additional item to the perception study that evaluates the subjective perception of understanding between the human and artificial interlocutors. Every item was evaluated on a 7 point Likert-scale.

For our perception study, we leverage Amazon's Mechanical Turk (MTurk) platform. The MTurk platform has been successfully used in the past for a wide range of perception experiments and has been shown to be an excellent resource to collect human ratings for large studies [5]. Each stimuli was evaluated by eight human raters that have a minimum approval rating of 98% and are located in the United States. The human raters were instructed that the conversations should be considered to be taken from a conversational rather than a written context: repetitions and pause fillers (e.g., *um, uh*) are common and no punctuation is provided. The human raters were paid 0.30USD per stimulus. As each stimulus was about 45 seconds in length and required evaluating 8 items using a Likert scale items. The average evaluation time was expected to be around 2 minutes. We observed that a small number of raters (N = 6) took more than 10 minutes to evaluate the stimuli and we considered them as distracted and hence removed them from the subsequent analyses. We kept the MTurk study active for five days and received a total of 330 responses. After removing the distracted raters, we have access to 324 valid ratings.

We conducted initial ANOVAs, with human ratings for the perception scales as our DVs and dialog source and type of appearance as the two IVs. When we observe significant main effects in the ANOVAs, we conducted follow-up t-tests to identify which conditions are responsible for the observed effects.

4.1.1 Stimuli Generation. Specifically, we generated 45 stimuli[3] to form a comprehensive dataset for the tested IVs. These IVs are

[2] http://www.cleverbot.com/

[3] Links for the generated stimuli are available at: https://goo.gl/JHcm8P

the source of dialog and type of appearance used for stimuli presentation. For each configuration, 5 stimuli are generated using a set of corresponding starting prompts.

(1) *What do you think is the most important thing to look for in a life partner?*
(2) *Would you commit perjury for a friend or family member?*
(3) *What do you think about computers in education?*
(4) *What is your favorite holiday?*
(5) *Do you like to cook?*

To ensure that there are adequate examples of reference responses, these starting prompts are chosen according to the Fisher corpus description file, containing a list of topics of conversation in the dataset. While the fact that the conversation topics were taken from the Fisher dataset may give Affect-LM (trained on the Fisher dataset) an implicit advantage over Cleverbot, the topics were purposely chosen to be as generic as possible to maximize the probability that it also exists in Cleverbot's response database.

The 5 starting prompts are used to provide consistent starting conditions for the 3 sources of dialog. Following the initial prompt, the interaction is allowed to continue naturally between a human and the dialog source for a maximum of 8 dialog turns. The conversation between the human and each chatbot is recorded in individual transcript files.

The transcripts are used to recreate the conversation dialog for the 3 types of appearance, which include both multi-modal and audio-only stimuli. The multi-modal stimuli are generated by capturing the video and audio output of the NADiA conversation application running on a mobile phone while the audio-only stimuli are generated by combining the audio output with a screenshot of the NADiA virtual human. The virtual human's facial mimicry is controlled by enabling or disabling the facial expression CNN. While participants are not able to fully appreciated the facial mimicry aspect of the conversational interaction, we posit that the human interlocutor's emotional expression is reasonable representation of the conversational responses, and the relatively short duration of the conversation limits the variance of possible affect states. The application supports dynamic response generation using various dialog sources, but the virtual human's responses are manually set according to the pre-generated transcript responses during stimuli generation for the purposes of consistency and reproducibility due to the non-deterministic nature of the chatbots.

4.2 Objective Evaluation - BLEU Score Evaluation

To complement the subjective evaluations (cf. Section 4.1) we evaluate the similarity between the automatically generated responses of Cleverbot and Affect-LM and the reference responses gathered from the actual conversations that were recorded in the Fisher dataset. Due to the nature of the chosen conversation topics, the reference responses are generic and represent common responses in the course of natural conversation. In particular, we leverage the BLEU score that is traditionally used to assess the quality of machine translation for this purpose [23]. As BLEU was originally designed for document-level translation, smoothing function 1 described by Chen and Cherry is applied during evaluation [7].

