The Sound of Silencing: Identities and Ideologies in Commercial Text-To-Speech

Alice Ross
UKRI CDT in NLP
University of Edinburgh
Edinburgh, Lothian, United Kingdom
alice.ross@ed.ac.uk

Nina Markl
University of Essex
Essex, United Kingdom
nina.markl@essex.ac.uk

Catherine Lai
School of Philosophy, Psychology and Language Sciences
University of Edinburgh
Edinburgh, United Kingdom
c.lai@ed.ac.uk

Lauren A. Hall-Lew
Linguistics and English Language
University of Edinburgh
Edinburgh, United Kingdom
lauren.hall-lew@ed.ac.uk

Abstract

Text-to-speech (TTS) technology allows the synthesis of speech that is frequently described as highly 'natural' and, in some contexts, indistinguishable from human speech. Voice interfaces using such synthesised speech are increasingly encountered in a wide range of contexts. Recognising that listeners are likely to hear human-like voices as belonging to different demographic/social groups, and that these social judgments exist within ideological frameworks, 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 upon sociolinguistic theory, we carry out a novel experiment to investigate these issues in a leading commercial TTS system, concluding that the system's output disproportionately reproduces white, male, US-accented speech when prompted to convey competence. This work aims to encourage further research applying sociolinguistic knowledge to the study of human-computer interaction with speech technology.

CCS Concepts

• **Human-centered computing** → **Ubiquitous and mobile devices**; **Natural language interfaces**; • **Computing methodologies** → *Natural language processing*; • **Social and professional topics** → **User characteristics**.

Keywords

voice user interfaces, speech synthesis, voice AI, language ideology, diversity and inclusion

ACM Reference Format:
Alice Ross, Nina Markl, Catherine Lai, and Lauren A. Hall-Lew. 2026. The Sound of Silencing: Identities and Ideologies in Commercial Text-To-Speech.

CHI EA '26, Barcelona, Spain

ACM ISBN 979-8-4007-2281-3/26/04
https://doi.org/10.1145/3772363.3798657

In *Extended Abstracts of the 2026 CHI Conference on Human Factors in Computing Systems (CHI EA '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3772363.3798657

1 Introduction

Contemporary applications of speech technology go well beyond its traditional use cases in accessibility: very human-like text-to-speech (TTS) voices are increasingly encountered in contexts such as customer service, education and training, social media and entertainment [13]. Several state-of-the-art models now offer users the ability to 'design' a custom synthesised voice using natural language prompting. Like other machine learning (ML) applications, including large language models and image generation models, these technologies are liable to reproduce and perpetuate existing social biases and stereotypes. Stereotypes disseminated via media are linked to real-life negative effects on members of stereotyped groups [2], and given that speech technology plays a growing role in today's social and mainstream media, we are motivated to interrogate the ideologies that it may be reinforcing.

Our project applies the sociolinguistic concepts of indexicality and markedness to the study of commercial TTS voices. We recognise that human-like synthesised voices reproduce cues that lead listeners to assign a social identity to the 'speaker' [56], that these imagined speakers can be perceived as belonging to different demographic/social groups [42], and that these social judgments exist within ideological frameworks [21]. Following [3], we therefore ask: which linguistic norms are upheld in synthetic voice development? Whose voices are *unheard* in speech technology, and whose are assumed to be 'standard', acceptable and appropriate? This work introduces a paradigm, informed by sociolinguistic theory, for investigating biases in the development and description of synthesised voices.

2 Background

Human-like TTS voices are often chosen for user interfaces due to designers' beliefs – or assumptions – that this reflects users' preferences [30]. While there is evidence to suggest that human-like synthesised voices are evaluated as more pleasant, credible, and likeable than robotic voices, approval generally does not reach the level of ratings given to actual human voices [10, 30, 54]. Research

in this field also tends to gloss over important questions: *which* humans' voices do we replicate? And which humans decide what sounds 'pleasant', 'credible', and so on? These questions concern us because human speech is characterised by variation, and much of the variation that we hear and discriminate is indexically linked to personal features of the speaker [12].

