Artificial Personality and Disfluency

Artificial Personality and Disfluency

Mirjam Wester[1], *Matthew Aylett*[1,2], *Marcus Tomalin*[3], *Rasmus Dall*[1]

[1]The Centre for Speech Technology Research, The University of Edinburgh, UK
[2] Cereproc Ltd., UK
[3] Cambridge University Engineering Department, University of Cambridge, UK
m.wester@inf.ed.ac.uk, matthewa@inf.ed.ac.uk, mt126@cam.ac.uk, r.dall@sms.ed.ac.uk

Abstract

The focus of this paper is artificial voices with different personalities. Previous studies have shown links between an individual's use of disfluencies in their speech and their perceived personality. Here, filled pauses (*uh* and *um*) and discourse markers (*like*, *you know*, *I mean*) have been included in synthetic speech as a way of creating an artificial voice with different personalities. We discuss the automatic insertion of filled pauses and discourse markers (i.e., fillers) into otherwise fluent texts. The automatic system is compared to a ground truth of human "acted" filler insertion. Perceived personality (as defined by the big five personality dimensions) of the synthetic speech is assessed by means of a standardised questionnaire. Synthesis without fillers is compared to synthesis with either spontaneous or synthetic fillers. Our findings explore how the inclusion of disfluencies influences the way in which subjects rate the perceived personality of an artificial voice.

Index Terms: artificial personality, TTS, disfluency

1. Introduction

Speech influences the personality impressions that listeners have about a speaker, especially in a zero acquaintance scenario [1]. These impressions, while not necessarily accurate still drive peoples' behaviour and attitude towards others [2]. Nass and Brave [3] describe how people respond to voice technologies as if responding to actual people and behave as they would in any social situation. Investigating how perceived personality can be manipulated in artificial speech is a crucial step in creating more satisfactory performing synthetic speech systems. Synthetic voices that convey appropriate personality traits may be more effective in fulfilling their function.

For almost a century now, speech as a personality trait has been investigated [4]. Speaking rate, loudness, voice quality and the effect of pausing are all factors that have been shown to influence perceived personality. A short survey of how nonverbal vocal behaviour influences the perception of personality can be found in [5]. In our work, we used filled pauses to create more conversational style speech synthesis [6, 7]. Investigating the inclusion of filled pauses and discourse markers in speech synthesis and the effect this may have on the perceived personality of the synthetic voice is the main focus of this paper.

A recent study [8], showed links between the use of filled pauses and discourse markers and perceived personality. Laserna and colleagues [8] analysed the use of the fillers *I mean, you know, like, uh,* and *um* and found that discourse markers were found to be used more commonly among women, younger participants and more conscientious people. Although there is no straightforward link between filled pauses and anxiety, with some studies showing a relationship [9] and others not [8, 10], it seems listeners' impressions are shaped by a speaker's use of filled pauses, linking it to anxiety and lack of preparation [11].

This study looks at the effect of the same fillers Laserna et al. investigated, *I mean, you know, like, uh,* and *um*, on the perceived personality of a unit-selection synthesis system. The text material we used was designed to produce a perceived variation in four of the big five traits (extroversion, agreeableness, conscientiousness, neuroticism and openness). In order for a synthetic system to produce disfluencies/fillers[1] they need to be inserted into the text, ideally automatically. In [6] we proposed a system which automatically inserts filled pauses. This system has been extended to also insert discourse markers [12]. To analyse how well our automatic system is producing valid fillers at valid insertion points in a sentence we need ground truth data. The method we have adopted in this paper is to use "acted" fillers. Subjects are given a text and are asked to imagine they are saying the sentence and to then insert fillers where they think they would do this in a real situation. Although this is an unusual task for subjects to carry out, we found –in [6]– that subjects are consistent with actual usage of filled pauses and show a very reasonable level of agreement with each other where to place filled pauses.

