Who owns your voice? Ethically sourced voices for non-commercial TTS applications

Kristen M. Scott
kristen.scott@m-iti.org
Madeira Interactive Technology Institute (M-ITI)
Funchal, Portugal

Simone Ashby
simone.ashby@m-iti.org
Interactive Technology Institute (ITI)
Funchal, Portugal

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

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

ABSTRACT

We examine the ethical questions surrounding voice donation for speech synthesis technology, including questions of voice ownership, identity and unintended consequences. This is examined specifically in the context of non-professional volunteer voice donors in small communities. We propose a multi-step informed consent process that more fully engages with TTS voice donors.

CCS CONCEPTS

• **Security and privacy → Privacy protections**; • **Human-centered computing**;

KEYWORDS

speech synthesis, text-to-speech, voice, privacy

ACM Reference Format:
Kristen M. Scott, Simone Ashby, David A. Braude, and Matthew P. Aylett. 2019. Who owns your voice? Ethically sourced voices for non-commercial TTS applications. In *1st International Conference on Conversational User Interfaces (CUI 2019), August 22–23, 2019, Dublin, Ireland.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3342775.3342793

1 INTRODUCTION

As part of an EU funded community radio project working with innovative low-barrier technologies, we are integrating locally accented text-to-speech (TTS) voices into a free/open technology stack for low-power FM radio stations. Our goal is to develop freely accessible developer tools for generating local, varied and representative TTS voices for use by other communities. In practice the development of hyperlocal TTS voices has raised some ethical questions that warrant significant attention. An early voice recording effort in one of the small, remote communities where we are launching stations and for which we are developing local DNN voices with the Idlak Tangle toolkit [14], went awry after the community member / amateur voice talent subsequently requested that their voice recordings from the session be deleted[1]. Upon further inquiry, this person admitted having entered the recording session with misgivings about the process, stating that they had heard 'something bad' about having their voice recorded and that the recording session, and overall objective of using these data to generate a TTS voice, gave them a sense of 'losing control' of their voice.

Rather than focus solely on how to make community members more comfortable or 'willing' to be part of the voice recording process, we acknowledged that there are significant and valid concerns to be explored regarding our collection effort. Discussions of ethics and privacy in speech synthesis tend to focus on issues specific to particular uses of the technology, e.g. personal assistants and privacy issues [8] or voice cloning and 'deep fake' concerns [5], along with extolling the benefits of TTS for enhanced accessibility and increased representation [13]. There are, however, some less recognized ethical considerations surrounding even purportedly ethical and neutral speech synthesis applications. In this paper we examine the question of what it truly means to 'sign away' control of one's voice to a TTS voice. The nature of TTS applications is that the voice donor cannot control how the synthetic voice, a near facsimile of their own, will ultimately be used. Given the intrinsic link between one's voice and identity, even with synthetic voices [19], is it ethical to ask people to contribute their voices to such a technology? Can the free use of another person's voice be ethically or legally solicited? Given the lack of clarity on the legal question, and the real possibility of unintended negative consequences, we propose a multi-step informed consent process that more fully engages with volunteer, non-professional TTS voice donors before, during and after the recording session. The larger ethical considerations discussed here may be relevant to the commercial voice synthesis process, however specific concerns and solutions for commercial voice building are outside the scope of this paper.

2 BACKGROUND

2.1 Legal ownership considerations

Kimppaa and Saarni [9] examine the implications of voice synthesis from a legal point of view, asking if a person has a natural right to

[1] no TTS voice was ever developed from the recordings

CUI 2019, August 22–23, 2019, Dublin, Ireland

ACM ISBN 978-1-4503-7187-2/19/08…$15.00
https://doi.org/10.1145/3342775.3342793

their voice, and whether TTS voice donors have the right to control use of their voice in any resulting applications. They examine possible ownership rights, including the possibility of applying the Intellectual Property Rights (IPR) protections of trademark, copyright or patent to one's voice. However, the authors acknowledge that rather than enforce an inherent right to one's own voice, these options require a cumbersome opt-in legal process.