2.1 Markedness and discrimination

Listeners can quickly identify an unseen, unfamiliar speaker's gender, age and race with reasonable accuracy, and in many cases, a speaker's accent reveals their first language and/or where they grew up [7, 32, 49, 57, 58]. This ability to infer characteristics from speech perception extends to synthesised as well as human speech [21, 25, 34, 54]. In line with the principle that computers are social actors [44], existing language attitudes and social biases also appear to be reproduced, even when listeners know that they are not hearing human speakers; these include accent bias [25], misogyny [24], and racialisation [21].

2.2 Competent or warm?

Research in sociology shows that stereotypical beliefs about social groups can be classified in a two-dimensional space trading off the qualities of (perceived) *competence* and *warmth* [16, 18], identifying cross-cultural patterns of ascribing higher status and agency to some groups (e.g. men, rich people), and social desirability or cooperativeness to subordinated others (e.g. women, poor people). Bias and prejudice based on such stereotypes can result in unfair expectations and treatment of individuals at school, work, and beyond, contributing to social problems like the gender pay gap [41]. [5] outlines how omitting groups of people from public-facing representation, e.g. in mass media, can negatively affect those groups by rendering them unimportant or unacceptable. [9] explains how language attitudes can impact outcomes in education and employment, health care, and legal processes. Further, we acknowledge that an unwillingness to focus on race in human-computer interaction (HCI) communities can mask and reinforce racial disparities [46]. One insidious way that racist and classist discrimination often manifests is through accent bias [35, 39]: 'marked' talkers may be subjectively judged as less intelligible or less competent, even when there is no objective evidence to support this.

2.3 Speech technology in social context

The last few years have provided compelling evidence that new language technologies can influence human speech production, both in experimental settings and 'in the wild' [47, 59]. This strand of research leads us to the related question: might they influence the way we *hear* and construct social meanings based on human voices?

Despite many efforts, there is no agreed-upon, universal correlation between specific voice qualities and the social meanings they convey [28]; such meanings seem to depend on cultural norms and/or factors about the listeners. Indexicality [9] – the process linking individuals' language use to their social identity – is typically a two-way process involving input from both speakers and listeners. Speakers use a particular variation that may convey their identification with a group; listeners notice the variation and use that information to place the speaker within their own social world. In the case of TTS, there is no human speaker to whom we can ascribe the expression of identity, but listeners still (consciously or otherwise) perceive speech features and their associated social meanings. In the framework of the commercial voice library, a human-like TTS voice with certain characteristics may be assigned text labels such as *professional* or *suitable for customer service*; these labels are applied as though they are objective; and, circularly, prompting this system to generate a 'professional' voice will be likely to result in a similar output. Our position is that there is no acoustic feature which inherently makes a voice professional. Natural language prompting in human-like TTS generation, though, associates acoustic features with personality descriptors (for example, in our dataset, a very high F0 appears to be linked with *kindness*), with the potential to reinforce stereotyped ideologies. Therefore, we explore cues to demographic group identity that listeners may hear in natural language prompted TTS voices where no such information was specified in the prompt.

3 Related work

First steps have recently been taken in research to quantify bias in promptable TTS models [29, 43]. However, existing work often lacks sociolinguistic context and is limited by considering only one possible source of bias: gender, oversimplified to a binary category and classified using so-called 'automated gender recognition' models. This approach is out of step with the prevailing view in social science that gender (like other attributes) is socially constructed, and the understanding that speaker identity is not static [45, 53]. Researchers also propose superficial 'debiasing' strategies that would mask, not remove, systemic bias [19]. Contributing a new perspective to the topic, we approach social identity as intersectional and multidimensional, and contend that it is important to consider other salient factors, such as race/ethnicity, language background, and age, along with gender. This is especially crucial if we are to avoid normalising *whiteness as default*, a documented problem in human-computer interaction [6, 46].

