Zara Returns: Improved Personality Induction and Adaptation by an Empathetic Virtual Agent

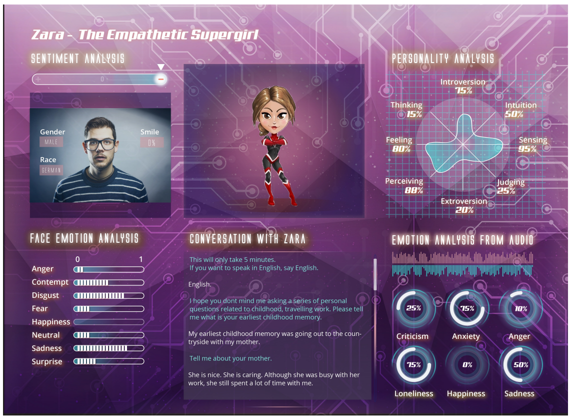
Farhad Bin Siddique†, **Onno Kampman**†, **Yang Yang**†, **Anik Dey**†*, **Pascale Fung**†*
†Human Language Technology Center
Department of Electronic and Computer Engineering
Hong Kong University of Science and Technology, Hong Kong
*EMOS Technologies Inc.
pascale@ece.ust.hk,
[fsiddique, opkampman, yyangag, adey]@connect.ust.hk

Abstract

Virtual agents need to adapt their personality to the user in order to become more empathetic. To this end, we developed Zara the Supergirl, an interactive empathetic agent, using a modular approach. In this paper, we describe the enhanced personality module with improved recognition from speech and text using deep learning frameworks. From raw audio, an average F-score of 69.6 was obtained from real-time personality assessment using a Convolutional Neural Network (CNN) model. From text, we improved personality recognition results with a CNN model on top of pre-trained word embeddings and obtained an average F-score of 71.0. Results from our Human-Agent Interaction study confirmed our assumption that people have different agent personality preferences. We use insights from this study to adapt our agent to user personality.

1 Introduction

According to Miner et. al's research (Miner et al., 2016), various world renowned virtual assistants respond inconsistently and impersonally to affect-sensitive topics such as mental health, domestic violence, and emergencies. This situation calls for a need in empathy and adaptive personality in the virtual agents (VA). In human-human interactions (HHI), personality compatibility is important for relationship satisfaction and interpersonal closeness (Long and Martin, 2000; Berry et al., 2000). With the rising number of interactive products in the market these days, it is important to emphasize machine adaptability to different users of varying needs. Here we propose an adaptive personality module that can be embedded in interactive dialog systems. In our new version of Zara, we start with recognizing user personality through speech and text based on the Big Five traits of personality (Gosling et al., 2003; Barrick and Mount, 1991). Our module includes two separate Convolutional Neural Network (CNN) models, one with real-time raw audio analysis and the other trained on text with pre-trained word embeddings. To justify our concept of adaptive personality, we did a user study after developing three virtual agents (two with distinct personalities and a robotic version as a control).



This will only take 5 minutes.
If you want to speak in English, say English.

English.

I hope you dont mind me asking a series of personal questions related to childhood, travelling work. Please tell me what is your earliest childhood memory.

My earliest childhood memory was going out to the countryside with my mother.

Tell me about your mother.

She is nice. She is caring. Although she was busy with her work, she still spent a lot of time with me.

Figure 1: System UI design

2 System Description

We are building our current work on top of our previous interactive system of Zara the Supergirl (Fung et al., 2015), with an updated user interface (see Figure 1). Zara assesses user personality based on the answers users provide at each turn. The entire conversation lasts around five minutes. The dialog is system-initiated when the user's face is detected. Zara asks a series of personal questions with increasing intimacy, including topics like childhood memory, travel, work-life, and affiliation towards human-agent interactions. At every

Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations, pages 121–126
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-4021

