Bot or Not: Exploring the Fine Line between Cyber and Human Identity

Mirjam Wester
CereProc Ltd.
Edinburgh, UK
mirjam@cereproc.com

Matthew P. Aylett
CereProc Ltd.
Edinburgh, UK
matthewa@cereproc.com

David A. Braude
CereProc Ltd.
Edinburgh, UK
dave@cereproc.com

ABSTRACT

Speech technology is rapidly entering the everyday through the large scale commercial impact of systems such as Apple Siri and Amazon Echo. Meanwhile technology that allows voice cloning, voice modification, speech recognition, speech analytics and expressive speech synthesis has changed dramatically over recent years. The demonstration, described in this paper, is an educational tool in the form of an online quiz called 'Bot or Not'. Using the quiz we have gathered impressions of what people realise is possible with current speech synthesis technology. The opinions of various groups regarding the synthesis of famous voices, sounding like a robot, and the difference between synthesis and voice modification were collected.

CCS CONCEPTS

• **Human-centered computing** → *HCI design and evaluation methods*;

KEYWORDS

speech synthesis, voice modification, identity

ACM Reference Format:
Mirjam Wester, Matthew P. Aylett, and David A. Braude. 2017. Bot or Not: Exploring the Fine Line between Cyber and Human Identity. In *Proceedings of 19th ACM International Conference on Multimodal Interaction (ICMI'17).* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3136755.3143027

1 INTRODUCTION

'Bot or Not' is an educational tool, in the form of a quiz, that has been designed to show-case and explain how technology of speech synthesis and voice modification has moved forward in recent years and to illustrate what is currently possible. It is based around the question: is this recording a bot or not?

There is a whole field of research which deals with detecting whether a speech sample is from a human or a machine, referred to as spoofing and anti-spoofing [3, 4] in the area of automatic speaker verification (ASV). Spoofing is when an attacker tries to manipulate a verification result by by mimicking a client speaker in person by means of for example, impersonation, replay, voice conversion or speech synthesis. Voice conversion and speech synthesis have, in recent years, become especially effective and efficient methods to use for spoofing attacks. In tandem with the advances in spoofing attacks, automatic spoofing detection algorithms are continually being updated and improved [1]. A comparison between automatic spoofing detection algorithms and humans in judging whether a speech sample was produced by a machine or a human (bot or not) showed that humans are not particularly good at this and are outperformed by automatic spoofing detection algorithms [2].

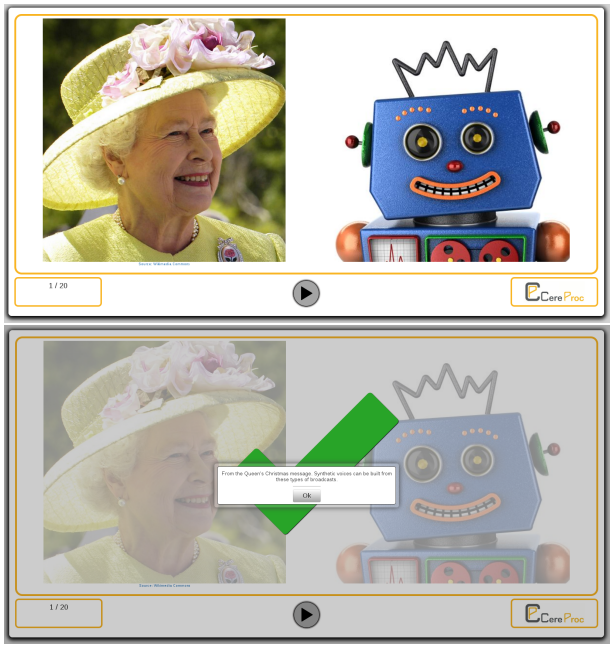

Figure 1: Screen shots of the first item in the quiz. Top: an audio file is played, the user decides whether it is bot or not and selects the corresponding photo. Bottom: After clicking the photo, feedback is presented in a text box.

Being able to tell whether we are dealing with a recording of speech or synthesised speech is relevant in domains beyond speaker verification. Synthesis of famous voices opens avenues for seemingly getting those famous voices to say things that they (the voices' owners) haven't actually said. Or alternatively, when designing an intelligent agent, is it the goal for the agent to sound as human-like as possible thereby running the risk of ending up in the 'uncanny valley' or inadvertently tricking human interlocutors to believe the speech is real, rather than synthetic?

ICMI'17, November 13–17, 2017, Glasgow, UK

ACM ISBN 978-1-4503-5543-8/17/11…$15.00
https://doi.org/10.1145/3136755.3143027

The fine line between human and cyber identity is one that is explored in 'Bot or Not'. Speech synthesis is a crucial component in achieving an intelligent agent and careful consideration of what the voice should sound like and its quality is paramount.

