Speech Synthesis for the Generation of Artificial Personality

Matthew P. Aylett, Alessandro Vinciarelli, *Member, IEEE*, and Mirjam Wester

Abstract—A synthetic voice personifies the system using it. In this work we examine the impact text content, voice quality and synthesis system have on the perceived personality of two synthetic voices. Subjects rated synthetic utterances based on the *Big-Five* personality traits and naturalness. The naturalness rating of synthesis output did not correlate significantly with any Big-Five characteristic except for a marginal correlation with openness. Although text content is dominant in personality judgments, results showed that voice quality change implemented using a unit selection synthesis system significantly affected the perception of the Big-Five, for example tense voice being associated with being disagreeable and lax voice with lower conscientiousness. In addition a comparison between a parametric implementation and unit selection implementation of the same voices showed that parametric voices were rated as significantly less neurotic than both the text alone and the unit selection system, while the unit selection was rated as more open than both the text alone and the parametric system. The results have implications for synthesis voice and system type selection for applications such as personal assistants and embodied conversational agents where developing an emotional relationship with the user, or developing a branding experience is important.

Index Terms—Personality, automatic personality perception, automatic personality recognition, automatic personality synthesis

1 Introduction

According to the Roman writer Publilius Syrus, "*speech is a mirror of the soul; as a man speaks, so is he*".[1] Roughly twenty centuries after this sentence was written, this intuition has been investigated scientifically in personality psychology. Speech, especially when it comes to nonverbal aspects, has been shown to be a personality marker, i.e., a physical trace of a speaker's personality [1]. Furthermore, speech appears to significantly influence the personality impressions that listeners develop about speakers, especially in zero acquaintance scenarios [2].

Cognitive sciences show that people attribute traits to others in less than a second after the first contact [3]. The process, mostly unconscious, is spontaneous and it takes place not only face-to-face, but also when people see or hear others in video and/or audio recordings [4]. Furthermore, recent research shows that the same effect extends to machines that display human-like features and behaviours like, e.g., social robots or embodied conversational agents (see Section 2 for more details). Therefore, it is possible to perform *Automatic Personality Synthesis* (APS) [5], to generate artificial behaviours that stimulate users to attribute predefined traits to the machines they interact with (e.g., to develop artificial agents for tutoring systems that are perceived as extrovert and conscientious).

This article focuses on speech-based APS and shows that listeners attribute different personality traits to a synthesizer depending on speech parameters, synthesis methodology and content of the message uttered. In particular, the experiments of this work show that voice quality change implemented using a unit selection synthesis system significantly affected the perception of the Big-Five. In addition a comparison between a parametric implementation and unit selection implementation of the same voices showed that parametric voices were rated as significantly less neurotic than both the text alone and the unit selection system, while the unit selection was rated as more open than both the text alone and the parametric system.

The main motivation behind the synthesis of personality-coloured speech is that impressions, while not necessarily being accurate, still drive people's behaviour and attitude towards others. Therefore, synthesizers that convey appropriate personality traits might be more effective in fulfilling their function. For example, users tend to think that speaking GPSs work better when their voices convey desirable traits (e.g., high extroversion and conscientiousness) [6]. Furthermore, there is evidence that users tend to favour machines that convey personality traits similar to their own (see Section 2). In this respect, the investigation of the interplay between speech parameters and attributed traits promises to be a crucial step towards increasing the satisfaction of interactive machines' users.

To the best of our knowledge, the most important novelties of this work are a systematic analysis of the relationship between synthetically generated voice quality modifications

[1] "The Moral Sayings of Publilius Syrus: A Roman Slave", first Century B.C.

- M.P. Aylett and M. Wester are with the School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom. E-mail: matthewaylett@gmail.com, mwester@inf.ed.ac.uk.
- A. Vinciarelli is with the Computing Sciences, University of Glasgow, Glasgow G12 8QQ, United Kingdom. E-mail: vincia@dcs.gla.ac.uk.

Manuscript received 16 Mar. 2017; revised 30 Aug. 2017; accepted 8 Oct. 2017. Date of publication 16 Oct. 2017; date of current version 29 May 2020.
(Corresponding author: Matthew Aylett.)
Recommended for acceptance by Y. H. Yang.
Digital Object Identifier no. 10.1109/TAFFC.2017.2763134

and the perception of personality, as well as a comparison of the effects of different state-of-the-art approaches to speech synthesis on perceived personality.

The rest of this paper is organized as follows: Section 2 presents an overview of previous work in APS and speech synthesis, Section 3 details the methods for assessing and measuring personality, Section 4 describes the textual and speech based stimuli used in the experiments, and Section 5 outlines our key research questions. Section 6 describes the overall experimental setup, with Section 7 presenting the results from the naturalness pre-evaluation, and Section 8 the results from APS. The final Section 9 discusses the results and draws some conclusions.

2 Previous Work

An extensive survey of Personality Computing, the domain revolving around technologies dealing with human personality, is available in [5]. This section focuses on APS, the different approaches to speech synthesis, and how synthetic voices can be modified to convey different affective aspects.

2.1 Speech-Based Automatic Personality Synthesis

The core-idea of APS is that the attribution of traits to others is a spontaneous, mostly unconscious, cognitive phenomenon and it takes place not only when people meet human others, but also when they interact with machines displaying human-like features and behaviours (e.g., social robots, artificial agents, etc.). Therefore, most APS works investigate whether the generation of specific artificial cues leads people to attribute predefined specific traits to a machine they interact with.

Psychologists show that vocal behaviour is the individual channel that covariates most with whole person judgments like the attribution of personality traits (compared to facial expressions and bodily postures) [2]. Hence, it is not surprising to observe that previous APS work often aims at personality-coloured speech synthesis, both as a single modality [6], [7], [8], [9], [10] and in conjunction with other behavioural channels [11], [12], [13], [14], [15], [16].

The earliest observations showing that human listeners tend to attribute personality traits to synthetic voices were proposed in [7]. However, no attempt was made in [7] to manipulate speech parameters with the goal of influencing the personality traits attributed to a synthetic voice. The experiments in [6], [8] show that increasing pitch, frequency range and speaking rate tends to lead to the attribution of higher Extraversion scores. Furthermore, the experiments show that listeners tend to prefer synthetic voices that sound closer in terms of personality traits (the *similarity-effect*). Later approaches [9], [10] extended the number of speech parameters (pitch, pitch range, intensity and speaking rate) that can be controlled to make a voice sound high or low along each of the Big-Five traits [17].

More recent work [18] using synthetic voices, examined the effect of fillers (in this case *um, uh, I mean, like* and *you know*) on the perceived personality of an artificial voice. The work found that including disfluencies affected the perceived personality by making the voice sound more neurotic, less conscientious, less open and less extrovert.

In several studies, speech synthesis is used in conjunction with other modalities to endow embodied conversational agents [11], [12], [13] or robots [14], [15], [16] with perceivable personality traits. The experiments of [11], [12] show that the joint manipulation of paralanguage (length and frequency of pauses, hesitations, etc.), body movements (gestures, fidgeting, etc.), gaze behaviour and facial expressions allows one to influence the traits attributed to an artificial agent. In a similar vein, the experiments proposed in [13] show that the joint manipulation of prosody styles (fast, loud and high-pitched versus slow, soft and low-pitched), gaze patterns and eyebrow movements influences the attributed Extraversion scores.

The experiments of Tapus and colleagues [14], [15] investigate the similarity effect observed in [6], [8] with social robots. In this case, prosodic features (pitch and volume) were manipulated in conjunction with proxemic cues and verbal content to make a robot appear more or less extrovert. As in the case of [6], [8], users manifested the tendency to like robots that they perceive closer in terms of their own personality. Similar experiments were proposed in [16], but no similarity-effect was observed. In this case, the artificial cues of the robots included speaking activity-waiting or not for the subject before talking, lexical choices and gaze behaviour.

2.2 Speech Synthesis

The state-of-the-art in speech synthesis has significantly changed since Nass's pioneering work [8] in 2001. Quality has significantly improved with commercial systems often being mistaken for natural voices when used in a neutral context, such as for announcements and navigational systems. There are currently two dominant approaches:

1) Unit Selection: New speech is generated by taking segments (or units) of these recordings, cutting them up and sticking them back together in a different order [19], [20], [21].
2) Parametric synthesis: A statistical model, typically using hidden Markov models or deep neural nets, is created from the recorded speech. At synthesis, the model generates parameters for a vocoder in order to create speech. [22], [23].

