Appendix: A Conversational Agent Framework with Multi-modal Personality Expression

SINAN SONLU and UĞUR GÜDÜKBAY, Bilkent University, Turkey
FUNDA DURUPINAR, University of Massachusetts, USA

A. SCENARIOS

A.1 Introduction Scenario

In this scenario, the user is asked to learn certain information about the agent including name, age, occupation, and the city agent lives in. We use a neutral outdoor scene setup for this scenario, and the upper body of the agent is visible to the user. The dialogue state machine of this scenario supports non-linear flows, as seen in Figure 1, yet the user could tend to follow a certain order. The user can end the conversation with a farewell at any time. Figure 2 shows still frames from the scenario.

The name, age, occupation, and city of the agent are randomized at each execution. OCEAN Alternatives of agent Dialogue Units have a minimal distinction in this scenario because the questions of the user are rather direct. An example Dialogue Unit (DU) with its OCEAN Alternatives is given in Table 1.

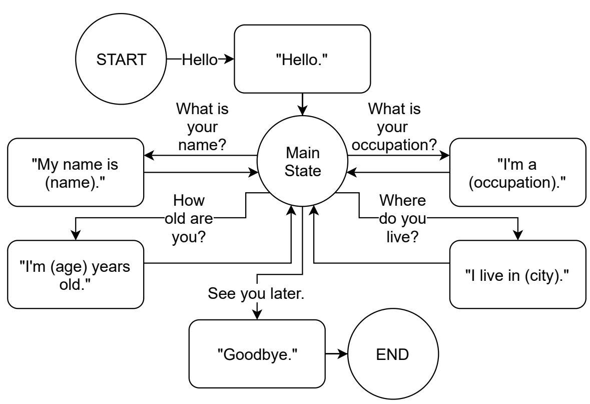


Fig. 1. The dialogue state machine of Introduction Scenario.

Fig. 2. Still frames from Introduction Scenario. We use an unconscientious agent (careless, negligent). White subtitles correspond to the agent's lines.

A.2 Fastfood Scenario

In this scenario, the user takes the role of a customer, with the purpose to order some food by talking to the cashier agent, in a fast-food restaurant setup. We do not give specific information about what to order to the user. We expect the user to ask what food the restaurant serves before ordering.

Authors' addresses: Sinan Sonlu, sinan.sonlu@bilkent.edu.tr; Uğur Güdükbay, gudukbay@cs.bilkent.edu.tr, Department of Computer Engineering, Bilkent University, Bilkent, Ankara, 06800, Turkey; Funda Durupınar, fundad@cs.umb.edu, Department of Computer Science, University of Massachusetts, 100 Morrisey Boulevard, Boston, MA, 02125, USA.

0730-0301/2020/1-ART1 $15.00
https://doi.org/10.1145/3439795

Table 1. Example OCEAN Alternatives for a DU from Introduction Scenario.

Type	Text
DU	"My name is (name)."
$O(+)$	"I am known as (name)."
$O(-)$	"I'm (name)."
$C(+)$	"My name is (name)."
$C(-)$	"Oh, well... My name is... (name)."
$E(+)$	"I'm (name), my friend."
$E(-)$	"(name)."
$A(+)$	"My name is (name), nice to meet you."
$A(-)$	"Why do you ask? It's (name)."
$N(+)$	"Um... I... I am (name)."
$N(-)$	"My name is (name)."

