The Sound of Silencing

Identities and Ideologies in Commercial Text-To-Speech

Alice Ross
Nina Markl
Catherine Lai
Lauren Hall-Lew

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

Whose voices are reproduced by a promptable TTS system when no demographic information is specified?

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

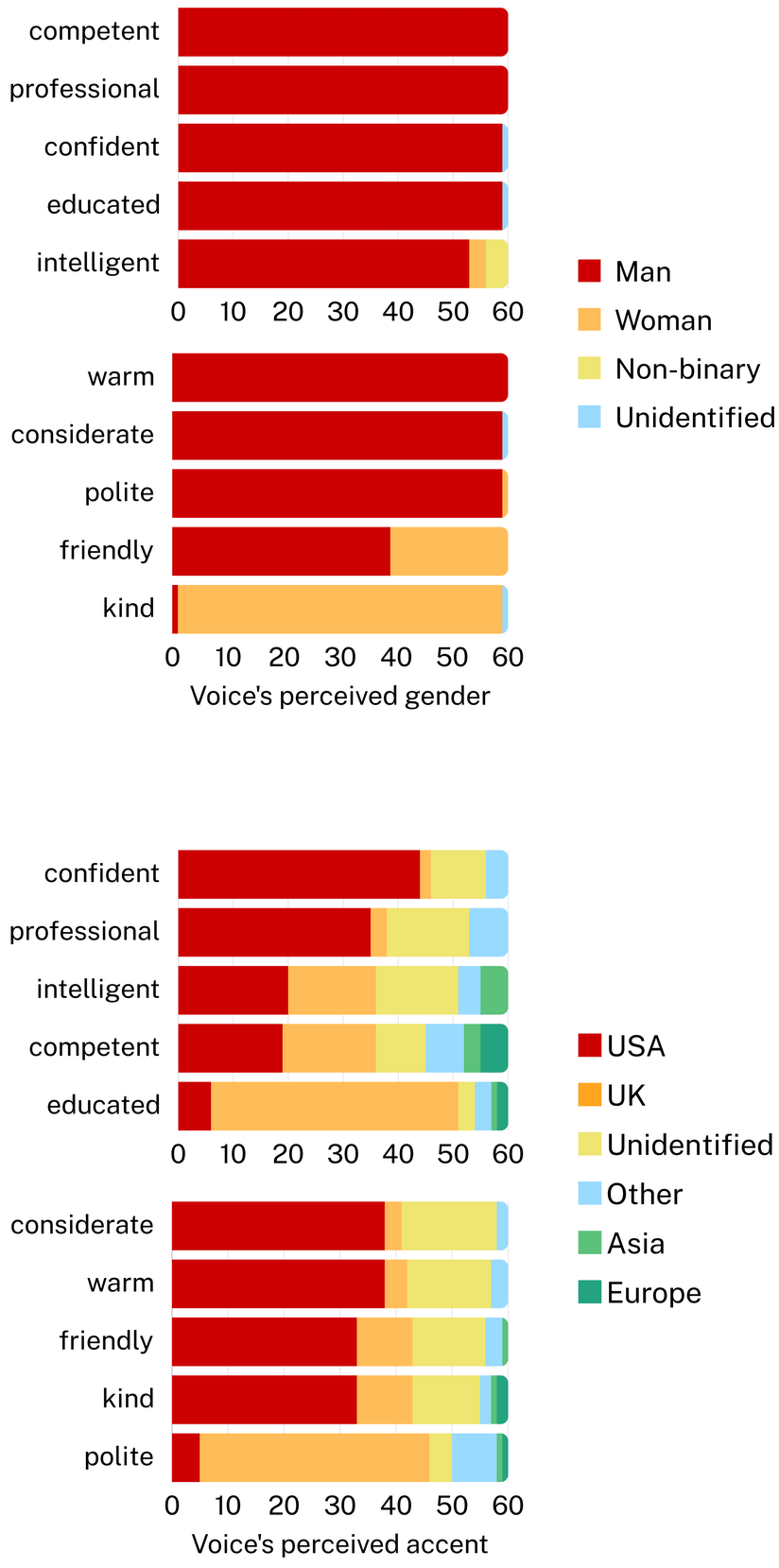
Prompt: 'a voice that sounds [adjective]' +
status: competent, confident, educated, intelligent, professional
solidarity: considerate, friendly, kind, polite, warm

Participants: 60 English-speaking Prolific users (26 men, 34 women) from US and Canada

The set of generated voices lacks diversity, regardless of the adjective used. Listeners perceived the 'speakers' as primarily:

93%	86.6%	60%
white	men	US accented

We see systematic links between some adjectives and demographic groups: *kind* is associated with (white, American and British) women; *polite* and *educated* with white British men.



overrepresentation of white, male, LI-accented speech...
or erasure of all the voices that are not reproduced?

audio samples

UK Research and Innovation
UKRI CENTRE FOR DOCTORAL TRAINING
The Centre for Speech Technology Research
THE UNIVERSITY of EDINBURGH School of Philosophy, Psychology and Language Sciences
University of Essex