Fillers were added to the texts at insertion points most frequently used by subjects. Next the texts were synthesised using a unit-selection system. We compared three conditions: synthesis without fillers, synthesis with synthetic fillers and synthesis with spontaneous fillers. A different set of subjects were asked to judge the naturalness and personality of the speech using a set of questions designed to evaluate the Big-Five personality traits. The question we consider in this paper is: Does including disfluencies in speech synthesis affect the perceived personality of an artificial voice?

The remainder of the paper is organised as follows: first we describe the text materials and how the fillers were inserted (by both subjects and automatically). Next, a description of the unit selection system and the generation of the fillers is given. This is followed by the setup of the perception test. Section 3 gives the results of the filler insertion, followed by naturalness ratings of the synthetic speech and finally the artificial personality trait results. In the final section, we discuss how the results can be used to inform the creation of artificial personalities and we conclude the paper with a few brief remarks concerning future work.

[1]The terms disfluencies and fillers, although strictly speaking not interchangeable, are used interchangeably in this paper to refer to filled pauses and discourse markers together.

2. Method

2.1. Text Materials

The text materials used in this study were crafted to elicit different personality traits. For more detail on how this was validated see [13]. The materials consist of eight paragraphs describing a person's view of their approach to a working environment (*About Myself* Text) and a second set of text materials consisting of negative and positive emotions in a speed dating context together with a neutral baseline in the form of recipe information (*Speed Dating* Text). Table 1 shows examples of the various text types.

About Myself
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.
About Myself
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.
Speed Dating (negative)
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.
Speed Dating (positive)
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.
Recipe Text (neutral)
Stir to combine thoroughly, then pour into the prepared baking pan. Bake for about 20 minutes in the top of the oven. (Alternatively, you can bake these in a muffin tray lined with paper cases.)

Table 1: *About Myself and Speed Dating text examples.*

2.2. Acted Insertion of Fillers

Twenty subjects were presented the 16 paragraphs (8 *About Myself*, 6 *Speed Dating* and 2 Recipe Texts) as separate sentences (42 total) and were asked to imagine they were speaking the sentences aloud and to decide where to insert one filler per sentence. They were given the fillers *I mean, you know, like, uh* and *um* to choose from.

2.3. Automatic Insertion of Fillers

In addition to the acted insertions, we used an automatic N-gram approach to inserting fillers. As in the acted scenario five fillers were considered. The language model was trained on 20 M words (1M sentences) of data from AMI [14], Fisher [15], Switchboard [16] and an unreleased corpus of British conversational telephone speech. In this work, the N-gram approach described in [6, 12] was used for filler insertion. The N-gram was a 6-gram, trained using the SRILM [17] toolkit with Knesser-Ney discounting. The automated system includes a disfluency parameter (DP) [12] which specifies the desired degree of disfluency (0= maximally fluent; 1=maximally disfluent). This parameter was set to 0.3, as it resulted in one disfluency per sentence on average, similar to the task the human subjects were asked to perform. *I mean* and *you know* were treated as single lexical units despite being phrasal structures.

2.4. Synthetic Speech System

When synthesising disfluencies one is faced with the problem that the read speech recorded for producing high quality synthesis does not contain disfluencies. On the other hand, spontaneous speech recordings which do contain myriad disfluencies do not result in high quality synthesis [18]. We opted to go with the middle ground and use CereVoice unit selection synthesis [19] which is based on a corpus of read speech and spliced spontaneous disfluencies from the same female speaker into the speech. The voice we used was a female Scottish voice called "Heather" [20].

When splicing spontaneous speech into unit selection synthetic speech which is based on read speech inevitably there are issues with differing recording conditions, differences in spectral characteristics and different contexts. We manually set amplitude and rate parameters to alleviate the problems arising from splicing as much as possible.

A second set of paragraphs containing disfluencies were generated using only unit selection. Although synthetic disfluencies are less natural sounding than spontaneous disfluencies the overall impression of the synthesis should be improved as both the fillers and text are based on the same recordings and the joins will be better.

