Synthesising Personality with Neural Speech Synthesis

Anonymous ACL submission

Abstract

Matching the personality of conversational agent to the personality of the user can significantly improve the user experience, with many successful examples in text-based chatbots. It is also important for a voice-based system to be able to alter the personality of the speech as perceived by the users. In this pilot study, fifteen voices were rated using Big Five personality traits. Five content-neutral sentences were chosen for the listening tests. The audio data, together with two rated traits (Extroversion and Agreeableness), were used to train a neural speech synthesiser based on one male and one female voices. The effect of altering the personality trait features was evaluated by a second listening test. Both perceived extroversion and agreeableness in the synthetic voices were affected significantly. The controllable range was limited due to a lack of variance in the source audio data. The perceived personality traits correlated with each other and with the naturalness of the speech.

1 Introduction

The law of attraction in human-robot interaction means users prefer social robots with similar personality traits to themselves (Park et al., 2012). Previous work has shown that it is possible to design a text-based chatbot with a pre-defined personality (Ahmad et al., 2020; Ruane et al., 2021), and matching the personality of the agent to the personality of the user can significantly improve the user experience (Smestad and Volden, 2019; Fernau et al., 2022). Personality in voice-based conversational agent is much less investigated, but the effect is no less significant. People attribute traits to others in less than a second after hearing them in video and/or audio recordings (Reeves and Nass, 1996; Uleman et al., 2008). The same effect extends to machines that display human-like features including embodied conversational agents (Nass and Brave, 2005). The perceived personality from speech is consistent across listeners (McAleer et al., 2014). This opens the possibility of generating synthetic voices that encourage users to attribute pre-defined traits to the artificial intelligence conversational agents they interact with.

Previous work (Aylett et al., 2017) has shown that personality can be manipulated with a speech synthesis system. The effect is restrained by the system used: unit selection is heavily constrained by the corpus recorded (though there have been advances in addressing this (Buchanan et al., 2018)), whilst HMM-based Speech Synthesis (HTS) is constrained by perceived naturalness. Neural speech synthesis systems such as Wavenet (Oord et al., 2016) and Tacotron (Wang et al., 2017a) has shown an improved ability to generate natural sounding output. This has led to advancement in expressive speech synthesis (Wang et al., 2017b, 2018; Zhang et al., 2019). However the focus is on manipulating the style of single utterances and is different from synthesising a voice with a consistent personality. Recent work (Shiramizu et al., 2022) achieved altering the social perception of synthetic speech by controlling single speech-based features such as pitch. It is interesting to see the effect of using neural speech synthesis system to manipulate the perceived personality traits of the output voice.

In this work the use of Big Five scores is explored for directly controlling the perceived personality of the synthetic speech. Big Five, or OCEAN model (John et al., 1999), is widely used the domain of human-computer interaction (Vinciarelli and Mohammadi, 2014). A condensed version (Rammstedt and John, 2007) that reduces the original 44 statements to ten while preserving a high level of accuracy was used.

2 Experiments

2.1 Big Five Rating of Source Voices

Our dataset comprised of 15 English native speaker voices taken from CereProc's voice bank. The voices varied by accent and gender, see Table 1.

Gender	Received Pronunciation	Scottish	Irish English	Total
Male	5	2	0	7
Female	5	2	1	8
Total	10	4	1	15

Table 1: Accent and gender distribution

For the listening tests, five news sentences were chosen for their content being emotionally neutral but can be read with different personalities (see Table 2). 28 English native listeners were recruited from Amazon Mechanical Turk (AMT) to rate the Big Five personality traits of each source voice. A web-based listening test was used to measure Big Five based on ten personality questions (Rammstedt and John, 2007) with an additional naturalness question using a 5-point Likert scale. Two slide bars were used to measure perceived age (10-70), and perceived gender (0-1, from woman to man). The system displayed the audio transcript and allowed participants to play the audio stimuli repeated times. Each participant listened to a subset of 5 speakers and for each of those speakers they listened to 5 audio examples. The audio order was randomised for each listener and each audio example was rated by nine or ten listeners.