Several scholars have examined social perceptions of TTS in other contexts. Nicole Holliday's pioneering work on listeners' judgments of US English Siri voices [21] highlights the influence of existing cultural stereotypes, such as the disenfranchised but entertaining young Black man. Another notable 'Black' TTS voice, the Spotify DJ, is discussed in [51], where 'his noticeable urban accent' is said to lend 'a "cool" vibe'; this perception may reflect a similar social bias about Black talkers. Using qualitative interview data, [50] report that out-group (non-Black) listeners failed to identify an African American English TTS voice as Black because it did not conform to their stereotyped expectations. Investigating accent in commercial English-speaking TTS systems, [42] find that American and British English accents are overrepresented, even when inputs specify African, Indian, and Australian accents, resulting in users feeling frustrated, annoyed, and excluded. The reproduction of 'standard' language ideologies through another speech technology, automatic speech recognition (ASR), is interrogated in [27, 37]. [56] call for examination of the socioindexical influence of synthesised speech in interaction, noting the media's role in reifying language ideologies; and [60] identify the need to explore bias and inequality

in speech technology through the lens of challenging linguistic and social power. With that aim, we introduce an experiment to interrogate which (types of) voices are reproduced when we prompt a commercial TTS system with no specific demographic information, and consider the potential repercussions of overuse of so-called standard varieties and exclusion of atypical speakers.

4 Experiment

Our experiment takes inspiration from [21] on perceptions of US English Siri voices, and from matched-guise testing, a well-established experimental method used in sociolinguistics to explore listeners' attitudes to speakers of marked language varieties [31].

Audio samples used in this study were created using the model ElevenLabs TTS v2, offered by a popular commercial TTS platform with over 5 million registered users as of 2025 [11]. We chose to study voices produced by a large speech model (LSM) as opposed to voice assistants (VA) as in [21], because these are especially interesting in social contexts: commercial TTS is increasingly used in interactions that would previously have been human-human, such as customer service, and becoming 'the voice of' public-facing entities like websites and museums. This contrasts with typical VA interactions, like setting timers and controlling appliances [36], which would otherwise be carried out using a visual interface. We chose Elevenlabs over other TTS models because of its popularity and availability: our experiment is easily replicable without specific training requirements. Additionally, marketing for this particular system highlights its supposed controllability and customisability; ElevenLabs' website [14] advertises its '[i]nfinite selection of AI voices' and proposes that users can create the ideal 'speaker' to suit their needs.

Using ElevenLabs' 'Voice Design' tool, we generated and recorded 30 synthesised voices. Our base prompt is 'a voice that sounds [adjective]', with a set of ten adjectives selected from existing language attitudes literature [8, 15, 20, 40, 55]. Half of the adjectives in the set are associated with 'status' (*competent, confident, educated, intelligent, professional*) and half are associated with 'solidarity', or social desirability (*considerate, friendly, kind, polite, warm*). Three different voices were generated for each adjective by re-running each prompt three times, with the spoken text each time a socially and emotionally neutral scientific fact [33]. We then edited the resulting 30 recordings to 2-3 words, around 1 second, each, allowing us to minimise any content effects and collect a large number of judgments while avoiding participant fatigue.

Sixty listeners were recruited via Prolific, meeting the following inclusion criteria: L1 English speakers who grew up in and were living in the US or Canada, had no hearing impairment, and had successfully completed ≥10 previous experiments on Prolific with ≥90% approval. Each participant listened to a subset (10) of the voices, balanced for adjective used, and for each they were asked to describe the speaker's: age (bands, categorical), gender (multiple choice), racial identity (multiple choice), accent (free text). Participants also provided demographic information about themselves, and they could optionally write in further comments about each voice and the study overall. The participants included 26 men, 34 women, 0 other genders. Their age groups ranged from 18 to 65+, and their racial identity was recorded as white (43), Black (9), Asian (4), Latino/-a/-x (3), and mixed (1). The median time taken to complete the experiment was 12 minutes, and the participants were paid 4 GBP, in line with Living Wage where the study was conducted.

In support of our methodology, we note that although listeners were always given the opportunity to select an 'other' option and input free text ('Let me type...'), for example when they felt they could not confidently answer the question, this was very rarely used (n = 9 of 600 observations, or 1.5%). We focus on how the voices are perceived by humans because perception is key in social interaction. Listeners' own identities may influence their perceptions, although we note that Holliday found no significant effects of listeners' demographic characteristics in [21]; we leave further investigation of this for future work. To identify salient characteristics in the perception of identity from speech, we had first conducted a norming study in which 15 participants listened to a subset (10) of the recordings and described, in free text, how they imagined speakers. 119 of these 150 responses mentioned sex or gender, including using gendered pronouns. 88 responses mentioned accent or place of origin; 82 age; 38 other visual features (height, hair, clothing), and 33 race or ethnicity. Hesitancy to discuss race is expected and does not negate our motivation outlined in S3 for including it.