Interest in modifying speech synthesis to affect the perception of emotion and to support more expressive speech has been an ongoing area of research with work in APS a key element of the research effort. In this previous work [6], [8], [14], [15], prosodic modification was constrained to pitch, amplitude and rate only. Voice quality, the perceptual colouring of a person's voice [24], is also an important factor in the perception of emotion in speech [25] as well as having a significant effect on the style and uniqueness of a speaker's voice [26]. However, unlike speech rate and pitch, which can be modified relatively easily using digital signal processing techniques such as *Pitch Synchronous Overlap Add* (PSOLA), modifying voice quality is more difficult, especially if it is important to retain naturalness.

In unit selection, rather than modifying speech to create different voice qualities, an alternative approach is to record different voice qualities and use them directly during concatenative synthesis. This approach has been applied to diphone synthesis [27] and has been extended to unit selection in the CereVoice system which uses pre-recorded voice quality sub-corpora in unit selection [28]. This is different from other unit selection approaches which have instead examined the use of sub-corpora of specific emotions, e.g., [29] where Happy, Angry and Neutral sub-corpora were incorporated into an emotional voice in Festival. Focusing on voice quality rather than specific emotions allows a combination of *Digital Signal Processing* (DSP) techniques and unit selection to craft a more varied and subtle set of speech styles [30].

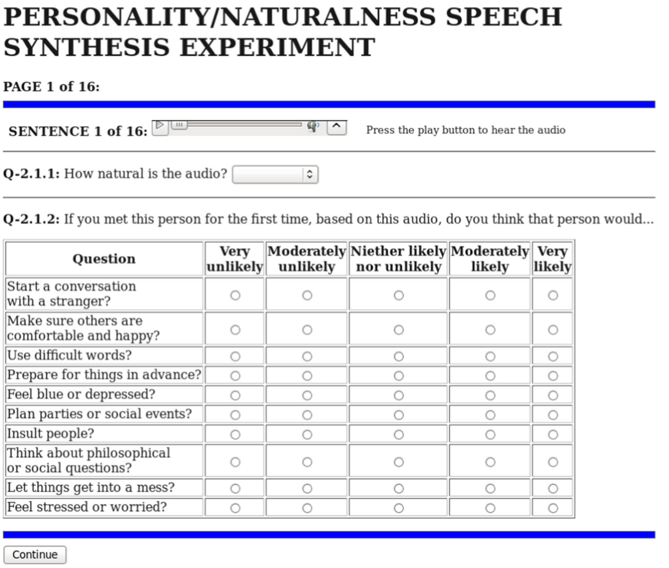
In parametric synthesis the speech is parameterised, modified and completely recreated using vocoding techniques (See [31], [32] for a review of common techniques). In current systems the vocoding process is lossy, resulting in degraded voice quality, and in some cases a perception of a buzzyness. In addition the parameters are modelled either using statistical techniques [22] or deep neural networks [23], [33]. This will typically merge frames with varying parameter values resulting in an mean value which often shows less variation removing phonetic and prosodic detail. This can result in speech sounding *dull* or *muffled*. Advances have been made to tackle these underlying problems, and although current parametric systems have previously been rated as less natural than unit selection [34] they are improving. The rapid development and innovation occurring in this field make it a challenge to choose or generalise from any one system. The system used in this study is very close to the HTS2007 open source distribution [22] with STRAIGHT analysis [35] and MLSA vocoding [36]. Many recent systems share similarities with HTS2007 and as such it acts as an excellent baseline. Many researchers in the field have access to this baseline which will allow them to both interpret and build on the results presented here.

3 Personality and Its Measurement

Personality is a psychological construct designed to capture stable individual characteristics that can explain and predict, possibly in quantitative terms, observable behavioural differences [37]. The literature proposes a wide spectrum of personality models (see, e.g., [38], [39] for extensive surveys). However, *trait*-based models are those that appear to be the most successful when it comes to the prediction of important life aspects such as, e.g., "*happiness, physical and psychological health, [...] quality of relationships with peers, family, and romantic others [...] occupational choice, satisfaction, and performance, [...] community involvement, criminal activity, and political ideology*" [40]. The main peculiarity of trait based models is that they represent personality as a point in a D-dimensional space where every dimension corresponds to a salient psychological phenomenon [41]. For this reason, the experiments of this work are based on the *Big-Five* model, the most popular and widely applied trait-based representation in both personality psychology and personality computing [5].

The dimensions of the Big-Five model correspond to the following aspects of observable behaviour [42]:

- *Extraversion*: Active, Assertive, Energetic, Outgoing, Talkative, etc.
- *Neuroticism*: Anxious, Self-pitying, Tense, Touchy, Unstable, Worrying, etc.
- *Conscientiousness*: Efficient, Organized, Planful, Reliable, Responsible, Thorough, etc.
- *Agreeableness*: Appreciative, Kind, Generous, Forgiving, Sympathetic, Trusting, etc.
- *Openness*: Artistic, Curious, Imaginative, Insightful, Original, Wide interests, etc.



PERSONALITY/NATURALNESS SPEECH SYNTHESIS EXPERIMENT

PAGE 1 of 16:

SENTENCE 1 of 16: Press the play button to hear the audio

Q-2.1.1: How natural is the audio?

Q-2.1.2: If you met this person for the first time, based on this audio, do you think that person would...

Question	Very unlikely	Moderately unlikely	Niether likely nor unlikely	Moderately likely	Very likely
Start a conversation with a stranger?	○	○	○	○	○
Make sure others are comfortable and happy?	○	○	○	○	○
Use difficult words?	○	○	○	○	○
Prepare for things in advance?	○	○	○	○	○
Feel blue or depressed?	○	○	○	○	○
Plan parties or social events?	○	○	○	○	○
Insult people?	○	○	○	○	○
Think about philosophical or social questions?	○	○	○	○	○
Let things get into a mess?	○	○	○	○	○
Feel stressed or worried?	○	○	○	○	○

Fig. 1. *Form for assessing Big-Five subject response to text or audio stimuli.*

The adjectives in the lists above are the tendencies associated to every trait. An individual is said to be *high* in a trait when her or his behaviour follows the tendencies associated to the trait itself. Otherwise, the individual is said to be *low* in the trait and her or his behaviour will be better described by the opposites of the adjectives above.

The position of an individual along the various traits can be measured with personality assessment questionnaires (see [37] for an extensive survey). These include a variable number of items - questions or statements about observable behaviour - associated to Likert scales. The questionnaire adopted in this work is the *Newcastle Personality Assessor* (NPA) [43]. Fig. 1 shows the online interface through which the assessments have been collected in this work. The answers can be mapped into numbers (e.g., answer "*Very Unlikely*" corresponds to -2 while answer "*Very Likely*" corresponds to 2) that can be appropriately summed and lead to a different *score* for every trait. The value of the scores measures how high or low an individual is along a given trait.

In this work, the questions of the NPA are expressed in third person because the experiments focus on the traits that the listeners attribute to synthetic speech. The reason behind this choice is twofold. On the one hand, self-assessments are not possible in the case of machines because these can convey personality impressions, but do not have a personality [5]. On the other hand, the traits that listeners attribute to synthetic speech are important because users tend to think that machines conveying a positive impression work better irrespectively of their actual performance [6].

In the experiments of this work, the NPA is filled by multiple assessors for the same speech sample. Hence, it is necessary to measure the agreement between different ratings assigned to the same stimulus. Following an approach typical of personality psychology [44], [45], the agreement (Cronbach's Alpha) is measured as fraction of ratings variance shared across assessors [44], [45]. The fraction can be

TABLE 1
About Myself (AM) Text

AM001: I like to bring order to everything I do. I think the details and facts are often missed by others, and I like to work based on concrete results. If faced by a problem I like to look at it logically and make a decision based on the specific problems at hand.
AM002: I'm good at encouraging others to work with each other and cooperate effectively. I think that if you look after and help colleagues you get the best out of them. If you do good work then the people around you will also become more motivated.
AM003: I'm great at getting people to work with each other and sorting out misunderstandings and conflict. If you concentrate on the common ideas and values you all share, you can find real insight and discover new possibilities.
AM004: I like to plan, provide direction and make sure everyone knows what their responsibilities are. I think its very important to be a good example to others, to be committed, and to work hard on doing things the right way to achieve your goals.
AM005: I'm good at encouraging others to contribute effectively. I think its important to enjoy your work and to be enthusiastic about what you do. If you can make work enjoyable then the people around you will also become more motivated and happier.
AM006: I'm great at helping others plan and cooperate to get things done. Its important to work out what can be done and the best way to do it. I like to work with others and help everyone come together behind a project.
AM007: I'm good at developing new strategies and approaches to a problem and I think being committed to what you do is very important. I love innovation and overcoming challenges.