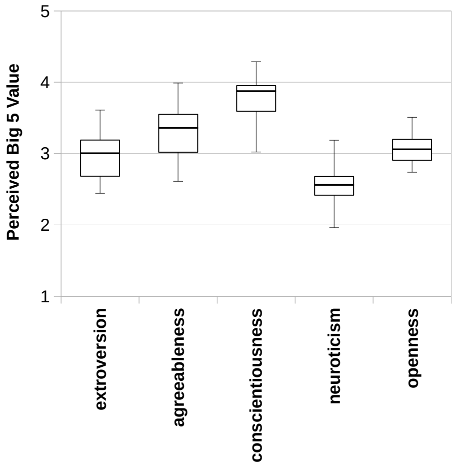

Figure 1: Box Plot of Big Five personalities averaged by voice.

Results were averaged by voice to give an overall personality score for that voice and are shown in Figure 1. Extroversion and agreeableness were chosen as the two personality traits to control as they showed the most variation.

Figure 2 shows the spread of the voices in the 1-5 Likert scale across both traits. The variation across the two traits in the database is between 2 and 4. This is expected as voice talents are often chosen on similar criteria, and the recording process for speech synthesis tends to avoid high energy emotional content which puts an artificial limit on the possible perceived personality variation within the voice. There is a positive correlation between the two traits (Pearson $r = 0.664$, $df = 13$, $p < 0.05$). The *r-squared* value is relatively low (0.441), meaning that although there is a significant positive correlation, it might not be linear or the data might not be enough to make an accurate prediction. Theoretically, the Big Five model is based on factor analysis which aims at producing independent dimensions (John et al., 1999), however, this is for *actual* personality and may not translate to independence in *perceived* personality.

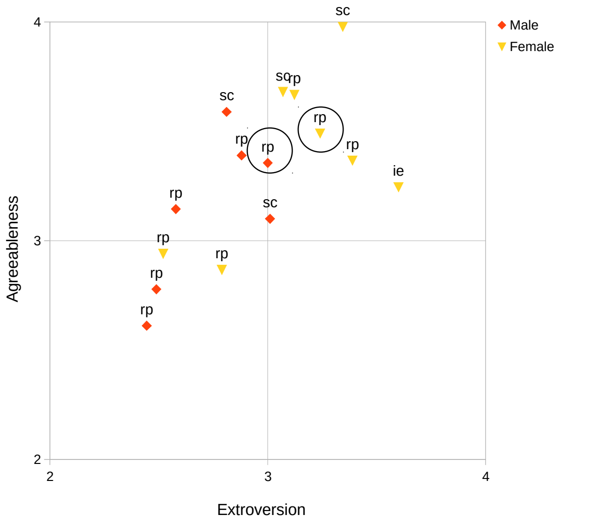

Figure 2: Distribution of average perceived extroversion and agreeableness by voice, target voices circled.

2.2 Building the Multiple Speaker Synthesis Voice

We used CereProc's Deep Neural Networks (DNN) speech synthesis system CereWave to build a multi-speaker voice. CereWave uses a recurrent neural network architecture to firstly produce prosody targets, and then produce an intermediate acoustic feature set. After predicting the acoustic features, it uses a custom neural vocoder to produce the final output waveforms. Its inputs include phonetic, linguistic, language, accent and speaker features, in which speaker features include age and gender. For this experiment, the personality dimensions chosen at the first stage (extroversion and agreeableness) are appended to the above features in the format of an average voice score on a 5-point Likert scale. Due to the time constraints of this research and its nature of pilot study, we limited the data to a total of 1,000 utterances from the neutral speaking style data (totalling approximately 2 hours of data), which puts an limitation on the naturalness.

When synthesising from an average voice, an original speaker specification can be used to generate synthesis sounding like that speaker. Two voices, one male and one female, close to the global mean for all voices in terms of extroversion and agreeableness, were chosen to synthesise stimuli (Male voice: mean extroversion 3.0, mean agreeableness 3.4; Female voice: mean extroversion 3.2, mean agreeableness 3.5). In addition, natural recordings for each of these speakers were used as a high naturalness anchor, and synthesis using a previous generation DNN system were used as a low naturalness anchor. Five utterances were synthesised for all synthesis conditions.