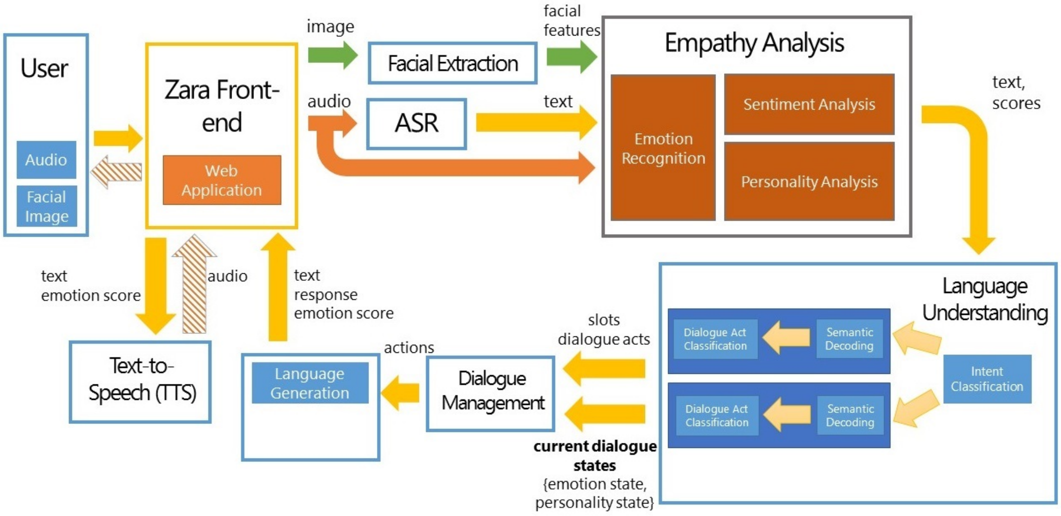


Figure 2: System architecture

utterance, the system gives an empathetic response based on the user's sentiment and emotion scores (Bertero et al., 2016; Fung et al., 2016). A separate module handles user-initiated queries, which chats with the user until the user desires to go back and finish the test. Abusive language - such as swearing, explicit sexist, or racist remarks - are also handled by a separate module. A simplified system architecture diagram is given in Figure 2. Zara is a web application, run on a server, and can be rendered on a browser.

3 Big Five Personality Induction

Personality is the complex pattern of individual differences, behavior, and thinking style. It is predominantly described with the Big Five model of personality (Goldberg, 1993), which defines five personality traits and rates a person along them. The traits are (abbreviations used in this paper are bolded): **Extra**version vs. introversion, **Agre**eableness vs. detachment, **Cons**cientiousness vs. carelessness, **Neur**oticism vs. emotional stability, and **Open**ness to Experience vs. cautiousness. Big Five personalities can be determined by self-reported assessments, such as the NEO-PI-R (Costa and McCrae, 2008). An automatic system of personality recognition is useful for situations where this is not practical.

Early work on automatic personality recognition was mostly concerned with text and audio-visual features. Argamon et al. focused on text-based personality recognition (Argamon et al., 2005), and used four sets of lexical features to determine a document's author's Extraversion and Neuroticism levels. In 2007, Mairesse et al. extracted lexical features from text, and prosodic features from speech clips (Mairesse et al., 2007). These were then mapped to the Big Five traits using different algorithms. Researchers using prosodic features commonly extract them using a pre-existing toolkit such as openSMILE (Eyben et al., 2010), after which they use classifiers such as SVMs to classify personality traits. More recently, in the ChaLearn Looking at People workshop of 2016, Gucluturk et al. experimented with a CNN approach on video snapshots and raw audio to train their model (Güçlütürk et al., 2016).

3.1 Personality perception from raw audio

We propose a method for automatically detecting someone's personality without the need for complex feature extraction upfront, as in (Mohammadi and Vinciarelli, 2012). This speeds up the computation, which is essential for dialog systems. Raw audio is inserted straight into a CNN. These architectures have been applied very successfully in speech recognition tasks recently (Palaz et al., 2015).