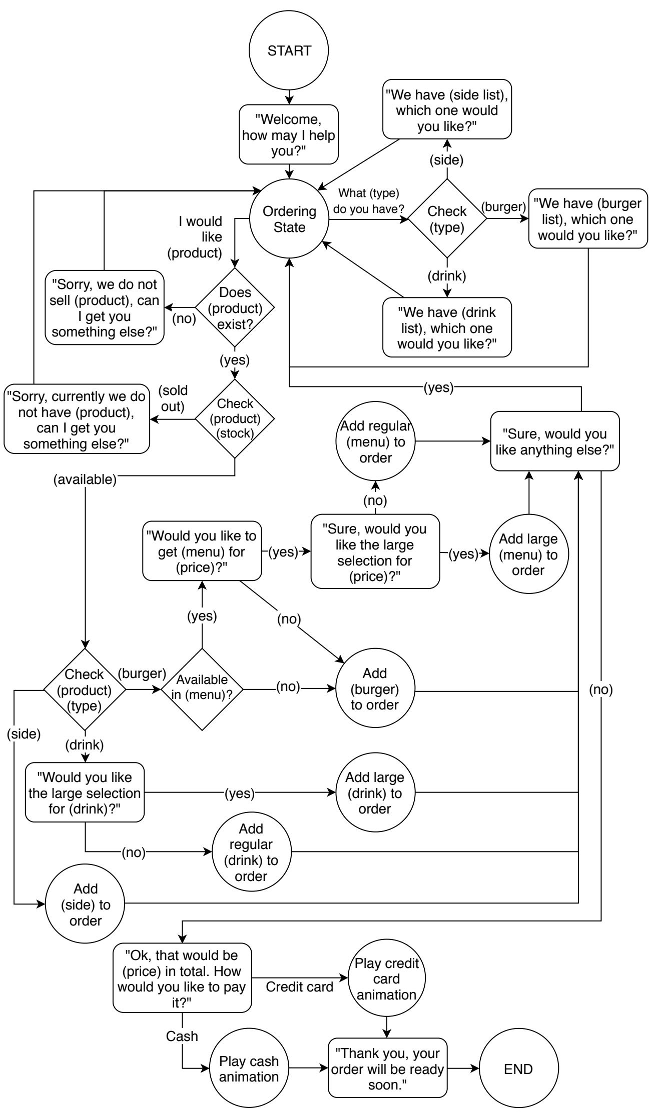


Fig. 3. The dialogue state machine of Fastfood Scenario.

Some products could be out of stock in this scenario. The cashier asks whether the user wants a menu or the big selection when it is applicable. At the end of the dialogue, the user has the option to pay with a credit card or cash. The scenario ends with the cashier preparing the order. Figure 3 gives the dialogue state machine of this scenario and Figure 4 shows still frames from the scenario.

Fig. 4. Still frames from Fastfood Scenario. We use an extravert agent (talkative, sociable). White subtitles correspond to agent's lines.

A.3 Passport Scenario

In Passport Scenario, the user takes the role of a passport officer. This scenario aims to question the visitor agent. Figure 5 shows the dialogue state machine of this scenario. We use an airport setup for this scenario, as seen in Figure 6. The agent's passport information includes visa and passport issue and expiration dates, and a visa type. To enter the country, the issue and expiration dates should be valid and the passenger's purpose of visit should be appropriate to the visa type. We show a guiding message on the screen according to the current state to help the user assume the role, but the user does not have to follow the guide and could decide on the final decision independently. We guide the user to ask the occupation and return ticket of the visitor as well.