2.3 Evaluating the Synthesis of Agreeableness and Extroversion

A second AMT listening test was carried out using the same interface and methodology described in section 2.1 with 18 participants. It is expected that synthesised voices' personality would not match the reference speakers exactly but should be similar. This was the case for the male voice but the synthesis process reduced both the perceived extroversion and agreeableness of the female voice (Male voice: mean extroversion 3.0, mean agreeableness 3.4; Female voice: mean extroversion 2.9, mean agreeableness 3.1).

Results were averaged over the 10 utterances (5 spoken by two voices) and a by-materials repeated measures MANOVA was carried out with perceived extroversion and agreeableness as the dependent variable. Target extroversion (tgt-e: low/high) and nested target agreeableness (tgt-a: low/high) were within-materials factors, with base synthesis voice (gender: male/female) as a between-materials factor. Both target factors were significant in a multivariate test (Wilks Lambda: tgt-a (F(2, 7)=21.258, p=0.001), tgt-e (F(2, 7)=11.422, p<0.01)), gender did not have a significant effect. Univariate tests with a Greenhouse-Geisser correction (sphericity not assumed) showed that target extroversion significantly affected perceived extroversion (tgt-e F(1, 8)=24.981, p=0.001) but not perceived agreeableness, whereas target agreeableness significantly affected both perceived agreeability (tgt-e F(1, 8)=47.399, p<0.001) and extroversion (tgt-e (F(1, 8)=34.561, p<0.001)).

In terms of the adjusted means by target groups, agreeableness has the desired effect on perceived agreeableness (tgt-a low: mean 2.922, Standard Error (SE) 0.048; high: mean 3.206, SE 0.039), but also significantly affected perceived extroversion (tgt-a low: mean 2.639, SE 0.055; high: 3.156, SE 0.06). Extroversion had the opposite affect on perceived extroversion as the higher target actually reduced perceived extroversion (tgt-e low: mean 3.019, SE 0.053; high: 2.775, SE 0.034).

The effect of trait targeting on speech rate, pitch and amplitude is also evalauted using Pearson's correlation analysis. Only speech rate had a significant effect (extroversion/words-per-second: $r(40)=0.29$, $p<0.05$), agreeableness/words-per-second: $r(40)=0.23$, $p<0.005$).

Figure 3 shows the average extroversion/ agreeableness by synthesis type. The manipulation targets are: '+e+a' to be positioned at 4,4; '+e-a' at 4,2; '-e+a' at 2,4; and '-e-a' at 2,2. It is shown that the perceived variation is much lower than this (between 2.5 and 3.5), and the spread does not form the pattern expected above.

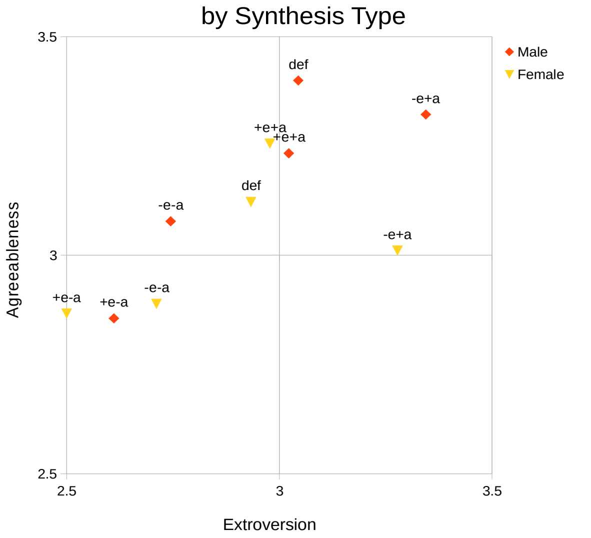

Figure 3: Distribution of average perceived extroversion and agreeableness for different synthesis types. 'def': synthesis with no personality modelling. '+': high(4), '-': low(2). 'e': extroversion, 'a': agreeableness.