Our CNN architecture is shown in Figure 3. The audio input has sampling rate 8 kHz. The first convolutional layer is applied directly on a raw audio sample $\mathbf{x}$:

$$\mathbf{x}_i^{\mathrm{C}} = \mathrm{ReLU}(\mathbf{W}_{\mathrm{C}}\mathbf{x}_{[i,i+v]} + \mathbf{b}_{\mathrm{C}}) \qquad (1)$$

where v is the convolution window size. We apply a window size of 25ms and move the convolution window with a step of 2.5ms. The layer

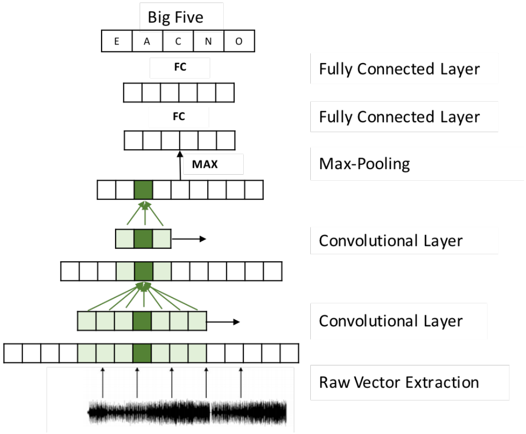


Figure 3: CNN to extract personality features from raw audio, mapped to Big Five personality traits

uses 15 filters. It essentially makes a feature selection among neighbouring frames. The second convolutional layer (identical to the first, but with a window size of 12.5ms) captures the differences between neighbouring frames again, and a global max-pooling layer selects the most salient features among the entire input speech sample and combines them into a fixed-size vector. Two fully connected rectified-linear layers and a final sigmoid layer output the predicted scores of each of the five personality traits.

The dataset trained on is the ChaLearn First Impressions dataset (Ponce-López et al., 2016). The dataset consists of 10,000 clips (15 seconds each) of YouTube bloggers talking about diverse topics. Each clip is annotated with the Big Five personality traits by Amazon Mechanical Turk workers. The dataset used a pre-defined split of a Training set of 6,000 videos, a Validation set of 2,000 videos, and a Test set of 2,000 videos. In our experiment we use the Training set for training (using cross-validation), and the Validation set for testing performance.

We implement our model using Tensorflow on a GPU setting. The model is iteratively trained to minimize the Mean Squared Error (MSE) between trait predictions and corresponding training set ground truths, using Adam (Kingma and Ba, 2014) as optimizer and Dropout (Srivastava et al., 2014) in between the two fully connected layers to prevent model overfitting.

3.2 Personality induction from text

CNNs have also gained popularity recently in efficiently carrying out the task of text classification (Kalchbrenner et al., 2014; Kim, 2014). In particular, using pre-trained word embeddings like word2vec (Mikolov et al., 2013) to represent the text has proved to be useful in classifying text from different domains. We have trained a model that takes as input text represented with the pre-trained word embeddings, followed by a single CNN layer, and then a max pooling, and a fully connected layer with softmax at the end to map the features to the binary classification task.

The convolution window sizes of 3, 4 and 5 were used in order to represent the n-gram features, and a total of 128 filters were trained in parallel. For regularization, the L2 regularization with lambda 0.01, and Dropout of 0.5 was used. We used rectified linear (ReLu) to add non-linearity, and the Adam optimizer was used for the training update at each step.

The datasets used for the training are taken from the Workshop on Computational Personality Recognition (WCPR) (Kosinski and Stillwell, 2012; Celli et al., 2014). The Facebook and Youtube personality datasets were combined and used for training. The Facebook dataset consists of status updates taken from 250 users with their personality labeled via a form the users filled up, and the Youtube dataset has 404 different transcriptions of vloggers which are labeled via perceived personality with the help of mechanical turk. A median split of the scores is done to divide each of the big five personality groups into two classes, and so the task was five different binary classifications, one for each trait.

We trained a SVM classifier using LIWC (Pennebaker et al., 2001) lexical features to treat as baseline (Verhoeven et al., 2013) in order to compare with our own model.