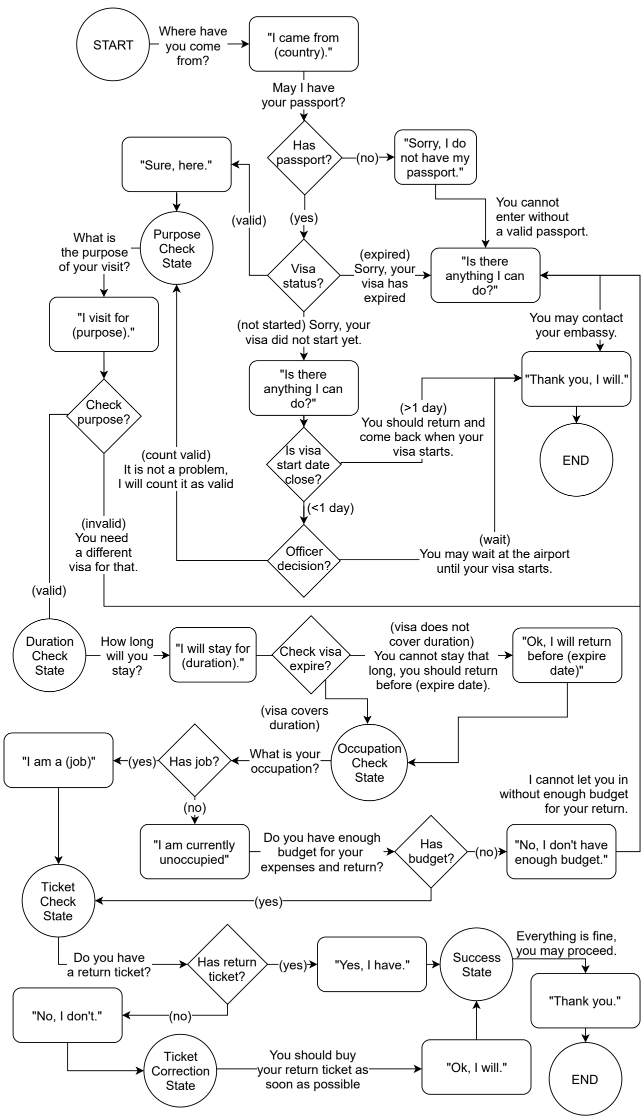


Fig. 5. The dialogue state machine of Passport Scenario.

Fig. 6. Still frames from Passport Scenario. We use an introvert agent (quiet, reserved). White subtitles correspond to agent's lines.

B. FACIAL SHAPE KEY WEIGHTS

Table 2 contains the weights of the facial shape keys per emotion.

Table 2. Emotion weights for facial expression keys

Key/Emotion	Anger	Disgust	Sadness	Surprise	Happiness
Brows Down	1.00	0.50	0.00	0.00	0.00
Cheek Puff	0.02	0.00	0.00	0.00	0.00
Frown	0.40	0.10	0.80	0.00	0.00
Mouth Down	0.10	0.10	0.10	0.00	0.00
Mouth Narrow	0.20	0.00	0.00	0.20	0.00
Squint	0.30	0.30	0.10	0.00	0.40
Brows Up	0.00	0.00	0.10	1.00	0.30
Eyes Wide	0.00	0.00	0.00	0.80	0.00
Mouth Open	0.00	0.00	0.00	0.30	0.10
Smile	0.00	0.00	0.00	0.00	1.00
Brows Outer Lower	0.00	0.00	1.00	0.50	0.00
Brows In	0.00	0.00	0.10	0.00	0.00
Jaw Backward	0.00	0.00	0.10	0.00	0.00
Nose Scrunch	0.00	0.80	0.00	0.00	0.00
Mouth Up	0.00	0.10	0.00	0.00	0.00
Jaw Forward	0.00	0.10	0.00	0.00	0.00
Upper Lip Out	0.00	0.50	0.00	0.00	0.00
Upper Lip In	0.00	0.20	0.00	0.00	0.00
Mid Mouth	0.00	0.20	0.00	0.00	0.00

C. ADDITIONAL EXPERIMENTS

C.1 Neutral Agent Experiment

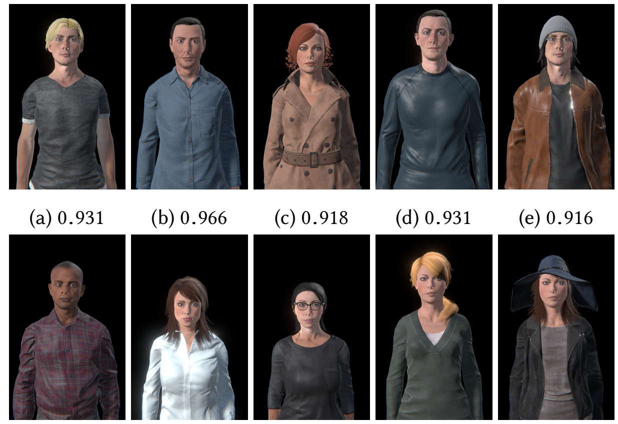
To use an agent model with minimal personality bias in the main experiments, we asked 25 participants to rate ten 3D human models manually created using Fuse [Adobe 2019]. These models, shown in Figure 7, were constructed using preset parts available in the software to give a distinct look per character. In this experiment, each sample is the image of a different agent. Using the same idle pose, we expect a slightly different mean OCEAN score for each 3D human model due to differences in appearance. Ideally, a neutral agent should have a score of 0.5 per OCEAN factor. We define the neutrality of an agent as:

$$N = 1 - \frac{\sum |OCEAN - 0.5|}{5},$$

$N = 1$ being the agent with all neutral OCEAN factors. The results of this experiment indicate that the agent in Figure 7 (b) is perceived as the most neutral one, with an average score of $N = 0.966$, and the agent in Figure 7 (e) is the least neutral, with an average score of $N = 0.916$. Figure 7 provides all agent images and their corresponding N values. Being found as the most neutral, the Figure 7 (b) agent is used in the rest of the experiments.