2.3.1 Effect on naturalness

A univariate repeated measures ANOVA with a Greenhouse-Geisser correction (sphericity could not be assumed) was carried out to explore the effect of trait targeting on perceived naturalness. Naturalness, initially recorded on a 1-5 Likert scale, was averaged by utterance for each synthesis type and used as the dependent variable. The model matched that used in the previous MANOVA. Target extroversion (tgt-e: low/high) and nested target agreeableness (tgt-a: low/high) were within-materials factors, with base synthesis voice (gender: male/female) as a between-materials factor. Target agreeableness was significant (F(1,8)=39.784, $p<0.001$) where a high target increased perceived naturalness (tgt-a low: mean 2.8, SE 0.073; high: mean 3.339, SE 0.052). There was also a significant effect for an interaction between voice and target extroversion (F(1,8)=5.967, $p<0.05$). This effect was caused by high target extroversion increasing perceived naturalness for the female voice (tgt-e*gender low: mean 2.289, SE 0.091; high: mean 3.178, SE 0.078) and reducing naturalness for the male voice (tgt-e*gender low: mean 3.144, SE 0.091; high: mean 2.967, SE 0.078).

Values for perceived extroversion, agreeableness and naturalness were averaged across subjects for each of the utterances in all four conditions (tgt-e: low/high, tgt-a: low/high) and for both male and female voices (40 data points in total). A Pearson correlation showed a significant positive correlation between perceived extroversion, perceived agreeableness and perceived naturalness. (extroversion/agreeableness: $r(40)=0.507$, p=0.001, extroversion/naturalness: $r(40)=0.641$, $p<0.001$, agreeableness/naturalness $r(40)=0.512$, p=0.001).

3 Discussion

This pilot study shows that using the personality traits to control the perceived personality of a synthetic voice is feasible with a modern DNN / neural vocoder system. Readers are invited to listen to sample natural and synthetic speech from the attached materials. Changing input features and manipulating the target for agreeableness both alter the perceived personalities in the expected direction. However, the range in agreeableness that can be controlled, as well as the lack of a similar result for extroversion, show that controlling perceived personality is a far from simple process.

Two limitations have compromised the results of the study: 1) The corpus used as a basis for this experiment was comprised of voices originally selected for being extrovert and agreeable, which can be seen from Figure 1 and Figure 2. With a machine learning approach this means when targets are set within outlying regions the system has to extrapolate the results which leads to unnatural results as they are not based on actual observations. This is shown for agreeableness where lower target scores (unseen in the data) generate stimuli rated lower for naturalness. In future work it will be important to source a corpus with a much wider variation in perceived Big Five personality traits. 2) The interaction between traits and naturalness appear to complicate perceived trait scores. In previous work, using actual vocal change in the data, or changing synthesis style, appeared to change Big Five without correlating with naturalness variation(Aylett et al., 2017). This work, however, shows a strong correlation between perceived agreeableness and perceived extroversion and naturalness. Such collinearity means it is difficult to produce stable results. The confounding effect is possibly intensified by using an average voice built with a limited amount of source data.

4 Conclusion and future work

To summarise our findings: 1) The prototype system showed a Big Five trait could be learned and controlled, though control may be limited in the controllable range. 2) Naturalness can interact with personality traits and ensuring the underlying average voice is as natural as possible is an important consideration. 3) Correlations across traits may interfere with final results.

The next steps would be to repeat the annotation and training with a dataset that contains a wide variety of speakers such as VCTK (Yamagishi et al., 2019), and apply the synthetic voice in a multi-turn voice-based conversational agent set-up. Methods of including personality features that are more sophisticated than concatenation on the input features can be explored, both in terms of architecture and training approaches (Gibiansky et al., 2017).

Further experiments can be using personality synthesis in speech together with text-based personality generation. This work suggests the possibility of making a chatbot speak in a voice with 1) pre-defined personality based on the generated text, which can be matching or mismatching, and 2) adaptive personality based on the personality of the user, as such adaptation is shown possible in text-based chatbots (Fernau et al., 2022). A multi-turn conversational set-up can also be used to experiment the consistency of synthetised personality.