We compared three conditions:

1. **FLU-SYN** Fluent synthetic speech.
2. **DIS-SPON** Disfluent spontaneous synthetic speech; disfluencies spliced from natural spontaneous speech.
3. **DIS-SYN** Disfluent synthetic speech; disfluencies produced with the read speech unit selection system.

2.5. Perceptual Experiment Setup

All the materials were synthesised three times and divided over three blocks. Within each block none of the 16 paragraphs occurred more than once and the three types of synthetic speech were balanced within each block (5-5-6). Forty-five subjects (15 per block) were asked to rate the naturalness of the speech they heard on a scale from 1 (Bad) to 5 (Excellent) and to answer a set of questions (Newcastle Personality Assessor) designed to assess individuals on the big five personality dimensions [21]. Table 2 lists the questions. The answers available to the subjects were "very unlikely", "moderately unlikely", "neither likely nor unlikely", "moderately likely" and "very likely".

Experiments were carried out using a web interface. The forty-five listeners were seated in sound isolated booths and listened to all samples using Beyerdynamic DT 770 PRO headphones. Listeners were remunerated for their time and effort.

3. Results

3.1. Acted and Automatic Filler Insertion

Fillers were placed into the 42 sentences by 20 subjects and the automatic system with the disfluency parameter set to 0.3. Table 3 shows results for both the subjects and the automatic system. The subjects use almost half of all the insertion points (IPs) available to them. The beginning and end of the sentence are also considered to be insertion points. As subjects were instructed to insert one filler per sentence the score here is near to 1. Setting the DP to 0.3 leads to one inserted filler per sentence

ID	Question
	If you met this person for the first time, based on this audio, do you think that person would...
1.	Start a conversation with a stranger?
2.	Make sure others are comfortable and happy?
3.	Use difficult words?
4.	Prepare for things in advance?
5.	Feel blue or depressed?
6.	Plan parties or social events?
7.	Insult people?
8.	Think about philosophical questions?
9.	Let things get into a mess?
10.	Feel stressed or worried?

Table 2: *Newcastle Personality Assessor Questions*

for the automatic system. Table 3 further shows that subjects agree on the exact same position 35% of the time. If a more liberal measure is applied, taking the three most used IPs as correct, the agreement raises to 69%. These results are substantially higher than in [6]. The automatic system shows 26% agreement with the subjects' top IP and this raises to 63% when the top three IPs are considered correct.

Fillers for generating the synthetic speech were inserted in each sentence at the insertion point most frequently used by the twenty subjects. The choice of filler was based on two factors: 1) which filler the subjects most frequently chose for that sentence as well as 2) keeping the five fillers as balanced as possible across all of the material.

	Pos	Used	Ins	Top	Top 3
Subjects	15.95	46.41%	0.97	34.76%	68.72%
Automatic	-	-	1.00	25.58%	62.97%

Table 3: *Mean values over all sentences for Possible IPs (Pos), Used IPs (Used), Inserted FPs (Ins), most (Top) and three most (Top 3) used IP agreements.*

3.2. Naturalness ratings

The raw data collected from each subject was a set of non-parametric opinion scores in the form of a naturalness rating from 1–5 and answers from 1–5 to the 10 question from the Newcastle Personality Assessor (NPA) [21]. Each personality trait score is based on two NPA questions, so scores range from 2 to 10.

Figure 1 shows the results of the MOS test for the *About Myself* texts (left) and for *Speed Dating* text (right). Wilcoxon rank tests showed no significant differences in naturalness ratings for the *About Myself* texts. For the *Speed Dating* texts Wilcoxon rank tests showed that there is no difference between the two types of disfluent speech systems but there is a significant difference between fluent and disfluent speech.

3.3. Artificial Personality Results

A by-materials repeated measures MANOVA across all five personality trait scores was carried out with system type (FLU-SYN, DIS-SPON, DIS-SYN) as a within-materials factor and text group (*About Myself* and *Speed Dating*) as between-materials factors. Pillai's Trace was used to determine significance. Both the between-materials factor text group (F(1, 714) = 96.32, $p < 0.001$) and system type were highly significant (F2, 714)= 6.390, $p < 0.001$). The MANOVA did not show a significant interaction between system type and text group.