5 Results

Fig. 1 shows a summary of the results grouped by adjective set. The majority of the voices (86.6%) were overwhelmingly judged to be men's voices. In the 'status' subgroup, all 15 voices (100%) were predominately classified as men's voices, while in the 'solidarity' subgroup, 11 of 15 (73%) were perceived as men's. The 'status' voices were more likely to be perceived as men than the 'solidarity' voices, and a generalized linear mixed effects model showed that this difference was significant (**p>0.001**). For race, the voices in both groups are strongly skewed toward being perceived as white: 28 of the 30 voices were classified as white by > 50%, and all by > 40%, of listeners. We see slightly more observations of other racial identities in the 'status' set than the 'solidarity' set (**p<0.05**). The majority of voices in both sets were perceived as US- or UK-accented. 18 (60%) of the 30 voices are perceived to be US-accented, and 9 (30%) UK-accented. The distribution of these accents across the two groups showed **no significant difference**. The most common age band attributed to 'status' voices was 35-44, compared to 25-34 for 'solidarity'; however, the patterns of responses to this question were very similar for all but a few voices, and so we do not discuss this result further here.

6 Discussion

Overall, regardless of which adjectives were used in prompts, the set of voices produced lacks diversity: whiteness, masculinity, and US and UK accents are clearly overrepresented. Examining specific items in our dataset yields some evidence of gender and accent bias, as discussed below.

6.1 Gender

Voices perceived as male are clearly overrepresented in the dataset, with 26 of 30 voices perceived as men's. The remaining four are all three 'kind' voices and one 'friendly'; that is, the prompt 'a voice that sounds *kind*' consistently produced voices that were perceived

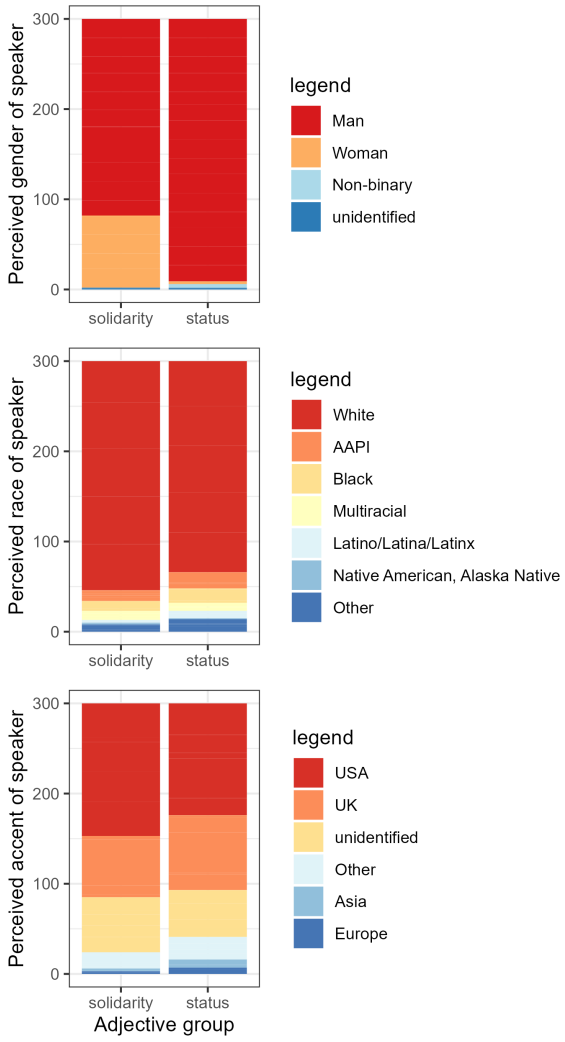

Figure 1: Counts of responses to 'How would you describe the speaker's...' (top-bottom:) gender; racial identity; accent. Responses are grouped by the set of adjectives used in prompts to generate the voices: 'status' (competent, confident, educated, intelligent, professional) and 'solidarity' (considerate, friendly, kind, polite, warm). Each set contains ten adjectives which were each used three times, for a total of 30 voices; each listener heard ten of the voices. Note that some responses have been recoded for easier visualisation: 'AAPI' combines two racial identity categories, 'Asian' and 'Native Hawaiian or other Pacific Islander'; accent data were recoded from free text inputs into broad categories.