TABLE 2
Speed Dating (SD) Negative and Positive Text

negative
NE001: I'm from West London, which is a part of town I really dislike. It was a real pain in the arse to get here, I can tell you. I used to like film until Hollywood ruined them all.
NE002: What a mess this place is. I'm sure the organiser has got it in for me. I've always had problems with people, either because they are stupid or jealous of me.
NE003: You don't seem to have made much effort. Though given the losers here I'm not surprised. You'd probably be happier watching TV at home.
positive
PO001: I'm from a lovely little suburb with lots of trees and parks. The train is very quick and it was no trouble to get here. I love going to the beach and spending time with my friends.
PO002: They've done a brilliant job at redecorating this bar. The people running it have been really nice to me. I always get on with people we have so much to share with each other.
PO003: I must say you are looking very nice tonight. Everyone is very nicely dressed and seem so successful. I expect you are looking forward to coming again.

measured by performing a two-way Analysis of Variance (ANOVA) of the ratings [46]. According to a survey of the literature [44], the average value of this fraction in experiments like those presented in this work is low (0.32 for Extraversion, 0.07 for Neuroticism, 0.13 for Conscientiousness, 0.03 for Agreeableness, and 0.07 for Openness). However, these levels of agreement are considered acceptable as long as they are statistically significant, i.e., as long as the raters agree beyond chance [44], [45].

4 The Stimuli

This section describes design and collection of the synthetic speech utterances, the *stimuli* hereafter, used in the experiments of this work. Following the methodologies of personality psychology, the stimuli have been assessed in terms of traits by a pool of human listeners, referred to as *assessors* in the rest of the article. Section 4.1 focuses on the design of the texts, Section 4.2 shows how the texts have been synthesized and Section 7 analyses the interplay between attribution of personality traits and judgment of *naturalness*, one of the metrics most commonly adopted to evaluate speech synthesis technologies.

4.1 Textual Stimuli

Tables 1 and 2 show the 13 texts that have been uttered with the synthetic voices used in this work (see Section 4.2). The textual stimuli are grouped into three main subsets, namely *About Myself* (AM), *Negative* (NE) and *Positive* (PO). Subset AM includes first person statements about professional life and attitude towards others in a working environment. Subset PO includes first person statements revolving around positive aesthetic judgments (about both people and places) and attitude towards others. Subset NE includes statements about the same topics as PO, but the tone is negative. The PO and NE subsets are taken from an imagined speed dating scenario.

Language psychology suggests that verbal messages do not convey only content, but also information about social and psychological phenomena: "*Words and language [. . .] are the very stuff of psychology [. . .] the very medium by which cognitive, personality, clinical, and social psychologists attempt to understand human beings.*" [47]. For this reason, the 13 textual stimuli adopted in the experiments (see Table 1) have been rated in terms of the Big-Five personality traits. The goal is, on the one hand, to verify that the texts attract a sufficiently wide spectrum of personality judgments and, on the other hand, to investigate the interplay between text and rendering (see Section 8).

The texts were chosen in order to give a broad coverage of personality perception. To evaluate how well this was achieved, 10 subjects read each text and evaluated them on the Big-Five scale. Fig. 2 shows the average trait scores attributed to the texts of the three subsets. Agreement across assessors was relatively high compared to previous work [44] (Cronbach's Alpha: 0.40 for Extraversion, 0.83 for Neuroticism, 0.60 for Conscientiousness, 0.66 for Agreeableness, and 0.60 for Openness).

Neutral texts were avoided because it is hard to define what neutral might be and also how we might expect a subject to respond to neutral text based on the Big-Five. For example answering the question "Would this person

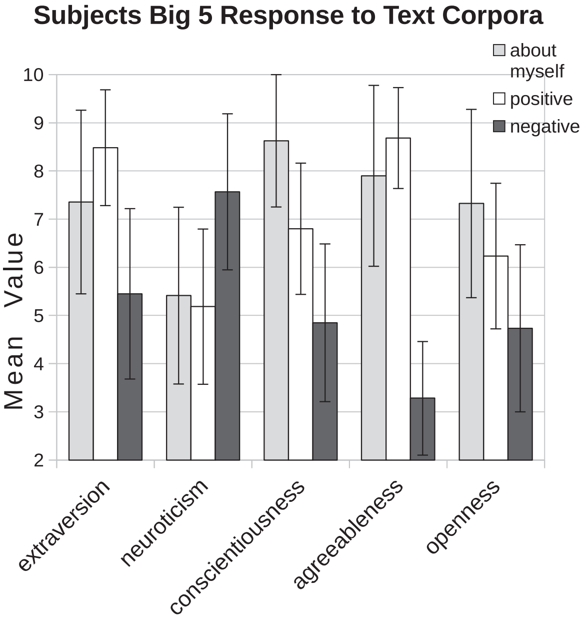


Fig. 2. *Subjects' Big-Five response to about myself and speed dating text corpora.*

start a conversation with a stranger?" is challenging based only on audio rendering the text "The light bulbs are three pounds forty".

4.2 Synthetic Speech Materials

Each textual stimulus has been rendered in eight different ways according to the scheme depicted in Fig. 3. The result is a set of $13 \times 8 = 104$ synthetic speech stimuli corresponding to all technically possible combinations of gender (*male* or *female*), synthesis approach (*parametric* or *unit selection*) and voice quality (*neutral*, *lax* or *tense*). The parametric voices were obtained by following the approach described in [22] with STRAIGHT [35] analysis, full context model tree building, and MLSA vocoding [36].

Within parametric approaches, voice quality (VQ) has generally not been directly modelled because of the difficulty of decomposing voicing into a set of parameters that could form a statistical model (e.g., [22]). Recent work is considering how to effectively model voice quality within parametric speech synthesis, e.g., [48], [49], [50], [51], [52], however this is atypical and in the work presented here the parametric system that is evaluated is very close to the design described in Zen et al. [22]. The approach does not directly model voice quality, but includes mixed excitation where noise representing frication is added to a pulse train to generate speech. Therefore in this work only neutral audio data was used to train the parametric system and no attempt was made to build Lax/Tense VQ models. In Section 9, we discuss how this work could be extended to parametric systems which model voice quality more effectively.

In contrast, unit selection approaches, by using original recorded speech, can retain voice quality. However, in unit selection, a set of features are computed for the target text in order to select appropriate units from the database, for example stress, phrase position, phonetic context. If there are few units matching these targets in the database, the system will be forced to either concatenate units which lack a smooth transition, or to select units which do not match the required specification, e.g., too short, wrong stress, etc. This will cause a perceptible error in the synthesis, a concatenation error, which reduces naturalness. The CereVoice[2] system used in this experiment is a good example of the state-of-the-art in unit selection and is used in many commercial applications from reading exam papers to students to producing multilingual announcements at Gatwick airport. As such, the neutral system is a fair representative of high grade unit selection systems in general. In order to create tense and lax synthesis output, the unit selection system blends substantial numbers of units based on tense and lax audio sub-corpora within the voice database. For neutral output these same units are prevented from being selected.

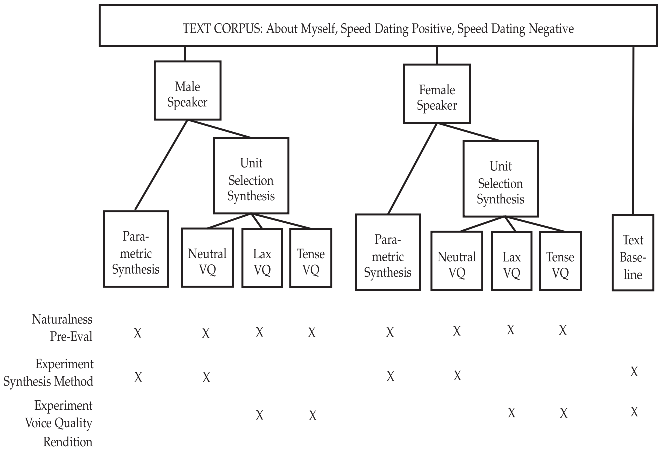

TEXT CORPUS: About Myself, Speed Dating Positive, Speed Dating Negative
Male Speaker
Female Speaker
Unit Selection Synthesis
Unit Selection Synthesis
Para-metric Synthesis
Neutral VQ
Lax VQ
Tense VQ
Para-metric Synthesis
Neutral VQ
Lax VQ
Tense VQ
Text Base-line
Naturalness Pre-Eval X X X X X X X X
Experiment Synthesis Method X X X X X
Experiment Voice Quality Rendition X X X X X

Fig. 3. The scheme shows how each text stimulus from the corpora has been rendered in 8 different ways. In the case of parametric synthesis, there are two different versions of the same textual stimulus, namely *male* and *female*. In the case of the unit selection synthesis, there are 6 different versions, 3 *female* (*neutral*, *lax* and *tense*) and 3 *male* (*neutral*, *lax* and *tense*). The crosses show which stimuli are used in the three experiments: Naturalness pre-evaluation Section 7, synthesis method Section 8.1, and voice quality rendering Section 8.2.