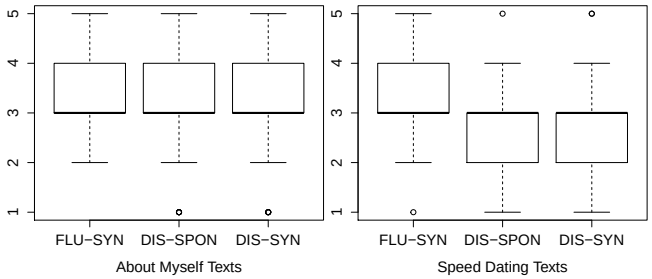

Figure 1: *Naturalness MOS scores for fluent (FLU-SYN), disfluent spontaneous fillers (DIS-SPON) and disfluent synthetic fillers (DIS-SYN) for About Myself and Speed Dating texts.*

3.3.1. Personality traits and system type

As the above MANOVA showed that system type affects personality trait scores, in the following section, we inspect the effect of the systems (FLU-SYN, DIS-SPON, DIS-SYN) on the personality traits for the two texts types separately.

The effect of the three types of synthesis on the ratings of personality traits was measured with five one-way ANOVAs with a x5 Bonferroni correction for the *About Myself* texts and are presented in Table 4. The results show that the systems score significantly different on conscientiousness and openness. In both cases it is the Fluent condition that scores higher. Including disfluencies in speech leads to lower perceived conscientiousness and openness.

Trait	F(2,357)	p-value
Extroversion	1.16	NS
Agreeableness	0.01	NS
Conscientiousness	11.29	< 0.001
Neuroticism	2.07	NS
Openness	6.22	< 0.001

Table 4: *About Myself: effect of systems on personality traits.*

Table 5 shows results from the ANOVAs for the *Speed Dating* texts with system type (FLU-SYN, DIS-SPON, DIS-SYN) as the within system type and text type (positive, negative, neutral) as between-materials factors. For the *Speed Dating* text we find again that the fluent system is rated more conscientious as well as more extrovert whereas the disfluent systems are rated more neurotic.

Trait	F(2,351)	p-value
Extroversion	7.22	< 0.001
Agreeableness	0.256	NS
Conscientiousness	10.37	< 0.001
Neuroticism	3.63	0.027
Openness	2.07	NS

Table 5: *Speed Dating: effect of systems on personality traits.*

Figure 2 presents boxplots for the personality traits that showed significant differences between system types, i.e., extroversion, conscientiousness, openness and neuroticism. The