3.3 Results

For personality perception from audio, our model outputs a continuous score between 0 and 1 for each of the five traits for any given sample. We evaluate its performance by turning the continuous labels and outputs into binary classes using median splits. Our classification performance is good when comparing, for instance, to the winner of the 2012 INTERSPEECH Speaker Trait sub-Challenge on Personality (Ivanov and Chen, 2012; Schuller et al., 2012). Table 1 shows the model performance on the ChaLearn Validation set for this 2-class problem. The average of the mean ab-

%	Average	Extr	Agre	Cons	Neur	Open
Accuracy	62.3	63.2	61.5	60.1	64.2	62.5
Precision	60.6	60.5	60.6	58.4	62.7	60.8
Recall	81.8	83.7	83.2	86.3	78.3	77.6
$F - Score$	69.6	70.2	70.1	69.6	69.7	68.2

Table 1: Classification performance on ChaLearn Validation dataset using CNN on raw audio

	Average	Extr	Agre	Cons	Neur	Open
CNN model	**71.0**	**70.8**	**72.7**	**70.8**	**72.9**	**67.9**
Baseline SVM	59.4	59.6	57.7	60.1	63.4	56.0

Table 2: F-Score results of CNN model on text compared to the baseline SVM with LIWC features

solute error over the traits is 0.1075.

For personality induction from text, the CNN model for text beats the F-score performance of the SVM baseline by a large margin (see Table 2). Our immediate future work will focus on combining the two models described in this paper and adding facial expression data to the input.

4 Adaptive Virtual Agent

Past research has shown the importance of VAs adapting to users in terms of culture (Al-Saleh and Romano, 2015), learning style (Liew and Tan, 2016), and social constructs (Youssef et al., 2015). These scenarios show improvement in user satisfaction and also better collaboration between the user and the VA (Hu et al., 2015; Liew and Tan, 2016). In addition, users prefer adaptive agents to non-adaptive VAs when completing a task (Hu et al., 2015). However, whether and how users would prefer adaptive personality in virtual agents has not yet been explored. To this end, we conducted a counter-balanced, three-by-three factorial within-subject study with 36 participants recruited from a local university.

Each participant filled in a big-five questionnaire and was shown three different virtual agents in random order. A Robotic VA was developed as control; it replies to affect-sensitive comments with impersonal responses like “I don’t understand” and “No answer detected.” The two personality-driven VAs were based on Clifford Nass’ work on computers with personalities (Nass et al., 1995). The Tough VA (TVA) and the Gentle VA (GVA) embody the traits of dominance and submissiveness dimension of interpersonal behavior (Kiesler, 1983) respectively. Below, adjectives that signify each dimension are listed:

- Dominance: able to give orders, assumes responsibilities, controlling, self-assertive
- Submissive: easily led, lets others make decisions, and avoids responsibilities

After interacting with each VA, the participants filled out a survey to indicate and explain their VA preference and user satisfaction. We looked at the correlations between users’ Big Five scores and their VA personality preference. Especially the Openness trait correlates with Submissive preference, where higher scores indicate an increased preference for submissiveness in an agent (see Figure 4). Our results also show that 77.78% of the participants found GVA more desirable to converse with, 16.67% with TVA, and only 5% preferred the Robotic version. This is in line with human interactions, where empathy in physicians directly improves patient satisfaction and recovery (Derksen et al., 2013). Therefore, our preliminary results broadly showed that users prefer personality adaptation in VAs.