If we do assume a person has ownership of their voice, even a synthesized version of it, it follows that some sort of permission for its use would have to be given by the owner, whatever the specific mechanism for ensuring that protection [9]. Current practice for commercial engagements with TTS voice donors involve release forms, or contracts, that detail allowed uses of the raw voice recordings and synthesized voices. In many cases voice donors essentially sign away their rights to the raw voice recordings collected and anything that may be generated from them. More recently, some commercial entities are drafting contracts that restrict use of the resulting TTS voice along a predefined set of parameters, e.g. the size of the marketing campaign or length of time the TTS voice can be used [17]. However, providing potential voice donors with a script or comprehensive outline of the content that will be read by the resulting TTS voice is not feasible or even achievable for more robust uses of TTS. Thus, a considerable loss of control over the donor's voice likeness will always feature in the consent request.

2.2 Balancing open data and technology with privacy and ethical concerns

By developing localised TTS resources and technologies that are free and open source, we aim to bridge the technology and information gap for speakers of non-dominant language varieties while helping to raise the prestige of a wider range of voices and improve user experience [3]. This requires an increased number of voice donors and the need for contributions from non-professional voice talent. We cannot expect such individuals to be well versed in their legal rights, or in the potential applications and uses of TTS technology.

Even voice talent professionals do not necessarily have a grasp of the scope and nature of participating in a speech synthesis project. A very public example of this comes from the original voice talent behind the automated personal assistant, Siri. In 2011, Susan Bennett was hired as a voice talent not by Apple, but by the company ScanSoft; she did not know that the recordings would be used for a personal assistant. She only found out with the release of the iPhone 4s, whose automated personal assistant bore a strong resemblance to her voice [2, 6]. According to Bennett she had no idea what was possible using her voice recordings, revealing "I just never imagined that technology could take my voice and make it say anything". Bennett added "Your voice is such a personal thing. It's like a fingerprint. It's your intellectual property. And suddenly, it's no longer yours...we signed a contract to do the work - but I don't think we had any clue of the ramifications" [2].

Concerns such as lack of informed consent, unforeseen and unforeseeable consequences, and a sense of loss of control of one's voice or even identity must be adequately addressed. Given the inherent link between voice and identity [1, 7, 13, 16] it is not surprising that people have strong reactions to the concept of a synthesised version of their voice being under the control, and potentially subject to the whims, of others.

These concerns are not unfounded. People often have non-neutral reactions to synthetic voices, including strong opinions about the quality, tone, or accents [3, 20], or frustrations with the functioning of applications using synthetic voices [4, 10, 15]. We are prone to treating synthesised voices, and computers in general, as if they are human [11, 12]. It is conceivable that in reacting to a syntheised voice, listeners might transfer this association onto the TTS voice donor. Such concerns seem especially relevant in the types of small communities we are focused on for this project.

3 INNOVATING INFORMED CONSENT

Can it be considered informed consent if the person is not given a vivid picture of the technologies they are consenting to be a part of? Can one be expected to 'imagine' the possible negative consequences of unfamiliar technologies and somehow consent to them. At a time when technological advancements (and their various socio-cultural consequences) are being unleashed at an accelerated pace, these are difficult questions even for the long practicing expert to address[18]. The solution we have settled on for our project is a multi-step information and consent process that seeks to convey both the benefits and possible risks of donating one's voice, prioritizes voice anonymisation and gives the donor final approval.

(1) A detailed information sheet and consent form is provided to the voice donor at least one week prior to the recording session.
(2) The consent form signed prior to recording grants consent only for voice recording for TTS voice development, not for the TTS voice to be used for any public purpose.
(3) Later, the donor is given an opportunity to hear examples of the resulting TTS voice, during which time they can request post-processing manipulations (e.g. affecting the overall pitch) to anonymise the voice.
(4) When the donor is comfortable with the TTS voice, they are asked to sign a release form permitting the use of the voice for achieving project objectives (i.e. for use in radio broadcasts).

4 CONCLUSION

While we have not determined the exact legal status of the right to one's voice, we acknowledge the personal and sensitive nature of the human voice. When recruiting TTS voice donors, we recommend a more comprehensive informed consent process and a shift in development in the field towards practices and features that could mitigate risks to voice donors, such as prioritising investigation into high quality anonymised voices.

ACKNOWLEDGMENTS

We gratefully acknowledge Grassroots Wavelength and LARSYS. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No H2020-ICT-2016-2017-780890

REFERENCES

[1] Marie-Cécile Bertau. 2008. Voice: A Pathway to Consciousness as "Social Contact to Oneself". *Integrative Psychological and Behavioral Science* 42, 1 (March 2008), 92–113. https://doi.org/10.1007/s12124-007-9041-8

[2] Zachary Crockett. 2019. A Q&A with the original Siri voice actor. Retrieved March 29, 2019 from https://thehustle.co/siri-voice-actor

[3] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is More Important Than Expertise: Accent Effects in Speech Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 1553–1556. https://doi.org/10.1145/1240624.1240859 event-place: San Jose, California, USA.