C.2 Facial Expression-Personality Experiment

There are notable studies in cognitive science that investigate the relationship between facial expressions and personality. Although the literature is more interested in the influence of the observer's facial expression, various examples such as Todorov et al. [2008] examine the correlation between the facial appearance of a person and others' inference of this person's personality. They measure trustworthiness (valence) and dominance (power) in computer-generated faces, rather than using a personality model such as OCEAN. They morph the whole face, thus the resulting faces vary in terms of both facial



(a) 0.931 (b) 0.966 (c) 0.918 (d) 0.931 (e) 0.916

(f) 0.919 (g) 0.950 (h) 0.938 (i) 0.938 (j) 0.949

Fig. 7. The 3D human models used in Agent Neutrality Experiment and the corresponding neutrality means.

expression and facial structure. They also investigate the potential role of the amygdala in face evaluation.

Knutson [1996] investigates how facial expressions of emotion affect subjects' interpersonal trait inferences with two experiments. He concludes that facial expressions carry both a target's internal state, as well as interpersonal information. Teijeiro-Mosquera et al. [2015] investigate the relation between facial expression and OCEAN personality inference in video blogs of real humans.

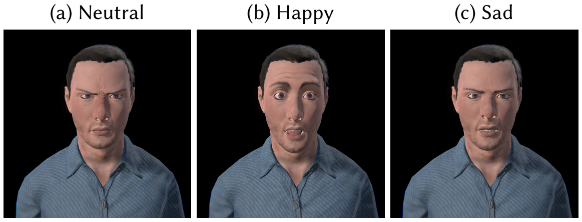
As these studies suggest, facial expressions influence personality inference. However, to our knowledge, a mapping between OCEAN personality and facial expression usage for virtual humans has not been established. We perform this experiment to quantify the relationship between facial expressions and the OCEAN personalities for virtual humans, where the expressiveness of the face is limited.

We ran this study based on images of the most neutral agent expressing different emotions. The agent's facial expression was set to neutral, happy, sad, angry, surprised, and disgusted in different samples. We showed the agent's face as a close-up portrait (see Figure 8). We preferred to use static images of the facial expressions for participants to focus on the expression itself rather than the animation. Keeping the facial expression of the agent still in animation would make it unnatural, and keeping it short would not give the participant enough time to perceive it.

Each sample was rated by 100 distinct participants. Figure 9 depicts the normalized OCEAN score distribution graph of each image. The scores of the neutral expression indicate a minor bias in openness, extraversion and neuroticism, and significant bias in conscientiousness. We take the neutral expression's OCEAN score as a base value and compare other expressions to the neutral expression to compensate for this bias.

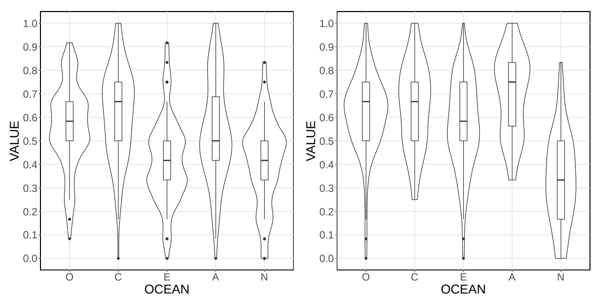
Table 3 shows the differences in means of user responses for each emotional expression and neutral expression per personality factor. We observe the highest difference for happiness in extraversion and agreeableness; for sadness in neuroticism and introversion; for anger in disagreeableness and neuroticism; for surprise in unconscientiousness and neuroticism; and for disgust in disagreeableness