5 Future Work

Our future work would involve improving Zara’s adaptation to user personality. To increase VA’s personality flexibility, we can do a modelling of different personalities from movie and TV series characters using machine learning techniques. An adaptive personality module may also require some form of personality compatibility matching between users and the VA. To improve, we can seek works in psychology on friendship and relationship compatibility based on personality (Huston and Houts, 1998) to extract and match features

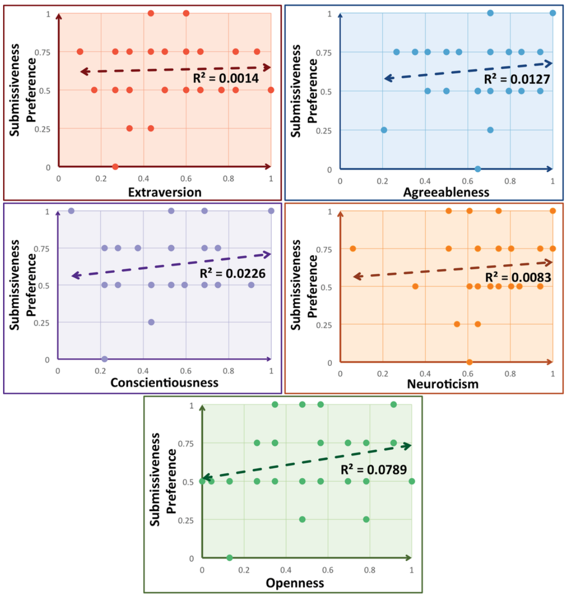


Figure 4: Correlation between user Big Five personality dimensions and VA personality

between the user and the VA. We are also designing more user experiments with more subjects and additional VA preference scales (apart from the Dominant-Submissive axis).

Furthermore, we have collected larger datasets to train our text model. One dataset contains 22 million status updates from 154,000 Facebook users, labeled with Big Five scores (Kosinski and Stillwell, 2012). Another way to improve the personality identification results is to train a separate ensemble model, which would take as input text, speech audio and facial features, and return the Big Five classifications as output.

6 Conclusion

We have described Zara the Supergirl, a multimodal interactive system that shows empathy to users and tries to assess their personality in the form of the Big Five traits, and have explained the personality modules used in detail. Training the classifier from raw audio directly without feature extraction enables real-time processing and hence makes it more suitable for applications like dialogue systems. We have also shown that our CNN classifier from text gives a much better performance when compared to the conventional SVM approach with lexical features. Our user study shows some interesting insights about the significance of personality adaptation in human-agent interactions. As we continue to build conversational systems that can detect and adapt to user personality, the interaction between user and machine can be more personal and human-like. This will in turn make it easier for people to consider virtual agents as their friends and counselors.

References

Mashael Al-Saleh and Daniela Romano. 2015. Culturally appropriate behavior in virtual agents: A review.

Shlomo Argamon, Sushant Dhawle, Moshe Koppel, and James W Pennebaker. 2005. Lexical predictors of personality type. In *Proceedings of the 2005 Joint Annual Meeting of the Interface and the Classification Society of North America*.

Murray R Barrick and Michael K Mount. 1991. The big five personality dimensions and job performance: a meta-analysis. *Personnel psychology* 44(1):1–26.

Diane S Berry, Julie K Willingham, and Christine A Thayer. 2000. Affect and personality as predictors of conflict and closeness in young adults' friendships. *Journal of Research in Personality* 34(1):84–107.

Dario Bertero, Farhad Bin Siddique, Chien-Sheng Wu, Yan Wan, Ricky Ho Yin Chan, and Pascale Fung. 2016. Real-time speech emotion and sentiment recognition for interactive dialogue systems .

Fabio Celli, Bruno Lepri, Joan-Isaac Biel, Daniel Gatica-Perez, Giuseppe Riccardi, and Fabio Pianesi. 2014. The workshop on computational personality recognition 2014. In *22nd ACM international conference on multimedia*. ACM, pages 1245–1246.

Paul T Costa and Robert R McCrae. 2008. The revised neo personality inventory (neo-pi-r). *The SAGE handbook of personality theory and assessment* 2:179–198.

Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. 2013. Effectiveness of empathy in general practice: a systematic review. *Br J Gen Pract* 63(606):e76–e84.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. opensmile - the munich versatile and fast open-source audio feature extraction. In *18th ACM international conference on multimedia*. ACM, pages 1459–1462.