The male speech stimuli were synthesized using 318 minutes of neutral material, 46 minutes of tense voice quality data and 40 minutes of lax voice quality data. The female speech stimuli were synthesized using 278 minutes of neutral material, 33 minutes of tense voice quality data and 29 minutes of lax voice quality data. For both male and female speech stimuli, the material consists of individually read sentences chosen to ensure good coverage of phonetic and prosodic variation. Texts described in Section 4.1 were used for synthesis output only. No matching audio was recorded or added to audio used to build voices.

Overall agreement across assessors for the speech stimuli are similar (and in general slightly higher) than for the initial text materials. (Cronbach's Alpha: 0.68 for Extraversion, 0.73 for Neuroticism, 0.63 for Conscientiousness, 0.60 for Agreeableness and 0.70 for Openness).

5 Research Questions

Perceived naturalness is a key aspect of speech synthesis that affects listener evaluation (even when they are asked to score, for instance, prosody or pleasantness etc.). Previous work has shown that generally parametric systems are not

[2] https://www.cereproc.com/en/products/sdk

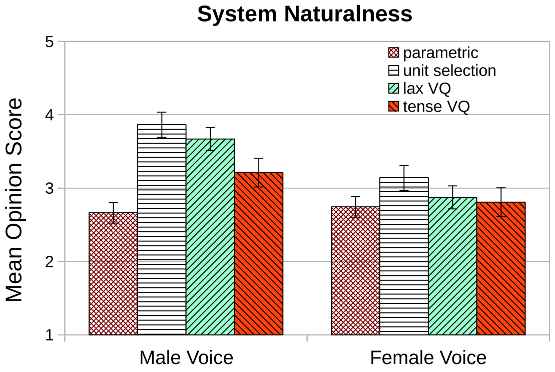


Fig. 4. *Naturalness*.

regarded as natural sounding as unit selection systems [34]. In addition, as different synthetic voices are built from different voice corpora, and digital signal analysis of different source voices can vary in quality, naturalness variation is also common between different synthetic voices built with the same system. To remove this potentially confounding effect of perceived naturalness caused by system differences we address the question: *"Are some, or all, Big-Five ratings directly related to naturalness ratings?"* (Section 7).

We then explore the differences in perceived personality for text rendered using a parametric speech synthesis system versus a unit selection system and the effect of different synthesised voice qualities. We have seen that voicing difference can affect the perception of emotion in a synthetic voice [27], [29], [30], and that this effect is independent of the effect of text content [53]. We compare unit selection systems based on tense and lax voices thereby quantifying how voice quality might influence perceived personality. The research questions these comparisons will answer are *"Does synthesis have an effect on the perceived personality of text when read to different renderings (i.e., different speech synthesis system types and voicing variation)? What are these effects?"* (Sections 8.1 and 8.2).

In order to consider the effect of various renderings of the text using different synthesis systems we obtained a baseline of the perceived personality of the text on its own (see Section 4.1). Thus we measure *change* in personality that different speech renditions produce as well as any interaction with textual material. Fig. 3 shows how we structured these different experiments across the speech and textual stimuli.

6 Experimental Setup

Subjects were asked to rate the naturalness of the speech they heard on a scale between 1 (Bad) and 5 (Excellent), and to answer the 10 questions from the NPA [43]. Each short spoken paragraph was presented using a web interface. Subjects could play the speech as many times as they wished, all questions had to be answered before moving on to the next stimuli. Fig. 1 shows a screen-shot of the web interface.

Subjects were organised into eight blocks so that no subject heard the same text spoken by more than one system. A total of 35 subjects took the experiment with a roughly even distribution across the eight blocks. Subjects had normal hearing and were native English speakers. The raw data collected from each subject was a set of non-parametric opinion scores in the form of answers from 1-5 to 10 questions chosen to rate Big-Five personality type, and a naturalness rating of the synthetic speech varying from 1 (not natural) to 5 (natural).

7 Naturalness Pre-Evaluation

A by-materials repeated measures ANOVA for overall naturalness was carried out with system type (Parametric, Unit Selection, Tense VQ, Lax VQ) as a within-materials factor and voice source (whether the audio was spoken by the female or male voice) and corpus type (whether the text was part of the *About Myself* corpus or the *Speed Dating* negative or positive corpus) as between materials factors.

Both system type and voice source factors produced significant results in this analysis (system type: F(1, 28)=52.968, p < 0.001), (gender: F(1, 28)=11.925, p < 0.005). See Fig. 4. The two source voices used in the study show a difference in naturalness. The female voice is less natural overall compared to the male voice across both parametric and unit selection systems. The biggest difference is for neutral unit selection (mean naturalness score of 3.1 versus 3.9). This large difference is the result of three factors: 1. Inherent differences in the ability of the source voice talent to produce consistent and clearly pronounced data; 2. Gender effects, Stevens et al. [54] shows different systems can be affected by gender differently, with female voices rated as less intelligible. This could be caused by the different genders presenting different challenges to signal processing such as formant tracking [55]; 3. Differences in the source database size. For the neutral voices the female speaker had 40 minutes less data (278 minutes of speech versus 318 minutes of speech in the male voice). Unit selection systems are very sensitive to database size with results in the Blizzard challenge causing up to 0.6 degradation in MOS for the same voice source [56].

We see no effect of text corpus - suggesting that there is no inherent difference in the quality of the synthesis across the corpora. This is important because synthesis systems based on source corpora will produce more natural results the closer the input text matches the corpora used. For this work, all three corpora are equally as challenging to synthesise naturally and no bias is caused by any of the three corpora matching the underlying corpus used in synthesis.

There is a significant difference in overall quality with the neutral concatenative system rated significant more natural than the parametric system. This follows consistent results from all previous Blizzard challenges [34]. The neutral system is also rated as more natural than the lax and tense systems. This variation is caused by less available data with appropriate voicing requiring concatenation of mis-matching segments in some circumstances and matches results obtained in Aylett et al. [53].

If we examine the correlation between Big-Five factors and naturalness across all system types (Table 3) we see only a weak positive relationship between naturalness and openness ($p < 0.05, r = 0.249$) predicting just over 6 percent of the variance.

These results support the hypothesis that changes in perceived personality presented in the next section are caused

TABLE 3
Naturalness Correlation with Big-Five Over All Data

Big-Five trait	Pearson's r
Extraversion	−0.123
Neuroticism	0.040
Conscientiousness	0.065
Agreeableness	−0.073
Openness	0.249

by underlying differences in the voice quality and synthesis style produced by the systems rather than the underlying impact of naturalness differences perceived by subjects. Furthermore, these results suggest that personality assessment effectively complements naturalness evaluation of synthetic speech (See Section 9 for a more detailed discussion of Big-Five as a potential standard for synthetic speech evaluation.

8 Experiments and Results

In Section 4.1, a strong effect of text was shown. What we are interested in investigating here is how speech synthesis alters the perception of text when read to different renderings.

Specifically, how does the subject's perception of the personality in the voice change by synthesising the text:

1) Using the different synthesis approaches (unit selection versus parametric)?
2) By altering voice quality in the synthesis system (tense versus lax)?

In order to avoid a type I error we carry out an initial by-materials MANOVA analyses over Big-Five scores averaged by subject responses to allow a parametric analysis based on the central limit theorem (mean of means). A significant result for Wilks' Lambda then licenses univariate ANOVA analysis with a following significant F score licensing post hoc tests. Comparison between synthesis method (unit selection versus parametric) and synthesis voice quality rendering (tense versus lax) are investigated separately. As all post hoc tests involve three means (across two synthesis methods or two synthesis renderings together with a text only baseline) a Fisher's protected least significance (LSD) post hoc test is used. The LSD test can cause type I errors for four or more means but both analytical [57] and empirical studies [58] have shown them to be robust with three means only.

In order to remove a spurious interaction between voice source and system caused by the text baseline results being identical across voice source, any significant voice source effects are re-analysed without the text system baseline and results are shown for synthesis only. All graphs show standard error and modified means based on the repeated measures analysis.

A by-materials repeated measures MANOVA across all five personality trait scores was carried out with *synthesis* method and rendition (the text analysis results *Text* and synthesis renderings, *Unit-Selection*, *Parametric*, *Lax* voice synthesised quality, and *Tense* voice quality) as within-materials factors and *corpus* type (*About Myself*, *SD positive*, *SD negative*) and *voice source* (*male/female*), as between-materials factors.