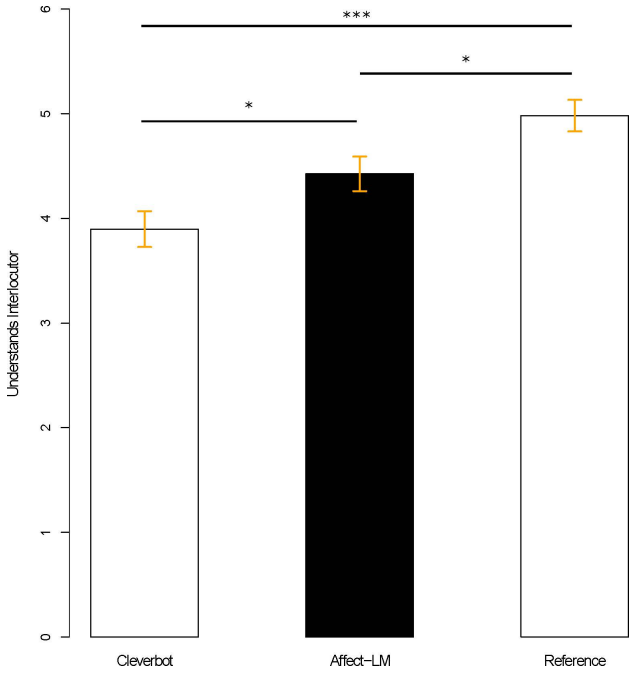

Figure 2: Human rater scores on seven-point Likert-scale with respect to whether the artificial interlocutor understands the human speaker. Significant differences as identified by follow-up t-tests is indicated with * for $p < 0.001$ and * for $p < 0.05$.**

We conducted an unpaired t-test to evaluate if the generated responses of Affect-LM or Cleverbot were closer to reference responses observed in the Fisher dataset.

5 RESULTS

5.1 RQ1 - Understands Interlocutor

For our first research question **RQ1**, we investigated whether the human raters evaluate NADiA's ability to understand the human interlocutor differently from the reference conversation that was recorded in the Fisher dataset and the Cleverbot baseline. The repeated measures ANOVA revealed a significant main effect for source of dialog (i.e., Affect-LM vs. Cleverbot vs. Fisher Reference; $F(2, 36) = 14.682$, $p < 0.001$, $\eta^2 = 0.449$). As expected, no significant effect for media type (i.e., Audio only vs. No Mirroring vs. Mirroring) was observed. Neither did we observe a significant interaction between the two IVs. Follow-up t-tests revealed that Fisher Reference ($M = 4.980$, $SD = 0.53$) significantly outperformed both Affect-LM ($M = 4.40$, $SD = 0.282$; $t(200) = 2.525$, $p = 0.01$) and Cleverbot ($M = 3.80$, $SD = 0.411$; $t(197) = 4.828$, $p < 0.001$). In addition, we observed a significant difference between Affect-LM and Cleverbot ($t(197) = -2.244$, $p = 0.026$). Figure 2 summarizes the observed differences in perceived understanding.

Complementary to the perceptual experiments on MTurk, we conducted an objective evaluation of the similarity between the Fisher Reference sentences and the dialog sources Affect-LM and Cleverbot. For this purpose we leveraged BLEU score [23]. The

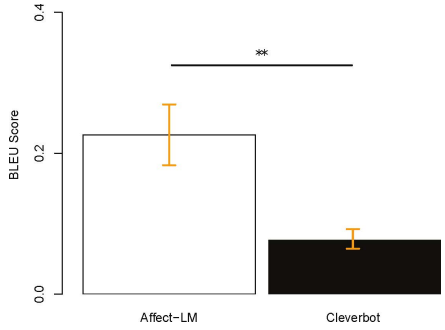

Figure 3: BLEU score comparison between dialog sources Affect-LM and Cleverbot. Both sources are compared to the Fisher Reference. Higher scores reflect increased similarity with Fisher Reference. Significant differences as identified with pairwise t-test is indicated with ** for p < 0.01.

t-test revealed that there is a significant difference between the BLEU score of Affect-LM (M = 0.226 , SD = 0.236) and Cleverbot (M = 0.077, SD = 0.076; t(29) = 3.277, p < 0.01), which signifies that the generated responses of Affect-LM were significantly closer to the Fisher Reference than those generated by Cleverbot. The result is summarized in Figure 3.