to be women's voices. This finding indicates that there may be a systematic association between women and these socially desirable qualities in the model, reflecting an attitude of 'benevolent sexism' which is both common and more often seen as acceptable compared to hostile sexism [52]. In terms of gender, then, the 'solidarity' set is less homogeneous than the 'status' set, but still dominated by men's voices, well in excess of the proportion of people who are men. We note that the voices perceived as women's voices were also perceived as white and US- or UK-accented: individual voices differed from the overall dataset according to gender or regional accent, but never both.

6.2 Race

Almost all of the voices were predominantly classified as white, suggesting that whiteness, like maleness, is heavily overrepresented in our generated dataset. We note that similar results arose in studies specifically focusing on Black TTS voices [21, 50]; whiteness may be unmarked and normalised in AI contexts more generally, at least for these listeners [6]. It should also be noted that our participants' own demographics (72% white) do not replicate the diversity of the population in the USA (62% white) [4] or globally. Only six of 30 voices were perceived as non-white by more than five listeners. Of these, five belong to the 'status' set and only one (*warm-c*) to 'solidarity', and all were perceived as men. This finding calls for further research, but it is possible that a view of (some) minoritised speakers as 'intelligent' and 'confident', given the context of speaking about science, may be reflected in the data. If whiteness is indeed highly correlated with social desirability in the model's output, this is cause for concern.

6.3 Accent

Listeners were asked to describe the speakers' accents in free text. We recoded responses into six broad levels: **USA** (e.g. 'Boston', 'Southern', 'Standard American', 'US'), **UK** (e.g. 'British', 'English'), **Asia** (e.g. 'Chinese', 'Japanese'), **Europe** (e.g. 'German', 'European'), **unidentified** (e.g. 'none', 'fluent', 'outdoorsy'), and **Other** (e.g. 'Australian', 'Mexican', 'Egyptian', and 'South African'). Instances of the 'Other' category were few (n = 43, 7% of observations). US and UK accents are clearly overrepresented in the dataset, together accounting for almost all of the voices. Those predominantly judged as UK-accented include all three 'polite' voices and all three 'educated' voices, along with *intelligent-a*, *friendly-a* (also described as Australian or Chinese by some listeners), and *competent-c* (also described as Australian or South African); as with gender, this finding suggests that specific adjectives are associated with Britishness in this model. Listener agreement is low on the accents of the remaining three voices.

7 Conclusion

ElevenLabs' TTS v2 model and 'Voice Design' system produced a set of synthesised voices given prompts that contain no demographic information, only positive personality traits. Analysing listeners' evaluation of these voices, we conclude that the model is disproportionately likely to reproduce white, male, US- or UK-accented speech when prompted to convey competence and other positive traits. Further, the output reveals apparently systematic associations between certain adjectives and demographic groups: *kind* with women; *educated* and *polite* with British men.

This skewed representation of (synthetic) speakers and their traits is not an isolated or neutral quirk. It arises from the collection, labelling and use of data in the system [38], and it exists in

the context of an era characterised by increasing dissemination of dangerous white supremacist, anti-migrant, and misogynist ideas [26]. If the use of TTS systems like this one continues to increase, particularly in education and entertainment, it seems inevitable that the overall set of voices an individual is exposed to will become homogenised in future. What we call 'overrepresentation' here can be alternatively construed as *erasure* of all the voices that are not reproduced (see [17]): considering the space of human(-like) voices, a subset of unnaturally fluent, Mainstream American-accented, white, and predominantly male voices are becoming normalised. In addition to the destructive impact of AI systems on voice acting and voiceover work [1], this new normal would influence the cognitive development and language attitudes of future generations, particularly towards already marginalised varieties and disfluent speakers.