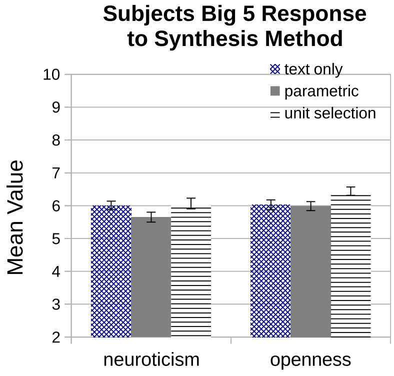


Fig. 5. *Mean neuroticism and openness by synthesis method. Error bars show standard error* ± 1.

Corpus type and *voice source* both showed a significant effect (*corpus*: $Wilk's\ Lambda = 0.008, F(10,32) = 31.895,\ p < 0.001$, *voice source*: $Wilk's\ Lambda = 0.383, F(5,16) = 5.517,\ p < 0.001$), with *synthesis* method and rendition showing a within-materials effect (*system*: $Wilk's\ Lambda = 0.457, F(20,316) = 3.368, p < 0.001$).

We would expect *voice source* to have a significant effect as different voices are well known to be perceived differently in terms of personality [4]. We also expect significant effects by *corpus* as the text was chosen to explicitly produce these differences and the highly significant effect on perceived personality by corpus group was a desired effect. Of more interest are the significant interactions between *system* type and *corpus* ($Wilk's\ \ Lambda = 0.379, F(40,334) = 2.082, p < 0.001$).[3]

These results show that the perception of personality is altered by *system* type, and alters differently by *corpus type*. We will now explore these effects in more detail by examining the effect of *synthesis method*-unit selection versus parametric approaches, and then looking at effects of voice quality modification on *synthesis rendition*.

8.1 The Effect of Synthesis Method, Unit Selection versus Parametric

Following the significant MANOVA a set of univariate ANOVAS are carried out with a by-materials repeated measures across all five personality trait scores with *synthesis method* (the text analysis results *Text* and synthesis renderings, *Unit-Selection*, and *Parametric*) as within-materials factors and *corpus* type (*About Myself*, *SD positive*, *SD negative*) and *voice source* (*male/female*) as between-materials factors.

8.1.1 *Synthesis Method*

Synthesis method was significant for neuroticism ($F(2,40) = 3.893, p < 0.05$) and openness ($F(2,40) = 6.754$, $p < 0.005$). Post hoc tests for neuroticism show *Parametric* < *Text* and *Parametric* < *Unit-Selection* ($LSD: p < 0.05$). For openness *Text* < *Unit-Selection* ($LSD: p = 0.001$) *Parametric* < *Unit-Selection* ($LSD: p < 0.01$).

Fig. 5 illustrates the subjects' responses to text only, to parametric synthesis and to neutral unit selection synthesis.

[3] *system* type and *voice source* were also significant ($Wilk's\ Lambda = 0.584, F(20,253) = 2.227, p < 0.005$). However further investigation of the *voice source* interaction with the *Text* only baseline removed showed it was a spurious result caused by using the same baseline used both *voice sources*.

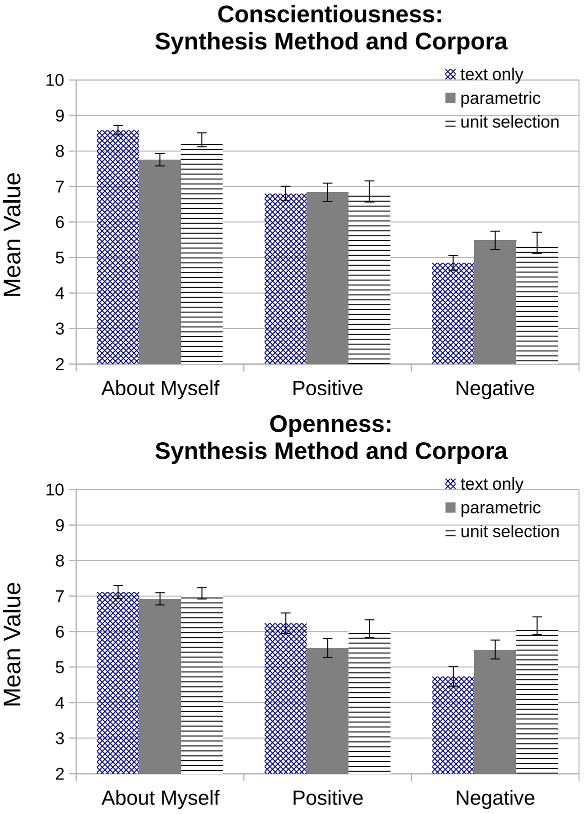


Fig. 6. *Interactions of synthesis method and corpus type. Means for conscientiousness and openness, error bars show standard error* ± 1.

Parametric synthesis is rated significantly less neurotic than either the original text, or the text synthesised using the unit selection system. For this Big-Five factor, we can interpret the result such that the unit selection synthesis successfully renders the text as it stands while the parametric system alters the perception of character.

Unit selection is seen as significantly more open than either the text alone, or the text synthesised with a parametric synthesiser. So in contrast, for openness, unit selection modifies the perception of character whereas parametric synthesis does not.

The findings above mean that there is a modulation effect between, in the case of Neuroticism, text and parametric synthesis and, in the case of Openness, text and unit selection. These results suggest that rendering a text with a synthesis approach tends to make personality impressions more socially appealing. In fact, both the decrease of Neuroticism scores and the increase of Openness scores go in the direction of a more socially desirable impression. In this respect, the results confirm the general observation that users tend to interact more comfortably with machines capable of displaying human-like behaviours [5], [59].

8.1.2 *Synthesis Method Interaction with Corpora*

Univariate ANOVAs show a significant interaction between *synthesis method* and *corpora* for conscientiousness $(F(4,40) = 5.447, p < 0.005)$ and also for openness $(F(4,40) = 7.311, p < 0.001)$. No significant effects have been observed for the other traits.

The Parametric method is rated as less conscientious than both Unit-Selection $(LSD: p < 0.025)$ and Text only $(LSD: p < 0.001)$ on the About Myself corpus, however for the more affective negative corpus parametric is rated more conscientious than Text only $(LSD: p < 0.025)$. For openness only the negative corpora gives significance to synthesis method differences with *Text* being rated as less open than *Parametric* $(LSD: p < 0.01)$ and *Parametric* less than *Unit-Selection* $(LSD: p < 0.025)$ and *Text* less than *Unit-Selection* $(LSD: p < 0.005)$. See Fig. 6.

Openness and Conscientiousness are socially desirable traits, i.e., people tend to consider individuals that are higher along these dimensions more positively. Both synthesis approaches tend to make the attributed traits more socially desirable in the case of the *Negative* corpus, the text subset for which the ratings along the two traits are the lowest. In contrast, the only modulation effect observed for the *About Myself* corpus, the subset for which the attributed traits are the most desirable, is that the parametric approach reduces the Conscientiousness ratings. The main pattern that emerges is that rendering a text with synthetic speech tends to smooth extreme personality judgments caused by differences in text content. One possible explanation is that a synthetic voice tends to make the stimulus more concrete, with less variation in personality than perceived from the text only where assessors are, in effect, imagining the vocal rendition of the text.

8.2 The Effect of Synthesised Voice Quality, Tense versus Lax

The experiments of this Section aim at investigating the interplay between voice quality of synthetic speech and attribution of personality traits. The expression *voice quality* accounts for "*perceptual and acoustic correlates of changes in the breathiness or pressed / laryngealized nature of the voicing sound source [...] harshness [...] soft / weak / whispered voice, falsetto and habitual settings of the vocal-tract configurations, such as a tendency toward an overall nasality quality*" [60]. The main reason for analysing the effects of voice quality is "*its function in communicating paralinguistic, linguistic and extra-linguistic information*" [25], including the stimulation of personality impressions [61].

Following the significant MANOVA in Section 8, a set of univariate ANOVAS are carried out with a by-materials repeated measures across all five personality trait scores with *synthesis rendition* (the text analysis results *Text* and synthesis renderings, *lax* voice quality, and *tense* voice quality) as within-materials factors and *corpus* type (*About Myself*, *SD positive*, *SD negative*) and *voice source* (*male/female*) as between-materials factors. In order to avoid multiple reporting errors, neutral unit selection audio stimuli from the first experiment are not included in this analysis.

8.2.1 *Synthesis Rendition*

Synthesis rendition has a significant effect on extraversion $(F(2,40) = 12.630, p < 0.001)$, conscientiousness $(F(2,40) = 5.045, p < 0.025)$, and agreeableness $(F(2,40) = 14.470, p < 0.001)$. No significant effects are observed for the other traits.

