Kind is the Opposite of Competent

Phonetic variation in English TTS voices

AUTHORS

Alice Ross
Lauren Hall-Lew
Nina Markl
Catherine Lai

AFFILIATIONS

UKRI Centre for Doctoral Training in Natural Language Processing
Institute for Analytics and Data Science, University of Essex
School of Philosophy, Psychology & Language Sciences
and School of Informatics, The University of Edinburgh
The Centre for Speech Technology Research

INTRODUCTION

- Text-to-speech (TTS) technology now allows the fast synthesis of 'human-like' speech.
- These outputs reproduce socioindexical cues based on information present in the training data (Holliday 2023).
- Several state-of-the-art models offer users the ability to 'design' custom voices using natural language prompting.
- **If a model is prompted to convey emotions or personality traits, what socioindexical cues result in the output?**

METHOD

Data: 30 synthesised voices generated using ElevenLabs TTS v2, a popular commercial TTS platform using natural language prompting.

Prompt: 'a voice that sounds [adjective]'

- **status**: *competent, confident, educated, intelligent, professional*
- **solidarity**: *considerate, friendly, kind, polite, warm*

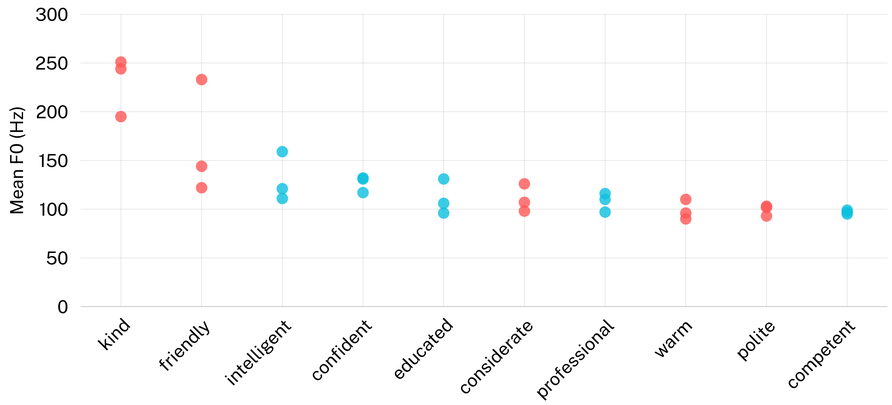
In the language attitudes literature, voices tend to be associated with 'status' or 'solidarity' features but not both. 'Solidarity', or social desirability, is often higher for socioindexically marked talkers.
Will TTS voices reflect this?

Each adjective → 3 different sentences, each a socially and emotionally neutral scientific fact (Lev-Ari & Keysar 2010).

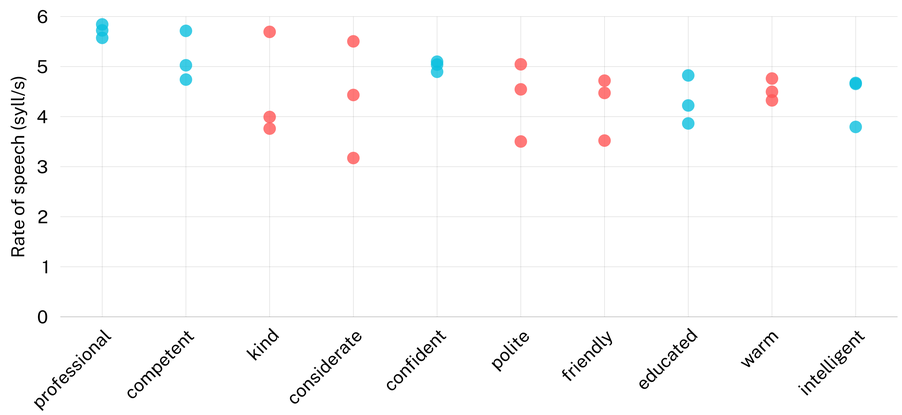
Measurements: F0 minimum, maximum, mean, and range, + RoS (syllables/sec, minus pauses).

RESULTS

Pitch: The *kind* voices, and one *friendly* voice, show very high F0 measures. The other 26 show an average F0 around the average for English-speaking men. The *competent* voices have particularly low F0.



Rate of speech: The *professional* voices were fast; the *warm* voices were slow. The two *kind* voices with extremely high F0 were also two of the slowest, while the *competent* voice with the lowest F0 values also has the second fastest RoS.



ADDITIONAL RESULTS

- For those utterances with intervocalic /t/, the *competent* and *confident* voices have an aspirated release, while the *kind* and *warm* voices have a flap.
- In other work (Ross et al. 2026), short extracts of the voices were evaluated by 60 North American English listeners. 93% of the voices were perceived as white, 60% as US-accented and 30% as UK-accented.

DISCUSSION

- More than 80% of the voices had a normative male F0.
- Rate of speech and F0 co-construct recognizable styles.
  - 2 of the system's 3 *kind* voices are slow and high, indexing a hegemonically feminine (CDS?) style.
  - All 3 *competent* voices are fast and low-pitched, indexing a hegemonically masculine style.
- This TTS model disproportionately reproduces speech indexing whiteness, maleness, and US- or UK-English accents, even when none of these features are specified.
- While some indexes of 'status' or 'solidarity' appear in outputs, so do some troubling gender stereotypes.

REFERENCES

- Holliday, N. (2023). Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication*, 8.
- 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)
- Ross, A., Markl, N., Lai, C., and Hall-Lew, L. (2026). 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*

UK Research and Innovation
UKRI Centre for Doctoral Training in Natural Language Processing
The Centre for Speech Technology Research
The University of Edinburgh School of Philosophy, Psychology and Language Sciences
University of Essex
The University of Edinburgh