Voice Puppetry: Speech Synthesis Adventures in Human Centred AI

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

Yolanda Vazquez-Alvarez
CereProc Ltd.
Edinburgh, UK
yv.alvarez@gmail.com

ABSTRACT

State-of-the-art speech synthesis owes much to modern AI machine learning, with recurrent neural networks becoming the new standard. However, how you say something is just as important as what you say. If we draw inspiration from human dramatic performance, ideas such as artistic direction can help us design interactive speech synthesis systems which can be finely controlled by a human voice. This "voice puppetry" has many possible applications from film dubbing to the pre-creation of prompts for a conversational agent. Previous work in voice puppetry has raised the question of how such a system should work and how we might interact with it. Here, we share the results of a focus group discussing voice puppetry and responding to a voice puppetry demo. Results highlight a main challenge in user-centred AI: where is the trade-off between control and automation? and how may users control this trade-off?

CCS CONCEPTS

• **Human-centered computing → Natural language interfaces**;

KEYWORDS

speech synthesis, social robots, personification

ACM Reference Format:
Matthew P. Aylett and Yolanda Vazquez-Alvarez. 2020. Voice Puppetry: Speech Synthesis Adventures in Human Centred AI. In *25th International Conference on Intelligent User Interfaces Companion (IUI '20 Companion), March 17–20, 2020, Cagliari, Italy*. ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3379336.3381478

1 INTRODUCTION

Copy resynthesis is a technique where the parameterisation of human voice audio is used to directly control a speech synthesis rendition. It is a well established technique used to develop and evaluate speech synthesis systems (e.g. [3, 5]).

In addition, copy resynthesis can be used as a basis for a *voice puppetry* system. Such a system allows natural speech input to control the output speech for a target voice [1, 2]. This contrasts with, but is related to, *voice morphing*, where a source speaker's voice is converted directly into a target speaker's voice without the requirement of a speech synthesis system. So called *Puppetry* systems are commonly used to control the rendering of graphics and lip-syncing, where human movements are mapped onto a potentially very different body form. The ability to extend this control to the vocal performance of an artificial character's voice, for example to tell a joke, despite ethical dangers, is an important area of research. For an example of puppetry see *https://tinyurl.com/wpzvcww* which demonstrates a female synthetic voice following a controlling male voice and *https://tinyurl.com/qps326s* where the same joke is retold by two celebrity voices, Donald Trump and Queen Elizabeth, with added effects.

To explore the use of voice puppetry a command line demo was created. To:

(1) Allow the user to record a phrase and re-record until they are happy with their performance.
(2) Use speech recognition to transcribe the phrase.
(3) Use a speech aligner developed by CereProc to analyse the duration and pitch of each sound in the user's input speech.
(4) Use the user's durations to control the rendition of a target voice, in this case the Queen Elizabeth II of England.
(5) Allow some control of the speech rate in the target rendition.
(6) Allow the user to record and puppeteer more phrases.

In order to inform a design process, this demo was used as a technical probe [4] with a focus group.

2 FOCUS GROUP

The focus group was recruited from staff with 2 members from a commercial background and 2 members from an engineering background. Although all members were familiar with speech synthesis technology they were not experts in this field. An experienced speech synthesis engineer and author of the tech probe moderated the group and an HCI (human-computer interaction) specialist took notes and advised on format (see Figure 1.1).

The discussion was split into three phases: Pre-demo discussion, interacting with voice puppetry demo (see Figure 1.2), and post-demo discussion. The goal of the focus group was to help understand how a voice puppetry service might be incorporated in order to provide a better service for customers.

2.1 Pre-Demo Discussion

The questions used to drive the discussions were: What might you understand by the term *voice puppetry*? What advantages can you imagine you might get from voice puppetry compared to speech synthesis? What disadvantages can you imagine in the use of voice puppetry?