Post hoc tests reveal a complex set of significant results. In general the *Tense* voice is less agreeable $(LSD: Tense < Text - p < 0.001, Tense < Lax - p < 0.001)$ and more conscientious $(LSD: Tense > Text - p < 0.025, Tense >$

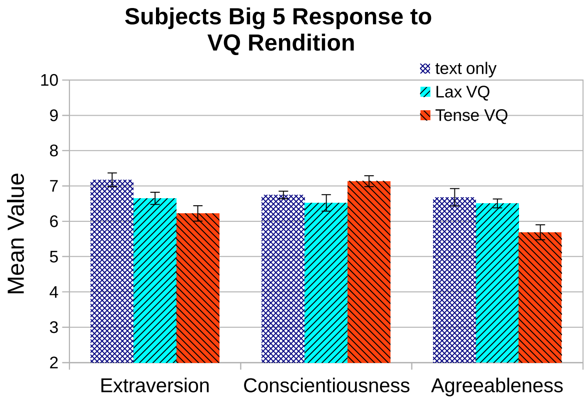


Fig. 7. *Mean of extraversion, conscientiousness and agreeable by synthesis rendition. Error bars show standard error* ± 1.

$Lax - p < 0.05$), while *Text* is perceived as more extravert than either *Lax* or *Tense* synthesis rendition ($LSD : Text > Lax - p < 0.025, Text > Tense - p < 0.001, Lax > Tense - p < 0.005$). See Fig. 7.

The overall pattern is that the tense voice tends to be perceived as the most conscientious, but the least extravert and agreeable. Vice versa, the textual only stimulus tends to be perceived as the lowest in Conscientiousness, but the highest in Extraversion and Agreeableness. The lax voice tends to follow the pattern of the textual only stimulus (the difference between the two is not statistically significant in the case of Conscientiousness and Agreeableness). These findings are in line with previous observations in the psychological literature under at least two main respects. The first is that voice quality has been shown to play a role in personality perception both in social psychology [61] and personality computing [5]. The second is that, on average, people that are perceived to be more competent (higher attributed Conscientiousness) tend to be perceived as socially colder (lower attributed Agreeableness and Extraversion) and, vice versa, people that are perceived to be less competent (lower attributed Conscientiousness) tend to be perceived as socially warmer (higher attributed Agreeableness and Extraversion) [62]. Thus, synthetic voices with non-neutral voice quality appear to reproduce the phenomenon known as *differentiation between task and social-emotional roles* [63], i.e., the tendency of people to be perceived as either *task-oriented* or *socially-oriented*. In particular, tense voices sound task-oriented while lax voices sound socially-oriented.

8.2.2 Synthesis VQ Rendering Significantly Interacts with Corpus Type

We see significant interaction between *corpus type* and *synthesis rendition* on the perception of conscientiousness ($F(2,40) = 6.279, p < 0.005$), agreeableness ($F(2,40) = 4.519, p < 0.005$) and openness ($F(2,40) = 4.913, p < 0.005$). No significant effects are observed for the other traits.

For conscientiousness in the *About Myself* corpus the *Tense* and *Lax* renditions are rated as less conscientious than *Text* ($LSD : Text > Lax - p < 0.005, Text > Tense - p < 0.025$). For the more affective negative corpus *Tense* is rated more conscientious than *Text* ($LSD : p < 0.005$). There are no significant difference for the positive corpus. See Fig. 8 top.

For agreeableness in the *About Myself* corpus all differences are significant ($LSD : Text > Lax - p < 0.05, Lax > Tense - p < 0.005, Text > Tense - p < 0.001$). In the *Positive* corpus *Text* is significantly more agreeable than *Lax* ($LSD : p < 0.05$) whereas the effect of the *Tense* VQ appears more varied, while in the *Negative* corpus only a significant effect remains between *Lax* and *Tense* ($LSD : Lax > Tense - p < 0.01$). See Fig. 8 middle.

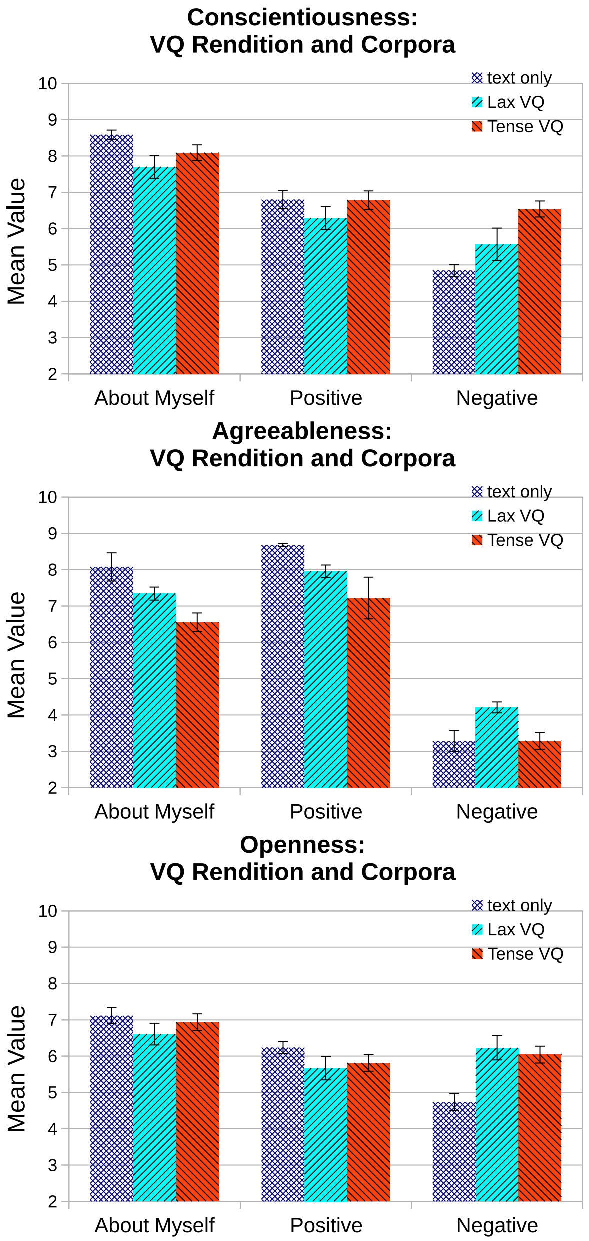


Fig. 8. *Mean of conscientiousness, agreeableness and openness by synthesis rendition and corpus type. Error bars show standard error* ± 1.

For openness there are no significant difference for the *About Myself* corpus, for the *Positive* corpus *Text* is just significantly more open than *Tense* ($LSD : Text > Tense - p < 0.05$) while for the *Negative* corpus *Text* is less open than *Lax* and *Tense* ($LSD : Text > Lax - p < 0.025, Text > Tense - p < 0.01$) See Fig. 8 bottom.

Synthesis VQ rendering has an impact on perceived traits, but also the process of synthesis, as with the results for different synthesis methods, appears to smooth extreme personality judgments caused by differences in text content. Again, possibly, a synthetic voice tends to

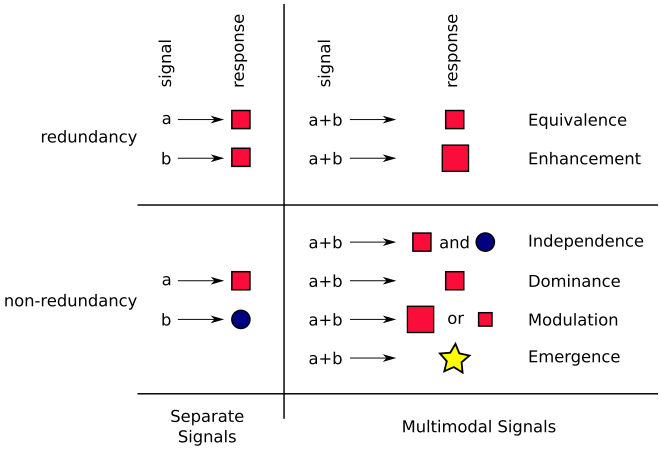


Fig. 9. *Partan Model: Classification of multimodal signals. Redundant signals are depicted above, non-redundant signals below. (Left) Responses to two separate components (a and b) represented by geometric shapes (the same shape indicates the same qualitative response; different shapes indicate different responses). (Right) Responses to the combined multi-modal signal*.

make the stimulus more concrete, with less variation in personality than perceived from the text only where assessors are, in effect, imagining the vocal rendition of the text.

9 Conclusion

This article focuses on the synthesis of personality-coloured speech. In particular, the experiments of this article show how different aspects of synthetic speech-text being uttered, synthesis approach and voice quality-interact to elicit the attribution of different personality traits. Furthermore, the experiments show that there is no statistically significant correlation between attributed traits and naturalness, a metric commonly adopted to evaluate the quality of synthetic speech. This suggests that naturalness and personality of synthetic speech are different phenomena that need to be addressed differently.

