IDEOLOGY, DISCRIMINATION, AND BIAS IN SYNTHESISED VOICE DESIGN

Whose voices are replicated? Which (types of) voices are unheard in speech technology? Which are assumed to be 'standard' and appropriate?

WORK IN PROGRESS

AUTHORS

Alice Ross
Dr. Nina Markl
Prof. Lauren Hall-Lew
Dr. Catherine Lai

alice-ross.github.io

AFFILIATIONS

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

INTRODUCTION

Contemporary TTS is frequently claimed to be extremely 'natural' and, in some contexts, indistinguishable from human speech. Voice interfaces using such synthesised speech (often combined with LLMs) are increasingly adopted in a wide range of contexts. We note a lack of diversity in popularly used English-speaking TTS voices, and caution that decisions taken in the design and deployment of voice interfaces risk perpetuating, or even exacerbating, existing social biases. Drawing on sociolinguistic theory, we design an experiment to investigate these topics in a leading commercial TTS system. We aim to provoke further work and conversations around applying linguistic knowledge to human-computer interaction with speech technology.

WHAT DO WE MEAN BY 'DISCRIMINATION'?

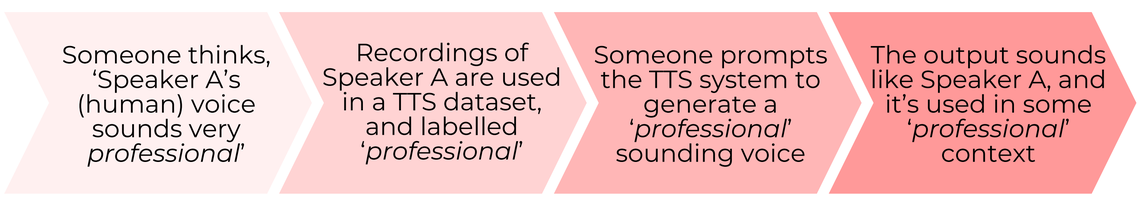
In the context of speech perception, *discrimination* has two meanings:

1. listeners' ability to notice and differentiate characteristics of voices (as with classification algorithms, we can determine whether an unseen speaker is familiar or a stranger; whether they are likely to be a man, woman, or child, etc...)

2. the lived reality of social hierarchies privileging some groups above others (in this sense, discrimination could lead an employer to choose one candidate over another, equally qualified for a job, because the first speaks fluently in an accent that the interviewer finds familiar and pleasant).

WHY DO WE CARE?

The links between voices and social constructs or personal qualities are indexical and subjective: there is no ground truth about what type of voice is *professional* (or, indeed, *male*). TTS design choices can become part of the process of reproducing biases and stereotypes, sometimes harmfully.



Good news for Speaker A. But what about other speakers, whose voices sound very different?

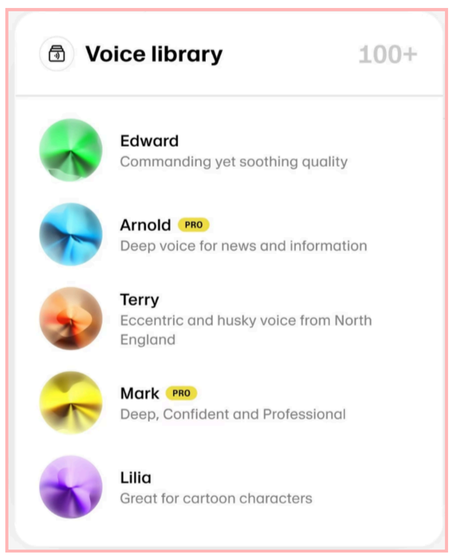
RESEARCH QUESTIONS

- **What social and personal characteristics do listeners identify when they hear 'human-like' TTS voices?**

- **Do linguistic ideologies and biases related to these characteristics influence the design, selection and popular use of these voices?**

CASE STUDY

Listeners can fairly accurately identify the gender, age, race, and, in many cases, regional accent of an unseen, unfamiliar speaker. Are cues to these identity features also audible in 'human-like' TTS voices?



ElevenLabs' home page illustrates some possible biases at work!
In this presumably curated selection of exemplar voices, we see:

🤔 5/5 European names

🤔 4/5 masculine names
and the feminine one is not 'commanding' or 'professional', but 'great for cartoon characters'

🤔 1/5 specifies a regional origin:
North England (marked in a British context for historically lower socioeconomic status), and the voice is 'eccentric'. Where are the others from?

To investigate which types of voices are treated as 'standard' in an industry-leading commercial TTS generation system, we can use prompts with no demographic/identity features specified: how does 'a voice that sounds professional' sound?

Stereotypes about demographic groups can often be classified in a two-dimensional space, trading off beliefs about (perceived) *competence* and *warmth*:

We will generate speech samples using the prompt: 'A voice that sounds [ADJECTIVE]' with adjectives from these two lists, then ask listeners for their judgments of the imagined speaker's gender, age, race, and accent. If listener agreement is relatively high (e.g. if 'professional' voices sound like middle-aged, white men), our results will provide evidence about characteristics hidden beneath the ideology of a 'standard' TTS voice.