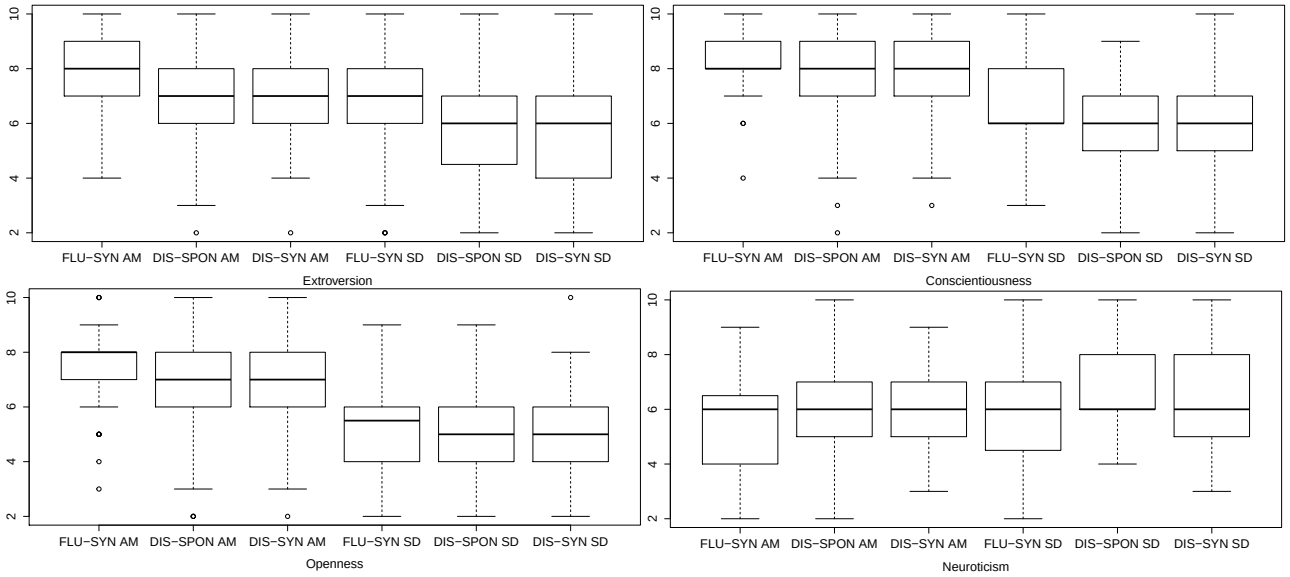

Figure 2: *Boxplots of NPA scores for the four personality traits (Extroversion, Conscientiousness, Openness and Neuroticism) that showed significant differences between fluent and disfluent synthetic speech on About Myself (AM) and Speed dating (SD) texts.*

results are presented separately for the *About Myself* and *Speed Dating* texts.

4. Discussion

The question we set out to answer with this study was: "Does including disfluencies in speech synthesis affect the perceived personality of an artificial voice?" We found that the *About Myself* and *Speed Dating* texts used in this study are a rich set of materials for exploring different personalities, they clearly elicit variation in perceived personality. The fluent system (without fillers) sounded more conscientious and more open on the *About Myself* texts and more extrovert, more conscientious and less neurotic on the *Speed Dating* texts. Or put differently, including disfluencies does indeed affect the perceived personality of an artificial voice, it makes the voice sound more neurotic, less open, less extrovert and less conscientious.

The objective of adding fillers was to make a more conversational and conscientious voice. The discourse markers in [8] were found to be used by more conscientious people, this did not come to the front in our study. Of course, in this study there was not a dialogue between the system and a user which is the scenario in which the discourse markers would potentially be able to have the same kind of effect as in [8]. In [8], the explanation for the association between the use of discourse markers and conscientiousness was that conscientious people are generally more thoughtful and aware of themselves and their surroundings. Thus, when having a conversation conscientious people use discourse makers to imply their desire to share or rephrase opinions to recipients. The finding that the voices with disfluencies were rated as more neurotic falls into line with the studies by Christenfeld & Creager [9] and Sherer & Sherer [22] where filled pauses have been considered as a reflection of anxiety.

Acted filler insertion, although a somewhat artificial task, leads to very reasonable agreement values between listeners. The automatic filler insertion system with a DP of 0.3 also results in very acceptable placement of fillers compared to human subjects. Although we used the human inserted fillers in the experimental material in the long term we will want to use an automatic system to predict this. One of the things that will require further investigation is personalisation of filler insertion. At present we took fillers that were based on an average of 20 subjects, whereas fillers are extremely personal, e.g., somebody who uses um, does not tend to use uh and gender and age also influence the use of filled pauses and discourse markers [23, 8]. Ideally we will want to be able to predict different profiles of filler usage and the personality type they are associated with.

The main focus here was the use of disfluencies to alter the perceived personality of a unit selection voice, we did not touch on the importance of the quality of the synthetic speech. However, in order to be able to create a convincing artificial personality, a synthetic voice that can produce naturally sounding spontaneous speech is paramount. Work is ongoing to achieve this [18, 6].

5. Conclusions

Speech influences the personality impressions that listeners develop about a speaker. This is not limited to people but extends to machines that display human-like features. This work has shown that including disfluencies in synthetic speech influences how subjects rate the perceived personality of an artificial voice. Future work will explore the manipulation of prosodic phenomena (pitch, speech rate, voice quality, inter-lexical pause duration) in conjunction with personalised filler profiles with the goal to further develop synthetic voices that convey appropriate personality traits.