5.2 RQ2 - Perceived Anthropomorphism

With respect to our second research question **RQ2** we observe the following: The repeated measures ANOVA revealed both a main effect for dialog source (F(2, 36) = 3.314, p < 0.05, η^2 = 0.155) and a main effect for media (F(2, 36) = 5.877, η^2 = 0.246). No significant interaction between the IVs was observed. Follow-up t-tests for dialog source, revealed that Fisher Reference (M = 3.877, SD = 0.283) was perceived as significantly more anthropomorphic than Cleverbot (M = 3.362, SD = 0.346; t(202) = 2.01, p < 0.05). There was no significant difference observed between Fisher Reference and Affect-LM (M = 3.425, SD = 0.233; p > 0.05), nor between Affect-LM and Cleverbot.

With respect to the type of media we observed a significant difference between the Audio Only (M = 15.636, SD = 1.709) and both Mirroring (M = 18.443, SD = 0.352; t(202) = -2.223, p < 0.05) as well as No Mirroring (M = 19.284, SD = 1.826; t(199) = -2.880, p < 0.01) conditions. No significant difference was observed between Mirroring and No Mirroring.

5.3 RQ3 - Pleasant Conversation Partner

We evaluate the perceived pleasantness of the conversation partner (textbfRQ3) by evaluating two ratings. First, the repeated measures ANOVA for the perceived pleasantness of the conversation partner revealed a main effect of dialog source (F(2, 36) = 4.355, p < 0.05, η^2 = 0.195). No other effects were observed. Follow-up t-tests reveal that there is a significant difference in perceived pleasantness between the Fisher Reference (M = 4.720, SD = 0.358) and Cleverbot (M = 4.098, SD = 0.269; t(196) = 2.453, p < 0.05). No significant difference was observed between Affect-LM (M = 4.271, SD = 0.214) and either Fisher Reference or Cleverbot.

Second, the repeated measures ANOVA for the "nice" characteristic of the conversation partner revealed a main effect for type of media (F(2, 36) = 3.430, p < 0.05, η^2 = 0.160). No other effects were observed. Follow-up t-tests revealed that the Mirroring condition (M = 5.128, SD = 0.157) was perceived significantly *nicer* than the Audio Only condition (M = 4.667, SD = 0.088; t(196) = -2.26, p < 0.05). No significant differences were observed with respect to the No Mirroring condition (M = 4.878, SD = 0.272).

6 DISCUSSION

Here, we summarize and discuss the results reported in Section 5 with respect to our research questions **RQ1-3**.

6.1 RQ1 - Understands Interlocutor

Out first research question is concerned with the ability of the conversational agent to sustain believable human conversation. We observe (cf. Section 5.1) that the dialog source Affect-LM was perceived to be significantly more understanding than Cleverbot. As expected, both artificial approaches are outperformed by the reference human response in the Fisher corpus, albeit only by one point on the Likert-scale in the case of Affect-LM. It is expected that when evaluated in longer conversations, Affect-LM's perceived realism will suffer due to its inability to store long-term context and generate consistent responses (due to softmax sampling).

The encouraging results from the subjective evaluation are further supported by a significantly higher BLEU score of Affect-LM over Cleverbot when compared to the Fisher Reference. In addition to providing a considerably higher similarity in responses, the observed BLEU score of $\approx$ 0.22 further serves as a sanity check that Affect-LM *did not* overfit on the Fisher dataset and produces meaningful responses that are independent of the training data[4].

