Kind is the Opposite of Competent: Phonetic variation in English TTS Voices

Alice Ross[1], **Professor Lauren Hall-Lew**[1], **Dr Catherine Lai**[1], **Dr Nina Markl**[2]
Affiliations: 1 - University of Edinburgh, 2 - University of Essex
Corresponding author: alice.ross@ed.ac.uk

Text-to-speech (TTS) technology now allows the fast synthesis of 'human-like' speech, reproducing cues that lead listeners to assign a social identity to the 'speaker'. The socioindexical information in TTS voices unsurprisingly replicates the socioindexical information present in the training data for those systems, along with a skewed representation of speaker types. Michel et al. (2025) show how English TTS systems are biased towards US and UK accents, even when explicitly prompted to produce African, Indian, and Australian accents, and that users consequently feel frustrated, annoyed, and excluded. TTS voices also demonstrate a biased representation of indexical associations, such as gender stereotypical voices for prompts specifying only occupation and not gender (e.g., *electrician* vs. *nurse*; Kuan & Lee 2025). Following this work, and Zellou and Holliday's (2024) call for research on bias and inequality in speech technology, we ask: What social indexicalities appear in TTS voices that only specify personality traits?

We used a popular commercial TTS platform's 'Voice Design' tool (ElevenLabs TTS v2) to generate recordings of 30 synthesised voices. Our base prompt is 'a voice that sounds [adjective]' where the adjectives include those associated with 'status' (*competent, confident, educated, intelligent, professional*) and those associated with 'solidarity' (*considerate, friendly, kind, polite, warm*). The language attitudes literature shows that voices tend to be associated with 'status' or 'solidarity' features, but usually not both, and we wondered if TTS voices would also pattern accordingly. Three different sentences were generated for each adjective, each a socially and emotionally neutral scientific fact (Lev-Ari & Keysar 2010). In separate work (Ross, et al., accepted) we analyse perceptions of the voices' regional accents. In this paper, we analyze variation in F0 minimum, maximum, mean, and range, rate of speech (syllables/second, minus pauses), and also note any variable connected speech phenomena.

The three *kind* voices, and one of the *friendly* voices, resulted in surprisingly high F0 measures, sometimes beyond the norm for the average English-speaking woman (e.g., 464 Hz). The other 26 voices show relatively similar average F0, around the norm for the average English-speaking man. The three *competent* voices all have especially low F0 measures. The *professional* voices were all fast and the *warm* voices were all slow. The two *kind* voices with extremely high F0 values were also two of the slowest voices, while the *competent* voice with the lowest F0 values also has the second fastest rate of speech in the whole sample. Rate of speech and F0 seem to interact, resulting in stylised social persona: The *kind* voices are slow and high-pitched, indexing a hegemonically feminine, almost Child-Directed Speech style, while the *competent* voices are fast and low-pitched, indexing a hegemonically masculine style. Furthermore, the 'default' voice indexes male gender, despite an equal balance of 'status' and 'solidarity' adjectives.

We also noted indicators of what might be called hypo- or hyperarticulation (despite no actual 'articulation'). For example, for utterances with intervocalic /t/, the *competent* and *confident* voices show an aspirated release, while the *kind* and *warm* voices have a flap. While the 30 sentences are not balanced for every such feature, these features enable a more nuanced interpretation of the stylization produced by this TTS system.

References

Kuan, C. Y., & Lee, H. Y. (2025, April). Gender bias in instruction-guided speech synthesis models. In *Findings of the Association for Computational Linguistics: NAACL 2025* (pp. 5387-5413).

Lev-Ari, S., & Keysar, B. (2010). Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology*, 46(6), 1093-1096.

Michel, S., Kaur, S., Gillespie, S. E., Gleason, J., Wilson, C., & Ghosh, A. (2025, June). “It’s not a representation of me”: Examining Accent Bias and Digital Exclusion in Synthetic AI Voice Services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency* (pp. 228-245).

Ross, A., Markl, N., Lai, C., & Hall-Lew, L. (accepted). The Sound of Silencing: Identities and Ideologies in Commercial Text-To-Speech. *CHI EA '26: Extended Abstracts of the ACM CHI Conference on Human Factors in Computing Systems*. Barcelona.

Zellou, G., & Holliday, N. (2024). Linguistic analysis of human-computer interaction. *Frontiers in Computer Science*, 6, 1384252.