All research data used in this paper is available to download from Edinburgh DataShare `http://hdl.handle.net/10283/787` [24].

Acknowledgements This work was partially supported by EPSRC under Programme Grant EP/I031022/1 (Natural Speech Technology).

6. References

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

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

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

[4] E. Sapir, "Speech as a personality trait," *American Journal of Sociology*, pp. 892–905, 1927.

[5] G. Mohammadi and A. Vinciarelli, "Automatic personality perception: Prediction of trait attribution based on prosodic features," *IEEE Transactions on Affective Computing*, vol. 3, no. 3, pp. 273–278, 2012.

[6] R. Dall, M. Tomalin, M. Wester, W. Byrne, and S. King, "Investigating automatic & human filled pause insertion for speech synthesis," in *Proc. Interspeech*, 2014.

[7] R. Dall, M. Wester, and M. Corley, "The effect of filled pauses and speaking rate on speech comprehension in natural, vocoded and synthetic speech," in *Proc Interspeech*, 2014.

[8] C. M. Laserna, Y.-T. Seih, and J. W. Pennebaker, "Um... who like says you know: Filler word use as a function of age, gender, and personality," *Journal of Language and Social Psychology*, vol. 33, no. 3, pp. 328–338, 2014.

[9] N. Christenfeld and B. Creager, "Anxiety, alcohol, aphasia, and ums." *Journal of Personality and Social Psychology*, vol. 70, no. 3, p. 451, 1996.

[10] G. Mahl, *Explorations in nonverbal and vocal behavior*. Routledge, 2014.

[11] N. Christenfeld, "Does it hurt to say um?" *Journal of Nonverbal Behavior*, vol. 19, no. 3, pp. 171–186, 1995.

[12] M. Tomalin, M. Wester, R. Dall, W. Byrne, and S. King, "A lattice-based approach to automatic filled pause insertion," *DiSS The 7th Workshop on Disfluency in Spontaneous Speech*, 2015.

[13] M. Aylett, M. Wester, and A. Vinciarelli, "Speech synthesis for the generation of artificial personality impressions," *Under preparation*, 2015.

[14] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI Meeting Corpus*," *Language Resources and Evaluation Journal*, vol. 41, no. 2, pp. 181–190, 2007.

[15] C. Cieri, D. Miller, and K. Walker, "The Fisher Corpus: A Resource for the Next Generations of Speech-to-Text Fisher," in *LREC*, Lisbon, Portugal, 2004.

[16] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *ICASSP*, San Francisco, CA, USA, 1992, pp. 517–520.

[17] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "SRILM at sixteen: Update and outlook," in *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*, 2011.

[18] S. Andersson, J. Yamagishi, and R. A. Clark, "Synthesis and evaluation of conversational characteristics in HMM-based speech synthesis," *Speech Communication*, vol. 54, no. 2, pp. 175–188, 2012.

[19] M. P. Aylett and C. J. Pidcock, "The CereVoice characterful speech synthesiser SDK," in *AISB*, 2007, pp. 174–178.

[20] S. Andersson, "Synthesis and Evaluation of Conversational Characteristics in Speech Synthesis," Ph.D. dissertation, University of Edinburgh, 2013.

[21] D. Nettle, *Personality: What makes you the way you are*. Oxford University Press, 2007.

[22] K. R. Scherer and U. Scherer, "Speech behavior and personality," *Speech evaluation in psychiatry*, pp. 115–135, 1981.

[23] E. K. Acton, "On gender differences in the distribution of um and uh," *University of Pennsylvania Working Papers in Linguistics*, vol. 17, no. 2, p. 2, 2011.

[24] M. Wester, M. Aylett, M. Tomalin, and R. Dall, "Artificial personality [dataset]," University of Edinburgh, School of Informatics, Centre for Speech Technology Research, 2015, http://dx.doi.org/10.7488/ds/254.