(a) Neutral (b) Happy (c) Sad

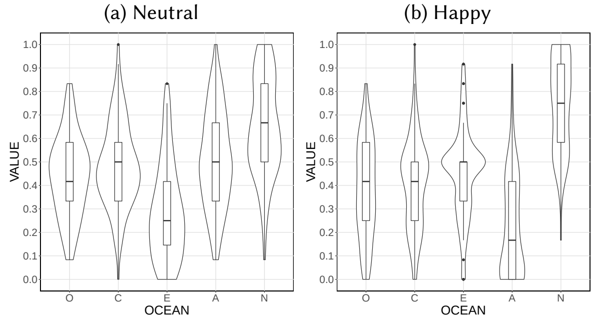


(d) Angry (e) Surprised (f) Disgusted

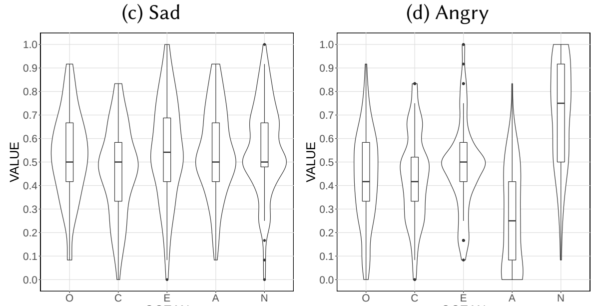
Fig. 8. The facial expressions for Facial Expression-Personality Experiment.



(a) Neutral (b) Happy



(c) Sad (d) Angry



(e) Surprised (f) Disgusted

Fig. 9. The OCEAN score distribution graphs of samples from Facial Expression - Personality Experiment.

and neuroticism. The results support our hypothesis about facial expressions being influential on the perceived personality.

Based on the statistically significant findings given in Table 3, we define additive facial expression values for each OCEAN factor extreme, included in the article. We determine the additive facial expression values based on the statistically significant influences of different facial expressions on each OCEAN factor. The sign and magnitude of the influence in Table 3 is the primary determinant. For differences of magnitudes higher than .3, .2, .1, and .05 we use the additive facial expression values of magnitudes .5, .35, .25, and .125, respectively, using the same sign. If there are multiple facial expressions with the same sign, strongly associated with an OCEAN extreme (i.e. with magnitude >.2), we limit their contribution to avoid unnatural blends. When an agent speaks, its facial expression is updated based on its personality and the emotion derived from the dialogue by IBM Watson NLP. For instance, an agreeable personality increases happiness and decreases anger and disgust values coming from the dialogue unit. The agent can still express anger or disgust; however, the emotion value calculated using IBM Watson NLP should be high enough to surpass the subtractive values, making agreeable agents less likely to express negative emotions.

Table 3. The differences in means of each facial expression and neutral facial expression. Statistically significant values with $p < 0.05$ are shown in bold.

Factor	Happy	Sad	Angry	Surprised	Disgusted
O	**0.057**	**-0.121**	**-0.177**	-0.089	**-0.128**
C	0.034	**-0.128**	**-0.203**	**-0.180**	**-0.175**
E	**0.157**	**-0.130**	0.037	**-0.020**	**0.063**
A	**0.155**	-0.033	**-0.311**	-0.152	**-0.279**
N	**-0.074**	**0.236**	**0.330**	**0.222**	**0.289**

C.3 Evaluation of Body Movement Modification Parameters

In our system, personality is conveyed through body motion via Laban Effort (LE) factors and Laban Shape Qualities (LSQ). A mapping between LE and personality has been established in the literature [Durupinar et al. 2017], and we adopt the same mapping although we represent the same Effort factors with different motion parameters. This follows the premise of [Durupinar et al. 2017] because the idea behind using Laban parameters is to decouple the implementation details of motion parameters from what they represent. As for an association between LSQ and personality, we hypothesize the correlations in Table 6 of the main article.

We performed Amazon Mechanical Turk [Amazon 2018] studies to validate (1) our LE implementation, (2) the effectiveness of our personality-LSQ mapping (cf. Table 6, main article) on improving the perception of related personality factors. We showed the participants the videos of two side-by-side agents performing the same set of actions with different motion styles. We used the Ten-Item Personality Inventory [Gosling et al. 2003] to collect evaluations. For instance, the question assessing openness was formatted as: "Which character looks more **open to new experiences & complex** and less **conventional & uncreative**". The participants were instructed to choose "Left", "Equal", or "Right" as their answer. The videos could be replayed as many times as desired. We randomized the left/right positioning of the two agents. We collected 40 unique responses for each video.

There was a total of 20 tasks, assessing three different settings. Participants were free to perform any number of tasks. There were 68 unique participants with an average age of 29.30 ± 6.46. The majority of the participants were from India (42.64%), followed by the USA (29.41%) and Italy (5.88%).