[4] Kerstin Fischer. 1999. Repeats, Reformulations, and Emotional Speech: Evidence for the Design of Human-Computer Speech Interfaces. In *HCI*.

[5] Nicholas Gardiner. 2019. Facial re-enactment, speech synthesis and the rise of the Deepfake. (2019), 79.

[6] Mark Hill and Susan Bennett. 2016. I'm The Voice Of Siri: And No, Apple Didn't Pay (Or Warn) Me. Retrieved April 29, 2019 from https://www.cracked.com/personal-experiences-2108-i-am-siris-voice-4-bizarre-realities.html

[7] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. 'Putting the Face to the Voice': Matching Identity across Modality. *Current Biology* 13, 19 (Sept. 2003), 1709–1714. https://doi.org/10.1016/j.cub.2003.09.005

[8] Joseph 'Jofish' Kaye, Joel Fischer, Jason Hong, Frank R. Bentley, Cosmin Munteanu, Alexis Hiniker, Janice Y. Tsai, and Tawfiq Ammari. 2018. Panel: Voice Assistants, UX Design and Research. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems (CHI EA '18)*. ACM, New York, NY, USA, panel01:1–panel01:5. https://doi.org/10.1145/3170427.3186323 event-place: Montreal QC, Canada.

[9] Kai K Kimppa and Tuomo I Saarni. 2008. RIGHT TO ONE'S VOICE? *Living, Working and Learning Beyond* (2008), 480.

[10] Jennifer Lai and John Vergo. 1997. MedSpeak: Report Creation with Continuous Speech Recognition. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems (CHI '97)*. ACM, New York, NY, USA, 431–438. https://doi.org/10.1145/258549.258829 event-place: Atlanta, Georgia, USA.

[11] Clifford Nass and Kwan Min Lee. 2000. Does Computer-generated Speech Manifest Personality? An Experimental Test of Similarity-attraction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '00)*. ACM, New York, NY, USA, 329–336. https://doi.org/10.1145/332040.332452 event-place: The Hague, The Netherlands.

[12] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153

[13] Esther Nathanson. 2017. Native voice, self-concept and the moral case for personalized voice technology. *Disability and Rehabilitation* 39, 1 (Jan. 2017), 73–81. https://doi.org/10.3109/09638288.2016.1139193

[14] Blaise Potard, Matthew P Aylett, David A Braude, and Petr Motlicek. 2016. Idlak Tangle: An Open Source Kaldi Based Parametric Speech Synthesiser Based on DNN.. In *INTERSPEECH*. 2293–2297.

[15] Silvia Schiaffino and Analía Amandi. 2004. User — interface agent interaction: personalization issues. *International Journal of Human-Computer Studies* 60, 1 (Jan. 2004), 129–148. https://doi.org/10.1016/j.ijhcs.2003.09.003

[16] Diana Sidtis and Jody Kreiman. 2012. In the Beginning Was the Familiar Voice: Personally Familiar Voices in the Evolutionary and Contemporary Biology of Communication. *Integrative Psychological and Behavioral Science* 46, 2 (June 2012), 146–159. https://doi.org/10.1007/s12124-011-9177-4

[17] Voice Tech Podcast. 2019. Voice Actors & Synthetics - David Ciccarelli. Retrieved April 21, 2019 from https://www.buzzsprout.com/159584/1037914

[18] Vivek Wadhwa. 2014. Laws and Ethics Can't Keep Pace with Technology. Retrieved April 4, 2019 from https://www.technologyreview.com/s/526401/laws-and-ethics-cant-keep-pace-with-technology/

[19] 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 the 19th ACM International Conference on Multimodal Interaction*. ACM, 506–507.

[20] J. Wilkie, M. A. Jack, and P. J. Littlewood. 2005. System-initiated digressive proposals in automated humanâĂŞcomputer telephone dialogues: the use of contrasting politeness strategies. *International Journal of Human-Computer Studies* 62, 1 (Jan. 2005), 41–71. https://doi.org/10.1016/j.ijhcs.2004.08.001