Pascale Fung, Dario Bertero, Yan Wan, Anik Dey, Ricky Ho Yin Chan, Farhad Bin Siddique, Yang Yang, Chien-Sheng Wu, and Ruixi Lin. 2016. Towards empathetic human-robot interactions. *arXiv preprint arXiv:1605.04072* .

Pascale Fung, Anik Dey, Farhad Bin Siddique, Ruixi Lin, Yang Yang, Wan Yan, and Ricky Chan Ho Yin. 2015. Zara the supergirl: An empathetic personality recognition system .

Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48(1):26.

Samuel D Gosling, Peter J Rentfrow, and William B Swann. 2003. A very brief measure of the big-five personality domains. *Journal of Research in personality* 37(6):504–528.

Yağmur Güçlütürk, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. 2016. Deep impression: audiovisual deep residual networks for multimodal apparent personality trait recognition. In *Computer Vision–ECCV 2016 Workshops*. Springer, pages 349–358.

Chao Hu, Marilyn A Walker, Michael Neff, and Jean E Fox Tree. 2015. Storytelling agents with personality and adaptivity. In *International Conference on Intelligent Virtual Agents*. Springer, pages 181–193.

Ted L Huston and Renate M Houts. 1998. The psychological infrastructure of courtship and marriage: The role of personality and compatibility in romantic relationships. *The developmental course of marital dysfunction* pages 114–151.

Alexei Ivanov and Xin Chen. 2012. Modulation spectrum analysis for speaker personality trait recognition. In *INTERSPEECH*. pages 278–281.

Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. 2014. A convolutional neural network for modelling sentences. *arXiv preprint arXiv:1404.2188* .

Donald J Kiesler. 1983. The 1982 interpersonal circle: A taxonomy for complementarity in human transactions. *Psychological review* 90(3):185.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882* .

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Michal Kosinski and David Stillwell. 2012. mypersonality research wiki.

Tze Wei Liew and Su-Mae Tan. 2016. Virtual agents with personality: Adaptation of learner-agent personality in a virtual learning environment. In *Digital Information Management (ICDIM), 2016 Eleventh International Conference on*. IEEE, pages 157–162.

M Valora Long and Peter Martin. 2000. Personality, relationship closeness, and loneliness of oldest old adults and their children. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences* 55(5):P311–P319.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of artificial intelligence research* 30:457–500.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. pages 3111–3119.

Adam S Miner, Arnold Milstein, Stephen Schueller, Roshini Hegde, Christina Mangurian, and Eleni Linos. 2016. Smartphone-based conversational agents and responses to questions about mental health, interpersonal violence, and physical health. *JAMA internal medicine* 176(5):619–625.

Gelareh Mohammadi and Alessandro Vinciarelli. 2012. Automatic personality perception: Prediction of trait attribution based on prosodic features. *IEEE Transactions on Affective Computing* 3(3):273–284.

Clifford Nass, Youngme Moon, BJ Fogg, Byron Reeves, and Chris Dryer. 1995. Can computer personalities be human personalities? In *Conference companion on Human factors in computing systems*. ACM, pages 228–229.

Dimitri Palaz, Ronan Collobert, et al. 2015. Analysis of cnn-based speech recognition system using raw speech as input. Technical report, Idiap.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates* 71(2001):2001.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *Computer Vision–ECCV 2016 Workshops*. Springer, pages 400–418.

Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, Gelareh Mohammadi, and Benjamin Weiss. 2012. The interspeech 2012 speaker trait challenge. In *INTERSPEECH*. pages 254–257.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research* 15(1):1929–1958.

Ben Verhoeven, Walter Daelemans, and Tom De Smedt. 2013. Ensemble methods for personality recognition. In *Proceedings of the Workshop on Computational Personality Recognition*. pages 35–38.

Atef Ben Youssef, Mathieu Chollet, Hazaël Jones, Nicolas Sabouret, Catherine Pelachaud, and Magalie Ochs. 2015. Towards a socially adaptive virtual agent. In *International Conference on Intelligent Virtual Agents*. Springer, pages 3–16.