The first group of tasks compared two opposite personality factors for each dimension, using only LE adjustments (see Figure 10).

Fig. 10. Introvert (left) and Extravert (right) agents using only LE adjustments.

Results indicate the success of our LE implementation (see Figure 11). For each personality dimension, we counted the number of responses for exact personality, opposite personality, and neutral answers. We assume the null hypothesis to be that the numbers of responses for the three options are distributed equally. The results of the two-tailed t-tests yield that the ratio of expected answers is significantly higher than the opposite and neutral answers with $p < 0.001$. The best-distinguished factor is extraversion, followed by openness, agreeableness, conscientiousness, and neuroticism.

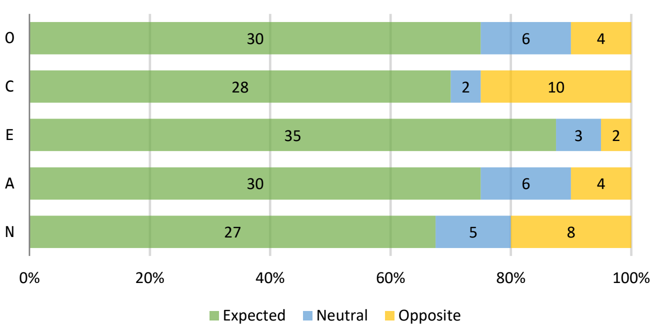


Fig. 11. Accuracy of the participants' personality perception comparing two opposite factors using only LE adjustments.

The second group of tasks compared two opposite personality factors for each dimension using both LE and LSQ adjustments (see Figure 12).

Fig. 12. Introvert (left) and Extravert (right) agents using both LE and LSQ adjustments.

We followed the same approach to analyze the responses and performed t-tests. The two-tailed p values for all the personality factors are less than 0.001. Thus, the additional LSQ adjustments significantly improve the performance, compared to using only LE adjustments (see Figure 13).

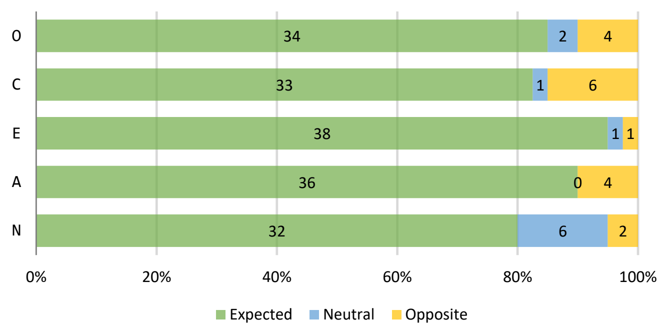


Fig. 13. Accuracy of the participants' personality perception comparing two opposite factors using both LE and LSQ adjustments.

Additionally, we report paired t-test results to compare the effect of using only LE adjustments to using both LE and LSQ adjustments. Table 4 shows the improvement rate of the LSQ adjustments for each factor with $p < 0.05$.

Factor	Accuracy			p-value
	LE	LE + LSQ	Improvement	
O	.75	.85	.10	.043
C	.70	.82	.12	.018
E	.87	.95	.07	.043
A	.75	.90	.15	.012
N	.67	.80	.12	.001

Table 4. Paired t-test results comparing LE adjustments to LE + LSQ adjustments.

Finally, the third group of tasks displayed two side-by-side agents expressing the same personality factor, where one agent is animated with only LE adjustments and the other agent with both LE and LSQ adjustments (see Figure 14). The question format was the same, again asking the participant to select the agent that best represents the personality trait in question.

Fig. 14. Introvert (left image) and Extravert (right image) agents, in each image the agent on left uses LE only, and the agent on right uses both LE and LSQ.

Figure 15 shows the response counts and rates for the expected, neutral, and opposite answers. Assuming the null hypothesis to be the random selection of these three options, we counted the number of responses for each group and performed two-tailed t-tests. For all the factors except N-, the ratio of expected answers is significantly higher than the opposite and neutral answers ($p < 0.001$). The poor performance of N- is possibly because we constrained the IK weights for LSQ anchors to prevent self collisions. Self collisions are especially salient in Retreating motion; thus the weights are limited the most, decreasing the impact of LSQ modifications.