As expected, the type of appearance (i.e., Audio only, No Mirroring, Mirroring) had no influence on the perceived capabilities of understanding the human interlocutor.

6.2 RQ2 - Perceived Anthropomorphism

Our second research question investigates the perceived anthropomorphism of NADiA with respect to the dialog source and conversational agent's appearance. To evaluate this effect, we conducted a perceptual study on MTurk using a five item scale of anthropomorphism [19]. We combined the five items of this scale to get at an overall score of anthropomorphism (i.e., the average over the items). As reported in Section 5.2, we found a significant main effect for both dialog source as well as type of appearance.

With respect to dialog source, we learned that this effect was mainly driven by the perceived difference in anthropomorphism between Cleverbot and the Fisher Reference. There was no significant difference between Fisher Reference and the proposed neural language model Affect-LM. This observation is encouraging and further supports the results discussed in Section 6.1.

As for type of appearance, the perception study conducted on MTurk reveals that, as expected, the animated versions of the virtual character (i.e., No Mirroring and Mirroring) appear significantly more anthropomorphic than the Audio Only condition. There was

[4]Please refer to https://goo.gl/JHcm8P to watch/listen to the actual stimuli that we generated for the MTurk study.

no significant difference between the No Mirroring and Mirroring conditions, indicating that our approach of using convolutional neural networks to directly mimic the human user did not improve the perceived anthropomorphism. In the future, we seek to investigate more sophisticated neural network based listening behavior generation [13] as well as manipulate the facial expressions of the virtual character to better match the affective content of the generated utterance [21].

6.3 RQ3 - Pleasant Conversation Partner

Our third research question pertains to the effect of both dialog source and type of appearance on the perceived pleasantness of the conversation partner. Our analyses comprised two separate items within our human perception study: (1) pleasantness of conversation and (2) perceived niceness of the artificial conversation partner.

With respect to pleasantness, we observed a main effect for dialog source, but not for the type of appearance. We observe that this effect is mainly driven by the increased pleasantness for the Fisher Reference. Fisher Reference is significantly rated as more pleasant than Cleverbot. There are no other significant results observed.

The observed results for the perceived niceness of the artificial conversation partner confirms our hypothesis. We observe that Mirroring is perceived as significantly nicer than the Audio only condition. However, there is no significant difference between Mirroring and No Mirroring conditions.

7 CONCLUSION

Within this work we evaluated our Neurally Animated Dialog Agent NADiA and compared its performance to both a human reference and a state of the art baseline both using subjective as well as objective evaluation criteria. We identified (**RQ1**) how well NADiA can understand its human interlocutor, (**RQ2**) how anthrompomorphic NADiA is perceived, and (**RQ3**) how pleasant the conversation with NADiA is perceived. For all three research questions we found encouraging results and could show that the here proposed NADiA architecture has potential to act as an enjoyable and empathic conversation partner. For future work we seek to improve its long-term memory capabilities by complementing the neural language generation module with a dedicated memory network and seek to improve its listening behavior by training a neural network that does not simply mimic the facial expressions of the human user. Overall, we believe that artificial conversational agents still have a long way to go to replace interpersonal human contact. However, we show that artificial neural architectures have the ability to uphold the illusion of understanding and some anthropomorphic characteristics during brief conversations.

REFERENCES

[1] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.

[2] Christoph Bartneck, Elizabeth Croft, and Dana Kulic. 2008. Measuring the anthropomorphism, animacy, likeability, perceived intelligence and perceived safety of robots. In *Metrics for HRI workshop, technical report*, Vol. 471. 37–44.

[3] Elisabetta Bevacqua, Sylwia Julia Hyniewska, and Catherine Pelachaud. 2010. Evaluation of a virtual listenerâĂŹs smiling behavior. In *Proceedings of the 23rd International Conference on Computer Animation and Social Agents, Saint-Malo, France*.