Three main factors have been considered, namely the content of the utterance being synthesised, the voice quality and the synthesis approach. The results can be interpreted in terms of the Partan model of multimodal communication [64], [65], a framework allowing one to interpret the interaction between multiple modalities in communication. According to such a model, the joint use of multiple modalities to convey the same message or impression can give raise to four main effects (see Fig. 9 adapted from [64]), namely *independence* (different modalities do not interact), *modulation* (different modalities interact to change the intensity of the message or impression), *dominance* (one of the modalities determines the message or impression being conveyed) and *emergence* (the modalities interact to produce a message different from the one conveyed by the modalities individually). The experiments of this work show that the only observed effect is modulation. In particular, the use of different voice qualities and or synthesis approaches tends to increase or decrease the trait scores attributed to a given text (at least for some traits). This is important because it suggests that nonverbal aspects of synthetic speech can enhance or attenuate the personality impressions conveyed by a text, but do not change them (at least in the range of the various factors considered in this work).

According to the experiments, the personality traits that people attribute to synthetic speech do not depend on naturalness. In other words, the traits that people attribute to synthetic speech are not an assessment of naturalness, but the result of the actual impression people develop about the artificial speaker. The only exception is openness-for this trait there is a statistically significant correlation with naturalness-but the effect is weak (naturalness explains only 6 percent of the variance in the attributed openness ratings). As a confirmation, several effects that have been observed appear to reproduce the results of speech perception in social psychology, including, e.g., the significant role of voice quality and the differentiation between task and social emotional roles. This result is important because it shows that the efforts aimed at making synthetic speech natural do not necessarily result into personality coloured voices. These latter can only be obtained by changing the characteristics of speech that actually appear to correlate with the attributed traits.

To the best of our knowledge, this is the first work that investigates in depth the interaction between multiple aspects of synthetic speech. Furthermore, this is the first work showing that traditional evaluation approaches based on naturalness do not necessarily account for the effectiveness of synthetic speech in conveying a certain personality impression. In this respect, this work provides actionable indications on how to allow synthetic speech to convey desired personality impressions. In this respect, future work will follow two main directions. The first is to verify how much the results of this work can be generalized and how the results of this work can help to generate artificial voices that convey impressions suitable for a particular application domain (e.g., producing a conscientious voice for an artificial tutor or an extravert voice for an artificial seller). The second is to test the effectiveness of personality-coloured speech in achieving specific goals like, e.g., to be more persuasive or to make an artificial agent more capable to establish empathic relationships with its users. This is in line with the indications of the literature showing that personality can act as a mediation variable, i.e., as a construct that can explain why certain behaviours are more effective than others to obtain certain interaction outcomes. In other words, the synthesis of personality-coloured speech can help to make artificial agents more effective at accomplishing the goals they are designed for.

Furthermore the corpus presented and evaluated here could help speech synthesis professionals to develop and evaluate future systems with reference to the baseline work carried out here. This is especially important in speech synthesis where systems are changing very quickly. In the space of a few years DNN based systems such as Merlin [66], Idlak Tangle [67] and Wavenet [68], have eclipsed much of the earlier HMM parametric work. Issues with voice modelling are still very much part of current synthesis research, and the framework and results we present here will help guide and support future work in APS with new systems as they become available.

ACKNOWLEDGMENTS

This research was funded by the Royal Society through a Royal Society Industrial Fellowship, and by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 (Aria VALUSPA). The work of A.Vinciarelli was supported by the EPSRC project EP/N035305/1.

REFERENCES

[1] K. Scherer, "Personality markers in speech," in *Social Markers in Speech*. Cambridge, U.K.: Cambridge Univ. Press, 1979, pp. 147–209.

[2] P. Ekman, W. Friesen, M. O'Sullivan, and K. Scherer, "Relative importance of face, body, and speech in judgments of personality and affect," *J. Personality Social Psychology*, vol. 38, no. 2, 1980, Art. no. 270.

[3] J. S. Uleman, S. A. Saribay, and C. M. Gonzalez, "Spontaneous inferences, implicit impressions, and implicit theories," *Annu. Rev. Psychology*, vol. 59, pp. 329–360, 2008.

[4] B. Reeves and C. Nass, *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[5] A. Vinciarelli and G. Mohammadi, "A survey of personality computing," *IEEE Trans. Affect. Comput.*, vol. 5, no. 3, pp. 273–291, Jul.–Sep. 2014.

[6] C. Nass and S. Brave, *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. Cambridge, MA, USA: MIT Press, 2005.

[7] J. Chin, "Personality trait attributions to voice mail user interfaces," in *Proc. Conf. Companion Human Factors Comput. Syst.: Common Ground*, 1996, pp. 248–249.

[8] C. Nass and K. M. Lee, "Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction and consistency-attraction," *J. Exp. Psychology: Appl.*, vol. 7, no. 3, pp. 171–181, 2001.

[9] M. Schmitz, A. Kruger, and S. Schmidt, "Modeling personality in voice of talking products through prosodic parameters," in *Proc. Int. Conf. Internet Things*, 2007, pp. 313–316.

[10] J. Trouvain, S. Schmidt, M. Schroder, M. Schmitz, and W. J. Barry, "Modeling personality features by changing prosody in synthetic speech," *presented at the Proc. Speech Prosody*, Dresden, Germany, 2006.

[11] E. De Sevin, S. Hyniewska, and C. Pelachaud, "Influence of personality traits on backchannel selection," in *Proc. Int. Conf. Intell. Virtual Agents*, 2010, pp. 187–193.

[12] M. McRorie, I. Sneddon, G. McKeown, E. Bevacqua, E. de Sevin, and C. Pelachaud, "Evaluation of four designed virtual agent personalities," *IEEE Trans. Affect. Comput.*, vol. 3, no. 3, pp. 311–322, Jul.–Sep. 2012.

[13] E. J. Krahmer, S. van Buuren, Z. Ruttkay, and W. Wesselink, "Audio-visual cues to personality: An experimental approach," in *Proc. AAMAS Workshop Embodied Agents Individuals*, C. Pelachaud, Z. S. Ruttkay, and A. Marriott, Eds., Melbourne, Australia, 2003, pp. 7–14.

[14] A. Tapus, C. Ţăpuş, and M. Mataric, "User—robot personality matching and assistive robot behavior adaptation for post-stroke rehabilitation therapy," *Intell. Serv. Robot.*, vol. 1, no. 2, pp. 169–183, 2008.

[15] A. Tapus and M. Mataric, "Socially assistive robots: The link between personality, empathy, physiological signals, and task performance," in *Proc. AAAI Spring Symp.*, 2008, pp. 133–140.

[16] S. Woods, K. Dautenhahn, C. Kaouri, R. Boekhorst, and K. Koay, "Is this robot like me? Links between human and robot personality traits," in *Proc. IEEE-RAS Int. Conf. Humanoid Robots*, 2005, pp. 375–380.

[17] R. McCrae, "The five-factor model of personality," in *The Cambridge Handbook of Personality Psychology*, P. Corr and G. Matthews, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 148–161.

[18] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality and disfluency," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3365–3369.

[19] A. Hunt and A. Black, "Unit selection in concatanative speech synthesis using a large speech database," in *Proc. Acoust. Speech Signal Process.*, 1996, pp. 192–252.

[20] R. A. Clark, K. Richmond, and S. King, "Festival 2 - build your own general purpose unit selection speech synthesiser," in *Proc. 5th ISCA Workshop Speech Synthesis*, 2004, pp. 147–151.

[21] J. Kominek, C. L. Bennet, B. Langer, and A. R. Toth, "The Blizzard challenge 2005 CMU entry - a method for improving speech synthesis systems," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 85–88.

[22] H. Zen, et al., "The HMM-based speech synthesis system (HTS) version 2.0," in *Proc. 6th ISCA Workshop Speech Synthesis*, Aug. 2007, pp. 294–299.

[23] H. Zen, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2013, pp. 7962–7966.

[24] J. Laver, "The phonetic description of voice quality," *Cambridge Stud. Linguistics London*, vol. 31, pp. 1–186, 1980.

[25] C. Gobl and A. N. Chasaide, "The role of voice quality in communicating emotion, mood and attitude," *Speech Commun.*, vol. 40, no. 1/2, pp. 189–212, Apr. 2003.

[26] M. Gordon and P. Ladefoged, "Phonation types: A cross-linguistic overview," *J. Phonetics*, vol. 29, no. 4, pp. 383–406, 2001.