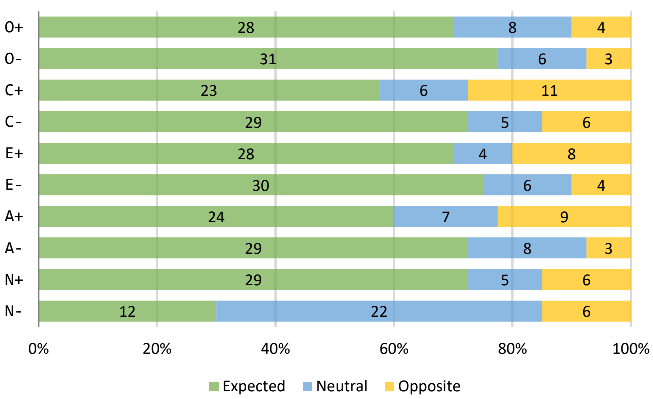


Fig. 15. Rate of the participants that chose the agent with both LE and LSQ adjustments over the agent with only LE adjustments, comparing the same polarity for each positive and negative OCEAN factor.

C.4 Naturalness Assessment of Agent Movement and Speech

We conducted a user survey to measure the naturalness of the agent in terms of its movement and speech for Models VF, VB, and VFB. We gave a 7 point Likert-scale for statements "The movements of this character feel natural" and "The speech of this character feels natural", changing between 1 (disagree strongly) and 7 (agree strongly). 50 Amazon Mechanical Turk workers participated in the study. The mean scores for each sample are depicted in Figure 16.

Although the speech of the agent remains unchanged, we see a trend towards judging the speech based on the naturalness of the movement. Facial expressions of the agent are perceived as more natural than other modalities, and the inclusion of body movement seems to have a negative effect. This may be due to some motion modification artifacts that cause self-collisions.

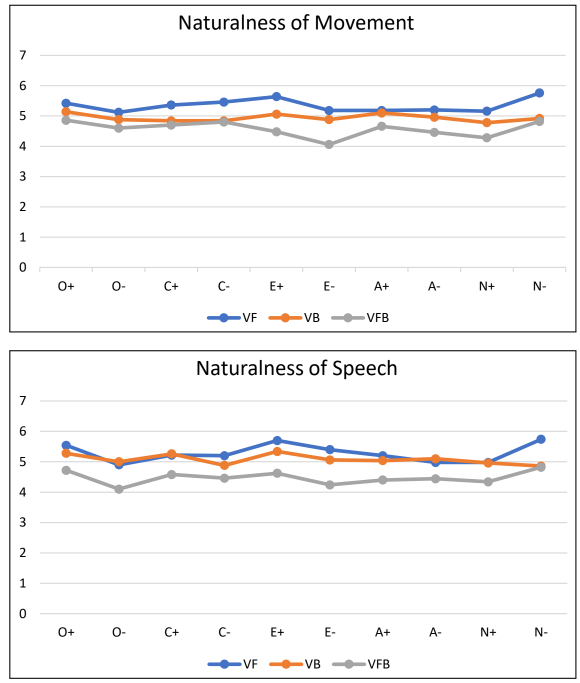


Fig. 16. Mean naturalness scores per sample for Models VF, VB and VFB. Scores are given on 7 point Likert-scale, 1 being the least natural and 7 being the most natural.

D. GRAPHS SUMMARIZING THE RESULTS

Figures 17 and 18 depict the mean differences between the scores of positive and negative samples for each model and each silent model, respectively. Higher values correspond to a better distinction of the opposite ends of the related OCEAN factor.

REFERENCES

Funda Durupinar, Mubbasir Kapadia, Susan Deutsch, Michael Neff, and Norman I Badler. 2017. PERFORM: Perceptual Approach for Adding OCEAN Personality to Human Motion Using Laban Movement Analysis. *ACM Trans. Graph.* 36, 1, Article 6 (2017), 16 pages.

Samuel D Gosling, Peter J Rentfrow, and William B Swann Jr. 2003. A Very Brief Measure of the Big-Five Personality Domains. *J. Res. Personality* 37, 6 (2003), 504–528.

Brian Knutson. 1996. Facial expressions of emotion influence interpersonal trait inferences. *J. Nonverbal Behav.* 20 (1996), 165–182.