2 'BOT OR NOT' QUIZ & FEEDBACK

'Bot or Not' was designed to provide information on speech synthesis in a fun and interactive manner. (https://cerevoice.com/apps/botornot) The quiz consists of 20 items, which all illustrate something relevant about speech synthesis. Information covered includes: using found data to create voices, unit selection, speech styles, identity, emotion/prosody, phonetic coverage, voice modification, vocoding, statistical speech synthesis, cross-lingual speech synthesis. The quiz starts with the following instructions: "Welcome to CereProc's Bot or Not. Try and guess if the audio is Synthesised or Recorded by clicking on the pictures." Fig. 1 (top) shows a screen shot of the pictures the user can choose between for the first item in the quiz. After selecting one of the pictures, feedback appears on screen explaining something of relevance regarding the speech synthesis used in the item (see the bottom of Fig. 1).

Feedback on 'Bot or Not' was collected by people doing the 'Bot or Not' quiz and giving their opinions on it through questionnaires in three different settings: a focus group meeting, at the British Science Museum Lates, and via CrowdFlower, a crowd-sourcing platform. In all three settings for collecting feedback, the reaction to the quiz was overwhelmingly positive. Here we discuss the scores from the Museum Lates Event and CrowdFlower.

119 participants took part in the quiz at the Museum Lates event and 25 people took part on CrowdFlower. The average score at the museum was 63.4% correct, with the lowest score 35% and the highest score 90% correct. The results from CrowdFlower were comparable with an average score of 57.6 %, low score 40% and high score 80%. Combined results per sentence are shown in Table 1. In this table, bold indicates significantly **incorrect** and italics indicate uncertainty, i.e., neither convincingly bot nor not. These results were in line with our intentions as trick questions were included to make the quiz more challenging and fun.

The lowest scoring item was the Stephen Hawking audio (#9). Almost 3/4 of participants scored Stephen Hawking's speech as his, rather than synthetic. It is very clear that the voice is synthetic but Stephen Hawking chose the words and played them. *"Can God make a stone that is so heavy that he can not lift it? I don't think it is very useful to speculate on what God might, or might not be able to do."* It seems the synthetic voice (despite having an American accent rather than English) is so much a part of Hawking's identity that it is not considered a bot. Another example is one of the items that was overwhelmingly classified as synthetic while it was recorded speech (#14). It concerned the sentence: *"Asquith oozes poise, even though the corsage chafed."* Admittedly a strange sentence, but it was read by a voice talent as one of the sentences which is recorded to increase the phonetic coverage needed to create a good quality synthetic voice.

3 DISCUSSION

'Bot or Not' as an educational tool to explore the fine line between human and cyber identity has been presented. It was an effective approach to gathering opinions on speech synthesis from the general public. By first running a focus group meeting, the tool was properly test-run prior to presentation at the Museum Lates and a short set of questions could be formulated to elicit more specific feedback. Although no explicit conclusions can be drawn from this experiment, it does highlight some of the ethical challenges that one must consider when including voice in an intelligent agent. Although human-like sounding synthesis was preferred over a robotic sounding voice, participants across the board want to know when they are hearing a synthetic voice, i.e., is it a bot or not?

Table 1: Results for synthetic (top) and recorded/human examples in the quiz (bottom). Type: description of the intended educational message of item, Number: position in quiz, % Correct: average score and p-value: significance.

Type	Number	% Correct	p-val.
Synthetic examples			
found data/famous voices	(2) (4) (7)	92, 85, 87	0.000
famous voice & content change	(8)	61	0.010
long-term use of synthesis	(9)	**28**	0.000
unit selection	(11)	81	0.000
unit selection with emotion	(13)	*45*	0.279
synthesis & voice modification	(17)	65	0.000
parametric synthesis	(19)	76	0.000
cross-lingual synthesis	(20)	63	0.003
Human/recorded examples			
found data/famous voices	(1) (5) (6)	82, 79, 76	0.000
famous voice, neutral prosody	(3)	62	0.006
human putting on funny voice	(10)	*47*	0.453
human pretending to be bot	(12)	61	0.010
human pretending to be child	(16)	**34**	0.000
semantically odd sentence	(14)	**30**	0.000
voice modification	(15)	40	0.024
vocoding	(18)	*56*	0.211

ACKNOWLEDGMENTS

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 645378 (ARIA-VALUSPA).

REFERENCES

[1] Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Zhizheng Wu, Federico Alegre, and Phillip De Leon. 2014. Speaker recognition anti-spoofing. In *Handbook of Biometric Anti-Spoofing*. Springer, 125–146.

[2] Mirjam Wester, Zhizheng Wu, and Junichi Yamagishi. 2015. Human vs machine spoofing detection on wideband and narrowband data.. In *Interspeech*. 2047–2051.

[3] Zhizheng Wu, Phillip L De Leon, Cenk Demiroglu, Ali Khodabakhsh, Simon King, Zhen-Hua Ling, Daisuke Saito, Bryan Stewart, Tomoki Toda, Mirjam Wester, and Junichi Yamagishi. 2016. Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 24, 4 (2016), 768–783.

[4] Zhizheng Wu, Nicholas Evans, Tomi Kinnunen, Junichi Yamagishi, Federico Alegre, and Haizhou Li. 2015. Spoofing and countermeasures for speaker verification: a survey. *Speech Communication* 66 (2015), 130–153.