Future work along this line of enquiry could compare the output of various TTS models (with different training data) given non-specific prompts, and focus on exploring the reasons behind the results. The overrepresentation of particular demographics (white, American, male) seen in this work suggests that 1) these traits are associated with positive value judgments in the NL promptable TTS system, and/or 2) they are very overrepresented in the model's training data, making most of the output sound like them. These explanations could be tested with further investigation. It would also be instructive to conduct more fine-grained analysis of *which* varieties within American English are represented, and of prosodic dimensions like speech rate and pitch range. In this way, we could examine more specific personas and styles, going beyond broad categories like 'a woman's voice'. Further research could also explore user experiences and different design routes, focusing on fairness and empowerment, for this and similar systems.

Bias in machine learning models like this one is not under-researched. However, work on the topic can be deepened and strengthened with analyses informed by social awareness of how biases and stereotypes contribute to material inequities and harms. Scholars including [3, 22, 23, 48, 56] are leading this move in the right direction. Along with them, our aim is to build and improve upon existing work by paying attention to the 'human' in human-computer interaction and considering potential repercussions of the decisions made in these systems' design and deployment.

Acknowledgments

The authors would like to thank Tanvi Dinkar, Gavin Abercrombie, Christian Ilbury, Éva Székely, Gustav Eje Henter, Cliodhna Hughes, and Cara Wilson for their insights and advice on this project. This work is supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1).

References

[1] Shm Almeda, Robin Netzorg, Isabel Li, Ethan Tam, Skyla Ma, and Bob Tianqi Wei. 2025. Labor, Power, and Belonging: The Work of Voice in the Age of AI Reproduction. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*. 1238–1249.

[2] Markus Appel and Silvana Weber. 2021. Do mass mediated stereotypes harm members of negatively stereotyped groups? A meta-analytical review on media-generated stereotype threat and stereotype lift. *Communication Research* 48, 2 (2021), 151–179.

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[4] United States Census Bureau. 2020. *Race and Ethnicity in the United States: 2010 Census and 2020 Census*. United States Census Bureau. https://www.census.gov/library/visualizations/interactive/race-and-ethnicity-in-the-united-state-2010-and-2020-census.html Accessed: 9 January 2026.

[5] Julie Carpenter. 2019. Why project Q is more than the world's first nonbinary voice for technology. *Interactions* 26, 6 (2019), 56–59.

[6] Stephen Cave and Kanta Dihal. 2020. The whiteness of AI. *Philosophy & Technology* 33, 4 (2020), 685–703.

[7] Cynthia G Clopper and David B Pisoni. 2004. Some acoustic cues for the perceptual categorization of American English regional dialects. *journal of phonetics* 32, 1 (2004), 111–140.

[8] Paolo Coluzzi. 2016. Attitudes towards Malay, English and Chinese among Malaysian Students: A Matched Guise Test. In *English in Malaysia*. Brill, 87–101.

[9] Justin T Craft, Kelly E Wright, Rachel Elizabeth Weissler, and Robin M Queen. 2020. Language and discrimination: Generating meaning, perceiving identities, and discriminating outcomes. *Annual Review of Linguistics* 6, 2020 (2020), 389–407.

[10] Scotty D Craig and Noah L Schroeder. 2017. Reconsidering the voice effect when learning from a virtual human. *Computers & Education* 114 (2017), 193–205.

[11] Hector Craigson. 2025. *Decoding ElevenLabs Revenue: A Deep Dive into Their Financial Growth*. https://techannouncer.com/decoding-elevenlabs-revenue-a-deep-dive-into-their-financial-growth/ Accessed: 30 November 2025.

[12] Penelope Eckert. 2008. Variation and the indexical field 1. *Journal of sociolinguistics* 12, 4 (2008), 453–476.

[13] ElevenLabs. 2024. *Customer Stories*. ElevenLabs. https://elevenlabs.io/blog?category=customer-stories Accessed: 27 April 2025.

[14] ElevenLabs. 2025. *Text to Speech (Speech Synthesis)*. ElevenLabs. https://elevenlabs.io/text-to-speech/ Accessed: 30 November 2025.

[15] Víctor Fernández-Mallat and Max Carey. 2017. A matched-guise study on L2, heritage, and native Spanish speakers' attitudes to Spanish in the State of Washington. *Sociolinguistic Studies* 11, 1 (2017), 175–198.

[16] Susan T Fiske. 2017. Prejudices in cultural contexts: Shared stereotypes (gender, age) versus variable stereotypes (race, ethnicity, religion). *Perspectives on psychological science* 12, 5 (2017), 791–799.