Adobe. 2019. Fuse CC. https://www.adobe.com/tr/products/fuse.html. Accessed: 2019-08-14.

Amazon. 2018. Mechanical Turk. https://www.mturk.com/. Accessed: 2019-08-14.

L. Teijeiro-Mosquera, J. Biel, J. L. Alba-Castro, and D. Gatica-Perez. 2015. What Your Face Vlogs About: Expressions of Emotion and Big-Five Traits Impressions in YouTube. *IEEE Trans. Affect. Comp.* 6, 2 (2015), 193–205.

Alexander Todorov, Chris P. Said, Andrew D. Engell, and Nikolaas N. Oosterhof. 2008. Understanding evaluation of faces on social dimensions. *Trends Cogn. Sci* 12, 12 (2008), 455–460.

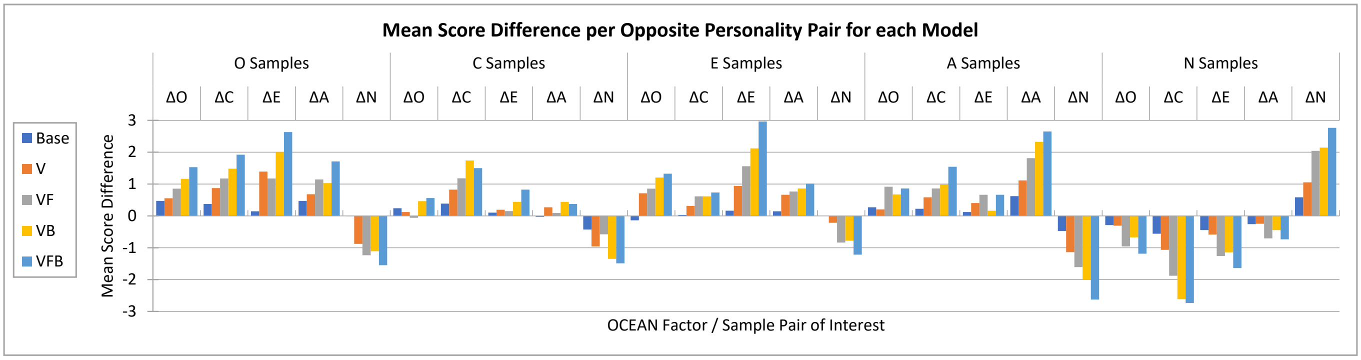


Fig. 17. The comparison of different models for each OCEAN factor with differences in means of positive and negative samples (higher magnitude is better).

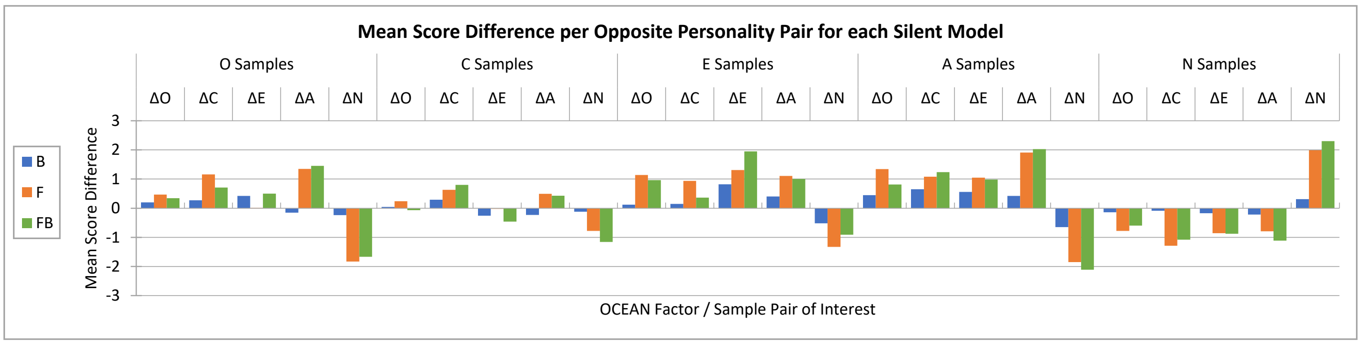


Fig. 18. The comparison of different silent models for each OCEAN factor with differences in means of positive and negative samples (higher is better).