References

Rangina Ahmad, Dominik Siemon, and Susanne Robra-Bissantz. 2020. Extrabot vs introbot: The influence of linguistic cues on communication satisfaction. In *AMCIS*.

Matthew P Aylett, Alessandro Vinciarelli, and Mirjam Wester. 2017. Speech synthesis for the generation of artificial personality. *IEEE Transactions on Affective Computing*.

Christopher G. Buchanan, Matthew P. Aylett, and David A. Braude. 2018. Adding personality to neutral speech synthesis voices. In *SPECOM*.

Daniel Fernau, Stefan Hillmann, Nils Feldhus, Tim Polzehl, and Sebastian Möller. 2022. Towards personality-aware chatbots. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 135–145.

Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou. 2017. Deep voice 2: Multi-speaker neural text-to-speech. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2962–2970. Curran Associates, Inc.

Oliver P John, Sanjay Srivastava, et al. 1999. The big five trait taxonomy: History, measurement, and theoretical perspectives. *Handbook of personality: Theory and research*, 2(1999):102–138.

Phil McAleer, Alexander Todorov, and Pascal Belin. 2014. How do you say ‘hello’? personality impressions from brief novel voices. *PloS one*, 9(3).

C. Nass and S. Brave. 2005. *Wired for speech: How voice activates and advances the Human-Computer relationship*. The MIT Press.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*.

Eunil Park, Dallae Jin, and Angel P del Pobil. 2012. The law of attraction in human-robot interaction. *International Journal of Advanced Robotic Systems*, 9(2):35.

Beatrice Rammstedt and Oliver P John. 2007. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212.

B. Reeves and C. Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge University Press.

Elayne Ruane, Sinead Farrell, and Anthony Ventresque. 2021. User perception of text-based chatbot personality. In *Chatbot Research and Design: 4th International Workshop, CONVERSATIONS 2020, Virtual Event, November 23–24, 2020, Revised Selected Papers 4*, pages 32–47. Springer.

Victor Kenji M Shiramizu, Anthony J Lee, Daria Altenburg, David R Feinberg, and Benedict C Jones. 2022. The role of valence, dominance, and pitch in perceptions of artificial intelligence (ai) conversational agents’ voices. *Scientific Reports*, 12(1):22479.

Tuva Lunde Smestad and Frode Volden. 2019. Chatbot personalities matters: improving the user experience of chatbot interfaces. In *Internet Science: INSCI 2018 International Workshops, St. Petersburg, Russia, October 24–26, 2018, Revised Selected Papers 5*, pages 170–181. Springer.

James S Uleman, S Adil Saribay, and Celia M Gonzalez. 2008. Spontaneous inferences, implicit impressions, and implicit theories. *Annu. Rev. Psychol.*, 59:329–360.

Alessandro Vinciarelli and Gelareh Mohammadi. 2014. A survey of personality computing. *IEEE Transactions on Affective Computing*, 5(3):273–291.

Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017a. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*.

Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A Saurous. 2017b. Uncovering latent style factors for expressive speech synthesis. *arXiv preprint arXiv:1711.00520*.

Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif A Saurous. 2018. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. *arXiv preprint arXiv:1803.09017*.

Junichi Yamagishi, Christophe Veaux, Kirsten MacDonald, et al. 2019. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit (version 0.92).

Ya-Jie Zhang, Shifeng Pan, Lei He, and Zhen-Hua Ling. 2019. Learning latent representations for style control and transfer in end-to-end speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6945–6949. IEEE.

A Appendix A: Sentences used in the listening tests

Sentence ID	Sentence
180	He also defended the company's policy of releasing new services and tools to users before they were finished products.
189	No charges were made, but two men have been thrown off the programme.
205	After a gruelling ten minute phone interview the reporter had a new job.
216	There is controversy around these findings: some people have tried to replicate them, although not using exactly the same methods, and got different results.
259	Even as voters drift away from party politics, they can get very heated and involved in the politics of the high street.

Table 2: Selected sentences for listening tests