[27] M. Scrhöder and M. Grice, "Expressing vocal effort in concatenative synthesis," in *Proc. Int. Congr. Phonetic Sci.*, 2003, pp. 2589–92.

[28] M. Aylett and C. Pidcock, "Adding and controlling emotion in synthesised speech," U.K. Patent GB2 447 263A, Sep. 10, 2008.

[29] G. Hofer, K. Richmond, and R. Clark, "Informed blending of databases for emotional speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2005, pp. 501–504.

[30] M. P. Aylett and C. J. Pidcock, "The CereVoice characterful speech synthesiser SDK," in *Proc. Artif. Intell. Simul. Behaviour*, 2007, pp. 174–178.

[31] Q. Hu, K. Richmond, J. Yamagishi, and J. Latorre, "An experimental comparison of multiple vocoder types," in *Proc. 8th ISCA Workshop Speech Synthesis*, 2013, pp. 155–160.

[32] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[33] Z.-H. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted Boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Trans. Audio Speech Language Process.*, vol. 21, no. 10, pp. 2129–2139, Oct. 2013.

[34] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, 2014, Art. no. e006.

[35] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, vol. 27, no. 3, pp. 187–207, 1999.

[36] S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 1983, pp. 93–96.

[37] G. Matthews, I. Deary, and M. Whiteman, *Personality Traits*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[38] D. Funder, "Personality," *Annu. Rev. Psychology*, vol. 52, pp. 197–221, 2001.

[39] P. Corr and G. Matthews, Eds., *The Cambridge Handbook of Personality Psychology*. Cambridge, U.K.: Cambridge Univ. Press, 2009.

[40] D. Ozer and V. Benet-Martinez, "Personality and the prediction of consequential outcomes," *Annu. Rev. Psychology*, vol. 57, pp. 401–421, 2006.

[41] I. Deary, "The trait approach to personality," in *The Cambridge Handbook of Personality Psychology*, P. Corr and G. Matthews, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2009, pp. 89–109.

[42] G. Saucier and L. Goldberg, "The language of personality: Lexical perspectives on the five-factor model," in *The Five-Factor Model of Personality*, J. Wiggins, Ed. New York, NY, USA: Guilford Press, 1996, pp. 21–50.

[43] D. Nettle, *Personality: What Makes You the Way You Are*. London, U.K.: Oxford Univ. Press, 2007.

[44] D. Kenny, L. Albright, T. Malloy, and D. Kashy, "Consensus in interpersonal perception: Acquaintance and the big five," *Psychological Bulletin*, vol. 116, no. 2, pp. 245–258, 1994.

[45] D. Kenny, "PERSON: A general model of interpersonal perception," *Personality Social Psychology Rev.*, vol. 8, no. 3, pp. 265–280, 2004.

[46] D. Howell, *Statistical Methods for Psychology*. Boston, MA, USA: Cengage Learning, 2012.

[47] Y. Tausczik and J. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Language Social Psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[48] T. Raitio, et al., "HMM-based speech synthesis utilizing glottal inverse filtering," *IEEE Trans. Audio Speech Language Process.*, vol. 19, no. 1, pp. 153–165, Jan. 2011.

[49] J. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 2, pp. 195–208, Apr. 2014.

[50] L. Chen, N. Braunschweiler, and M. J. Gales, "Speaker and expression factorization for audiobook data: Expressiveness and transplantation," *IEEE Trans. Audio Speech Language Process.*, vol. 23, no. 4, pp. 605–618, Apr. 2015.

[51] O. Watts, Z. Wu, and S. King, "Sentence-level control vectors for deep neural network speech synthesis," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 2217–2221.

[52] Y. Ohtani, K. Mori, and M. Morita, "Voice quality control using perceptual expressions for statistical parametric speech synthesis based on cluster adaptive training," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2258–2262.

[53] M. Aylett, B. Potard, and C. C. J. Pidcock, "Expressive speech synthesis: Synthesising ambiguity," in *Proc. 8th ISCA Workshop Speech Synthesis*, Aug. 2013, pp. 133–138.

[54] C. Stevens, N. Lees, J. Vonwiller, and D. Burnham, "On-line experimental methods to evaluate text-to-speech (TTS) synthesis: Effects of voice gender and signal quality on intelligibility, naturalness and preference," *Comput. Speech Language*, vol. 19, no. 2, pp. 129–146, 2005.

[55] K. Mustafa and I. C. Bruce, "Robust formant tracking for continuous speech with speaker variability," *IEEE Trans. Audio Speech Language Process.*, vol. 14, no. 2, pp. 435–444, Mar. 2006.

[56] C. L. Bennett and A. W. Black, "The Blizzard challenge 2006," in *Proc. Blizzard Challenge*, 2006, http://www.festvox.org/blizzard/blizzard2006.html

[57] A. J. Hayter, "The maximum familywise error rate of fisher's least significant difference test," *J. Amer. Statistical Assoc.*, vol. 81, no. 396, pp. 1000–1004, 1986.

[58] J. R. Levin, R. C. Serlin, and M. A. Seaman, "A controlled, powerful multiple-comparison strategy for several situations," *Psychological Bulletin*, vol. 115, no. 1, 1994, Art. no. 153.

[59] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social Signal Processing: Survey of an emerging domain," *Image Vis. Comput. J.*, vol. 27, no. 12, pp. 1743–1759, 2009.

[60] D. Klatt and L. Klatt, "Analysis, synthesis, and perception of voice quality variations among female and male talkers," *J. Acoustical Soc. America*, vol. 87, no. 2, pp. 820–857, 1990.

[61] K. R. Scherer, "Personality inference from voice quality: The loud voice of extroversion," *Eur. J. Social Psychology*, vol. 8, pp. 467–487, 1978.

[62] P. Slater, "Role differentiation in small groups," *Amer. Sociol. Rev.*, vol. 20, no. 3, pp. 300–310, 1955.

[63] P. Burke, "The development of task and social-emotional role differentiation," *Sociometry*, vol. 30, no. 4, pp. 379–392, 1967.

[64] S. Partan and P. Marler, "Communication goes multimodal," *Sci.*, vol. 283, no. 5406, pp. 1272–1273, 1999.

[65] S. Partan and P. Marler, "Issues in the classification of multimodal communication signals," *Amer. Naturalist*, vol. 166, no. 2, pp. 231–245, 2005.

[66] Z. Wu, O. Watts, and S. King, "Merlin: An open source neural network speech synthesis system," in *Proc. 9th ISCA Workshop Speech Synthesis*, 2016, pp. 202–207.

[67] B. Potard, M. P. Aylett, D. A. Baude, and P. Motlicek, "Idlak Tangle: An open source Kaldi based parametric speech synthesiser based on DNN," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2016, pp. 2293–2297.

[68] A. V. D. Oord, et al., "WaveNet: A generative model for raw audio," arXiv:1609.03499, 2016, https://arxiv.org/abs/1609.03499

Matthew P. Aylett has been involved in speech technology and HCI as a student and researcher since 1994. He founded Cereproc Ltd in 2006 with the aim of creating commercially available, characterful speech synthesis. In 2007 Cereproc released the first commercial synthesis to allow modification of voice quality for adding underlying emotion to voices. He has remained active both commercially, where he dictates Cereproc's technical strategy, and academically, as a research fellow in The School of Informatics, University of Edinburgh, where he was awarded a Royal Society Fellowship in 2012 looking at speech and personification. He has substantial commercial engineering and product development management experience together with a broad international research background in prosody, dialogue engineering, affective computing, novel interface design, and psycholinguistics.

Alessandro Vinciarelli is full professor with the University of Glasgow, where he is member of the School of Computing Science and affiliate academic of the Institute of Neuroscience and Psychology. His main research interest is social signal processing, the computing domain aimed at modelling, analysis and synthesis of nonverbal behaviour in human-human and human-machine interactions. He has authored more than 130 publications, including one book and 35 journal papers. According to Google Scholar, his works have attracted more than 5,000 citations. Furthermore, he is or has been associated editor of several journals (including the *IEEE Signal Processing Magazine* and *Cognitive Computation*) and he has co-chaired more than 25 workshops and conferences, including the IEEE International Conference on Social Computing and the ACM International Conference on Multimodal Interaction. Last, but not least, he is the co-founder of Klewel, a knowledge management company recognised with national and international awards. He is a member of the IEEE. (http://www.dcs.gla.ac.uk/vincia).

Mirjam Wester received the PhD degree from the Radboud University, in Nijmegen, The Netherlands, in 2002. From 2003-2016, she was a research fellow in the Centre for Speech Technology Research, University of Edinburgh, United Kingdom. Currently, she is senior research linguist with CereProc, Edinburgh. Her research interests focus on taking knowledge of human speech production and perception and applying it to speech technology.