An inclusive approach to creating a palette of TTS voices for gender diversity

Éva Székely[1]*, Maxwell Hope*[2]

[1] *Division of Speech, Music and Hearing, Royal Institute of Technology, Sweden*

[2] *Department of Linguistics and Cognitive Science, University of Delaware, USA*

szekely@kth.se, maxhope@udel.edu

Abstract

Current text-to-speech (TTS) systems primarily support binary gender voices, often not representing the diversity of gender identities such as transgender and nonbinary individuals. This limitation is significant for users of Speech Generating Devices (SGDs) who seek voices that authentically reflect their identities. Our research introduces a novel methodology for constructing a controllable palette of gender-expansive TTS voices. Utilising recordings from 14 gender-expansive speakers, we apply Constrained Principal Component Analysis (PCA) to extract gender-independent speaker identity vectors, allowing for the modulation of acoustic Vocal Tract Length (aVTL) and emergent vocal properties in a neural TTS framework. We detail the design process influenced by community input from nonbinary SGD users, emphasising increased diversity and control over voice characteristics. We evaluate the system with a series of objective metrics to assess quality, speaker consistency and to measure the extent to which aVTL is modifiable for each speaker. Additionally, we perform and a community-based qualitative evaluation using online interviews, to evaluate the system's capacity to generate adjustable and diverse vocal qualities. This work takes a step toward more inclusive voice technology that transcends traditional gender binaries, with the aim to support self-expression for gender-expansive individuals.

References

Hope, M, and Lilley, J. (2022). Gender expansive listeners utilize a nonbinary, multidimensional conception of gender to inform voice gender perception. *Brain and Language,* vol. 224, p. 105049, 2022.

Székely, É., Gustafson, J. and Torre, I. (2023). Prosody-controllable gender-ambiguous speech synthesis: a tool for investigating implicit bias in speech perception. In Proc. Interspeech, pp. 1234–1238.

Johnson, K. (2020). The δf method of vocal tract length normalization for vowels. Laboratory Phonology, vol. 11, no. 1.

Snyder, D., Garcia-Romero, D., Sell, G., Povey, D. and Khudanpur, S. (2018) X-Vectors: Robust DNN Embeddings for Speaker Recognition. In Proc. ICASSP, pp. 5329-5333.