Voice puppetry was understood to be *"Someone having their voice finely controlled"* P01, *"The direct manipulation of the output through the input"* P02, *"Infusing the synthesis with some characteristics from*

IUI '20 Companion, March 17–20, 2020, Cagliari, Italy

ACM ISBN 978-1-4503-7513-9/20/03.
https://doi.org/10.1145/3379336.3381478

Figure 1: 1. Focus group environment with 4 members, a moderator and a note taker. 2. Focus group member running the simple command line driven voice puppetry demo.

another speaker" P04. Advantages included being able to make synthetic voices sound more human by the fine control pitch and speed: *"That has been quite lacking before"* P01, and applications in media: *"great potential in dubbing for games"* P04, *"Dubbing Movies"* P03. The main disadvantage was seen as the ability to fake a person's identity: *"Deep Fake"* P02, *"It could be used for very sinister means in the wrong applications."* P02.

In discussing control, the term *energy* was used and related to emotion, that it *"can covey passion about a topic"* P01. The discussion often reiterated the unique identity of a person's voice *"Someone voice is just as unique to them as their appearance"* P02.

2.2 Voice Puppetry Demo

At this early stage, to avoid pitch normalisation issues, the demo only modelled the speed of delivery and not pitch. Users found this a little frustrating. In addition, the use of ASR caused some problems for some speakers who found recognition rates affected performance. Issues with alignment could also cause phrase final words to be cut or shortened inappropriately when re-spoken by the Queen's voice.

2.3 Post-demo Discussion

The questions used to drive the discussion were: How useful would the demo be to help generate synthetic prompts? What were your feeling in the way it reproduced your voice as the Queen? Did using the demo change your views on voice puppetry?

The demo helped give insight into the process: *"It gives you a good idea of how it works"* and *"How difficult it is to get it right"* P01, and it was functional, a participant noted *"It works"* P03. Their were some issues with speech recognition and a more interactive design where participants could correct recognition results would have been preferred but overall the use of the recogniser was perceived as *"Useful"* P03.

There was no suggestion that it felt odd or weird to recreate a phrase in another voice. It was noted that the output *"was a similar speech rate... The timing was very good"* P01. The puppetry of pitch and melody was not implemented in this demo and this lack of functionality was noted by one of the participants: *"I expected it to pick up more"* P03.

The participants had little to add on their overall views of voice puppetry except that it gave more insight into concrete applications such as making *"virtual GPs... more convincing"* P01, but with significant ethical issues *"Whether it's ethical or not is a different question"* P01.

3 DISCUSSION

Voice puppetry has clear applications in media and gaming. However, it comes with significant ethical considerations. Building a user-centred puppetry system is a significant challenge. It requires both an automatic means of capturing a vocal performance from the user and speech technology to render this performance in a different voice. The design issues raised are as follows:

- The ability to review, modify and improve the result is a complex interaction and key to user satisfaction.
- The underlying puppetry technology is imperfect and the interaction design needs to reduce the impact of errors.
- The issue of individual vocal identity is a sensitive one and requires careful handling.

The fact that the puppetry was a semi-manual process and thus prevented a fully automatic solution was not an issue raised by the focus group. Overall the idea of users curating voice output was readily accepted and understood. An example of a joke created using this puppetry system for a more convenctional speech synthesis voice can be found here *https://tinyurl.com/uamps4n*. The commercial synthesis voice to build this can be accessed here *https://www.cereproc.com/en/support/live_demo* and choose Heather-CereWave.

ACKNOWLEDGMENTS

This work was supported by the European Union's Horizon 2020 program under Grant Agreement No 780890.

REFERENCES

[1] Matthew P Aylett, David A Braude, Christopher J Pidcock, and Blaise Potard. 2019. Voice Puppetry: Exploring Dramatic Performance to Develop Speech Synthesis. In *Proc. 10th ISCA Speech Synthesis Workshop*. 117–120.

[2] Yuan-Yi Fan, Soyoung Shin, and Vids Samanta. 2019. Evaluating expressiveness of a voice-guided speech re-synthesis system using vocal prosodic parameters. In *Proceedings of the 24th International Conference on Intelligent User Interfaces: Companion*. ACM, 67–68.

[3] Wendy J Holmes. 1989. Copy synthesis of female speech using the JSRU parallel formant synthesiser. In *First European Conference on Speech Communication and Technology*. 2513–2516.

[4] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 17–24.

[5] Oytun Turk and Marc Schroder. 2010. Evaluation of expressive speech synthesis with voice conversion and copy resynthesis techniques. *IEEE Transactions on Audio, Speech, and Language Processing* 18, 5 (2010), 965–973.