[4] Timothy W Bickmore and Rosalind W Picard. 2005. Establishing and maintaining long-term human-computer relationships. *ACM Transactions on Computer-Human Interaction (TOCHI)* 12, 2 (2005), 293–327.

[5] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling. 2011. Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on psychological science* 6, 1 (2011), 3–5.

[6] Justine Cassell. 2000. *Embodied conversational agents*. MIT press.

[7] Boxing Chen and Colin Cherry. 2014. A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU.. In *WMT@ ACL*. 362–367.

[8] Mathieu Chollet, Magalie Ochs, Chloé Clavel, and Catherine Pelachaud. 2013. A multimodal corpus approach to the design of virtual recruiters. In *Affective Computing and Intelligent Interaction (ACII), 2013 Humaine Association Conference on*. IEEE, 19–24.

[9] Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text.. In *LREC*, Vol. 4. 69–71.

[10] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, et al. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[11] Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. 2012. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *European Conference on Computer Vision*. Springer, 525–538.

[12] Hao Fang, Hao Cheng, Elizabeth Clark, Ariel Holtzman, Maarten Sap, Mari Ostendorf, Yejin Choi, and Noah A Smith. 2017. Sounding Board–University of WashingtonâĂŹs Alexa Prize Submission. *Alexa Prize Proceedings* (2017).

[13] Will Feng, Anitha Kannan, Georgia Gkioxari, and C Lawrence Zitnick. [n. d.]. Learn2Smile: Learning Non-Verbal Interaction Through Observation. ([n. d.]).

[14] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. *arXiv preprint arXiv:1704.06851* (2017).

[15] Sayan Ghosh, Eugene Laksana, Stefan Scherer, and Louis-Philippe Morency. 2015. A multi-label convolutional neural network approach to cross-domain action unit detection. In *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 609–615.

[16] Shizhong Han, Zibo Meng, Ahmed-Shehab Khan, and Yan Tong. 2016. Incremental Boosting Convolutional Neural Network for Facial Action Unit Recognition. In *Advances in Neural Information Processing Systems*. 109–117.

[17] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2009. Measuring acceptance of an assistive social robot: a suggested toolkit. In *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*. IEEE, 528–533.

[18] Davis E. King. 2009. Dlib-ml: A Machine Learning Toolkit. *Journal of Machine Learning Research* 10 (2009), 1755–1758.

[19] Christine Lisetti, Reza Amini, Ugan Yasavur, and Naphtali Rishe. 2013. I can help you change! an empathic virtual agent delivers behavior change health interventions. *ACM Transactions on Management Information Systems (TMIS)* 4, 4 (2013), 19.

[20] Steven R Livingstone, Katlyn Peck, and Frank A Russo. 2012. Ravdess: The ryerson audio-visual database of emotional speech and song. In *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Science*.

[21] Stacy Marsella, Yuyu Xu, Margaux Lhommet, Andrew Feng, Stefan Scherer, and Ari Shapiro. 2013. Virtual character performance from speech. In *Proceedings of the 12th ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. ACM, 25–35.

[22] Michael L. Mauldin. 1994. Chatterbots, Tinymuds, and the Turing Test Entering the Loebner Prize Competition. In *Proceedings of the Twelfth AAAI National Conference on Artificial Intelligence*. 16–21.

[23] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 311–318.

[24] Ari Shapiro. 2011. Building a character animation system. *Motion in Games* (2011), 98–109.

[25] Catherine J Stevens, Bronwyn Pinchbeck, Trent Lewis, Martin Luerssen, Darius Pfitzner, David MW Powers, Arman Abrahamyan, Yvonne Leung, and Guillaume Gibert. 2016. Mimicry and expressiveness of an ECA in human-agent interaction: familiarity breeds content! *Computational cognitive science* 2, 1 (2016), 1.

[26] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.

[27] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[28] Richard S Wallace. 2009. The anatomy of ALICE. In *Parsing the Turing Test*. Springer, 181–210.