First indications for speaker individuality and speech intelligibility in state-of-the-art artificial voices

Claudia Roswandowitz[1,2,3], Thayabaran Kathiresan[2], Elisa Pellegrino[2], Volker Dellwo[2], Sascha Frühholz[1,3,4]

[1] Department of Psychology, University of Zurich, Zurich, Switzerland
[2] Phonetics and Speech Sciences, Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
[3] Neuroscience Center Zurich, University of Zurich and ETH Zurich, Zurich, Switzerland
[4] Center for Integrative Human Physiology (ZIHP), University of Zurich, Switzerland

Introduction

In modern environment, situations in which humans encounter synthesized voices become more and more frequent. Individualist artificial voices have the potential to make human-computer interactions more natural and personalized. One application, among others, are individualist speech assistants for ALS or throat cancer patients who permanently lost their voices. However, to date there is little evidence on how vocal-identity and semantic speech information, both being fundamental for social interactions, are preserved in synthesized voices. First findings from speaker similarity rating studies suggest that indexical vocal information are poorly preserved in synthesized voices compared to the corresponding natural speaker identity (Lorenzo-Trueba et al., 2018). To address the open question, we used a modern voice-synthesis algorithm and tested speaker and speech recognition of artificial and natural speakers' voices.

Methods

To generate artificial voices, we used *sprocket* - an open-source voice conversion software using a Gaussian mixture model framework and a vocoder-free speech wave synthesis technique (Kobayashi & Toda, 2018). For sprocket, the voice conversion challenge in 2018 revealed second-best sound quality scores for same-speaker pairs and the sixth place for speaker similarity rating among 23 submitted conversion systems (Lorenzo-Trueba et al., 2018). Based on a parallel dataset (read speech material of ~30 minutes duration), Sprocket first learns the idiosyncratic acoustical features of a target and source speaker (i.e. MFCC, pitch) and then merges these features with the linguistic material of the source speaker. With this, we created high-quality artificial copies of four natural male target speakers (Standard German speakers, age range 19-34 years). As source speaker, we recorded speech material of one professional male speaker (Standard German speaker, 46 years).

Our experiment included a speaker familiarization task that was followed by the main experiment; the speaker speech matching task. 27 participants (mean age 25.58 years, 19 females) have first learned our four male speakers and after successful speaker familiarization (above 80%), 25 participants conducted the speaker speech matching task. This experiment comprised two task conditions: a speaker and a speech task and two voice conditions: natural and synthesized voices. In the speaker task, participants were asked to memorize the first target sound and to decide after each of the following test sentences (i.e. 12 test sentences per block, 20 blocks per condition) whether the sentence was spoken by the target speaker identity or another speaker. We presented 75 different 2-word declarative German sentences (e.g. "Er fällt.", "Er fehlt.") spoken by the previously familiarized speakers. The speech task follwed the same structure and the same sounds were presented, but this time participants matched the verbal content of a test sentence to the target sound, irrespective of who was speaking. We presented phonologically similar sentence to ensure comparable task difficulty between the speaker and speech task. The speaker speech task was conducted in an MRI environment and sounds were presented via active noise-cancelling headphones effectively reducing external MRI-induced sounds. To test for task and voice manipulation effects, we fitted linear mixed-effects models with fixed (task, voice condition) and random slopes (participants by voice condition) terms as implemented in the lme4 package (Bates et al., 2014) in the R environment (Team, 2019).

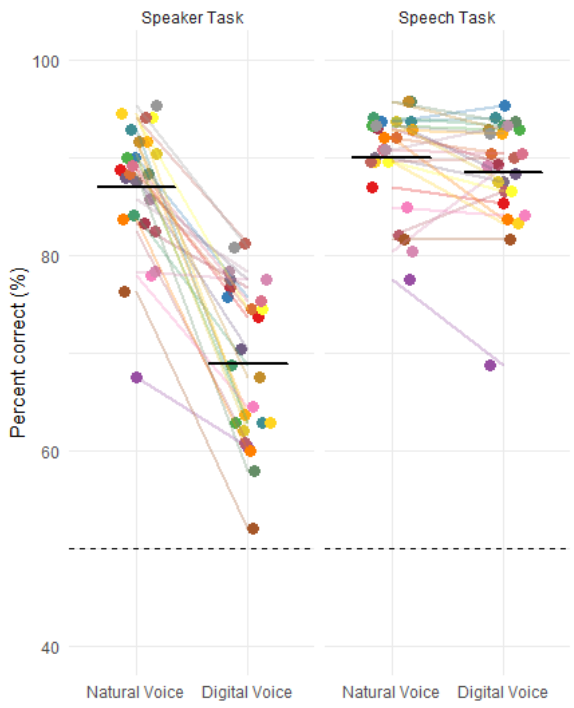

Figure 1. Individual results of the speaker and speech task. Solid black lines indicate mean percent correct for each task and voice manipulation. Dashed line indicate chance performance at 50%.

Results

In both tasks and voice manipulations, participants performed above chance level (i.e. 50%) indicating that speaker identity and speech recognition is possible for natural as well as synthesized voices (Figure 1). Next, we tested for an interaction between the speaker and speech task and the applied voice manipulation. As predicted, the interaction between task and voice manipulation (t = 9.04, p < 0.001) was significant, suggesting that the voice manipulation modulated the speaker and speech task differently. This difference is apparent in the speaker task with lower recognition performance for synthesized voices compared to the natural voices (t = 13.44, p < 0.001). Whereas in the speech task performance was not statistically different for the synthesized and natural voices (t = 1.03, p = 0.31). Overall, the model accounted for 82 % of the total variance in the data.

Conclusion

Our findings suggest that modern voice synthesis algorithms preserve socially relevant vocal attributes, especially semantic verbal information. However, we observed a marked reduction in identity recognition when listening to synthesized in contrast to natural voice identities. Our findings suggest that voice synthesis results in vocal versions of the natural speaker that are hardly accepted as natural variations of the corresponding natural speaker identity. Interestingly, speech recognition was largely unaffected by the voice manipulation. Our findings open new research avenues on social interactions in the digital age which may have important theoretical implications for current voice models but also technical application such as speech synthesize and automatic speaker recognition systems.

References

Bates, D., Maechler, M., Bolker, B., Walker, S., & Haubo Bojesen Christensen, R. (2015). lme4: Linear mixed-effects models using Eigen and S4.

Kobayashi, K., & Toda, T. (2018). *sprocket: Open-Source Voice Conversion Software*. https://doi.org/10.21437/odyssey.2018-29

Lorenzo-Trueba, J., Yamagishi, J., Toda, T., Saito, D., Villavicencio, F., Kinnunen, T., & Ling, Z. (2018). The voice conversion challenge 2018: Promoting development of parallel and nonparallel methods. arXiv preprint arXiv:1804.04262.