[17] Susan Gal and Judith T. Irvine. 2019. *Signs of Difference: Language and Ideology in Social Life*. Cambridge University Press. doi:10.1017/9781108649209

[18] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[19] Hila Gonen and Yoav Goldberg. 2019. Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 609–614.

[20] Karolina Hansen, Tamara Rakić, and Melanie C Steffens. 2017. Competent and warm? *Experimental Psychology* (2017).

[21] Nicole Holliday. 2023. Siri, you've changed! Acoustic properties and racialized judgments of voice assistants. *Frontiers in Communication* 8 (2023), 1116955.

[22] Nicole R Holliday. 2025. Socially prescriptive speech technologies: Linguistic, technical, and ethical issues. *The Journal of the Acoustical Society of America* 158, 6 (2025), 4361–4369.

[23] Maxwell Hope. 2024. *Creation, Perception, and Use of Gender Expansive Synthetic Voices*. University of Delaware.

[24] Ryan Blake Jackson, Tom Williams, and Nicole Smith. 2020. Exploring the role of gender in perceptions of robotic noncompliance. In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*. 559–567.

[25] Allison Jones and Georgia Zellou. 2024. Voice accentedness, but not gender, affects social responses to a computer tutor. *Frontiers in Computer Science* 6 (2024), 1436341.

[26] Catarina Kinnvall and Ted Svensson. 2022. Exploring the populist 'mind': Anxiety, fantasy, and everyday populism. *The British Journal of Politics and International Relations* 24, 3 (2022), 526–542.

[27] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. 2020.

Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences* 117, 14 (2020), 7684–7689.
[28] Jody Kreiman. 2024. Information conveyed by voice quality. *The Journal of the Acoustical Society of America* 155, 2 (2024), 1264–1271.
[29] Chun-Yi Kuan and Hung-yi Lee. 2025. Gender bias in instruction-guided speech synthesis models. In *Findings of the Association for Computational Linguistics: NAACL 2025*. 5387–5413.
[30] Katharina Kühne, Martin H Fischer, and Yuefang Zhou. 2020. The human takes it all: Humanlike synthesized voices are perceived as less eerie and more likable. evidence from a subjective ratings study. *Frontiers in neurorobotics* 14 (2020), 593732.
[31] Wallace E Lambert, Moshe Anisfeld, and Grace Yeni-Komshian. 1965. Evaluation reactions of Jewish and Arab adolescents to dialect and language variations. *Journal of personality and social psychology* 2, 1 (1965), 84.
[32] Nadine Lavan, Paula Rinke, and Mathias Scharinger. 2024. The time course of person perception from voices in the brain. *Proceedings of the National Academy of Sciences* 121, 26 (2024), e2318361121.
[33] Shiri Lev-Ari and Boaz Keysar. 2010. Why don't we believe non-native speakers? The influence of accent on credibility. *Journal of experimental social psychology* 46, 6 (2010), 1093–1096.
[34] Kevin D Lilley, Ellen Dossey, Michelle Cohn, Cynthia G Clopper, Laura Wagner, and Georgia Zellou. 2024. Social evaluation of text-to-speech voices by adults and children. *Speech Communication* (2024), 103163.
[35] Rosina Lippi-Green. 2012. *English with an accent language, ideology, and discrimination in the United States* (2nd ed.. ed.). Routledge, London ; New York.
[36] Irene Lopatovska, Katrina Rink, Ian Knight, Kieran Raines, Kevin Cosenza, Harriet Williams, Perachya Sorsche, David Hirsch, Qi Li, and Adrianna Martinez. 2019. Talk to me: Exploring user interactions with the Amazon Alexa. *Journal of Librarianship and Information Science* 51, 4 (2019), 984–997.
[37] Nina Markl. 2022. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 521–534.
[38] Nina Markl. 2022. Mind the data gap (s): Investigating power in speech and language datasets. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. 1–12.
[39] Nina Markl and Catherine Lai. 2023. Everyone has an accent. In *Interspeech 2023*. ISCA, 4424–4427.
[40] Timothy McTiernan and Robert Knox. 1979. Irish students' stereotypes about some national and subnational groups within Ireland and Great Britain. *Social Behavior and Personality: an international journal* 7, 1 (1979), 49–64.
[41] Loes Meeussen, Aster Van Rossum, Colette Van Laar, and Belle Derks. 2022. *Gender Stereotypes: What Are They and How Do They Relate to Social Inequality?* Springer International Publishing, 79–86. doi:10.1007/978-3-030-93795-9_7
[42] Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. 2025. "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*. 228–245.
[43] Aarish Shah Mohsin, Mohammad Nadeem, Shahab Saquib Sohail, Tughrul Arsalan, Mandar Gogate, Nasir Saleem, and Amir Hussain. 2025. Investigating Gender Bias in Text-to-Audio Generation Models. In *Proc. Interspeech 2025*. 3369–3373.
[44] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 72–78.
[45] Robin Netzorg, Alyssa Cote, Sumi Koshin, Klo Vivienne Garoute, and Gopala Krishna Anumanchipalli. 2024. Speech After Gender: A Trans-Feminine Perspective on Next Steps for Speech Science and Technology. In *Proc. Interspeech 2024*. 3075–3079.
[46] Ihudiya Finda Ogbonnaya-Ogburu, Angela DR Smith, Alexandra To, and Kentaro Toyama. 2020. Critical race theory for HCI. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–16.
[47] Rachel Ostrand, Victor S Ferreira, and David Piorkowski. 2023. Rapid Lexical Alignment to a Conversational Agent. In *Proc. Interspeech 2023*. 2653–2657.
[48] Ameena L Payne, Tasha Austin, and Aris M Clemons. 2024. Beyond the front yard: The dehumanizing message of accent-altering technology. *Applied Linguistics* 45, 3 (2024), 553–560.
[49] Massimo Pettorino, Antonella Giannini, et al. 2011. The Speaker's Age: A Perceptual Study.. In *ICPhS*. 1582–1585.
[50] Claudio Santos Pinhanez, Raul Fernandez, Marcelo Carpinette Grave, Julio Nogima, and Ron Hoory. 2024. Creating an African American-Sounding TTS: Guidelines, Technical Challenges, and Surprising Evaluations. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 259–273.
[51] Ido Ramati. 2024. Algorithmic Ventriloquism: The Contested State of Voice in AI Speech Generators. *Social Media+ Society* 10, 1 (2024), 20563051231224401.
[52] Miguel Ramos, Manuela Barreto, Naomi Ellemers, Miguel Moya, and Lúcia Ferreira. 2018. What hostile and benevolent sexism communicate about men's and women's warmth and competence. *Group Processes & Intergroup Relations* 21, 1 (2018), 159–177.
[53] Ariadna Sanchez, Alice Ross, and Nina Markl. 2024. Beyond The Binary: Limitations and Possibilities of Gender-Related Speech Technology Research. In *2024 IEEE Spoken Language Technology Workshop (SLT)*. 526–532. doi:10.1109/SLT61566.2024.10832234
[54] Simon Schreibelmayr and Martina Mara. 2022. Robot voices in daily life: Vocal human-likeness and application context as determinants of user acceptance. *Frontiers in Psychology* 13 (2022), 787499.
[55] Jessica L Spence, Matthew J Hornsey, Eloise M Stephenson, and Kana Imuta. 2024. Is your accent right for the job? A meta-analysis on accent bias in hiring decisions. *Personality and Social Psychology Bulletin* 50, 3 (2024), 371–386.
[56] Éva Székely, Jura Miniota, and Míša Michaela Hejná. 2025. Will AI shape the way we speak? The emerging sociolinguistic influence of synthetic voices. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*. 335–340.
[57] Erik R Thomas and Jeffrey Reaser. 2004. Delimiting perceptual cues used for the ethnic labeling of African American and European American voices. *Journal of sociolinguistics* 8, 1 (2004), 54–87.
[58] Melanie Weirich and Adrian P Simpson. 2018. Gender identity is indexed and perceived in speech. *PLoS One* 13, 12 (2018), e0209226.
[59] Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of Large Language Model's influence on human spoken communication. arXiv:2409.01754 [cs.CY] https://arxiv.org/abs/2409.01754
[60] Georgia Zellou and Nicole Holliday. 2024. Linguistic analysis of human-computer interaction. *Frontiers in Computer Science* 6 (2024), 1384252.