Computers in Human Behavior Reports 19 (2025) 100762

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Evaluating trustworthiness across ethnically diverse human and commercial synthesised voices: A comparative study

Constantina Maltezou-Papastylianou [a,b,*], Reinhold Scherer [b], Silke Paulmann [a]

[a] Department of Psychology and Centre for Brain Science, University of Essex, Colchester, CO4 3SQ, United Kingdom
[b] Brain-Computer Interfaces and Neural Engineering Laboratory, School of Computer Science and Electronic Engineering, University of Essex, Colchester, CO4 3SQ, United Kingdom

ARTICLE INFO

Keywords:
Trust
Speech acoustics
Ethnic minority
Voice assistants
Synthesised speech
Intelligent agents
Accents

ABSTRACT

This study examined trustworthiness perceptions in the tone of voice of human and real-world synthesised voices, focusing on the impact of acoustic features, speaker and listener ethnicities, listener biases toward voice-based intelligent agents and speaker nature (human vs synthesised). Speech rate, mean pitch, harmonics-to-noise ratio, jitter, shimmer, cepstral peak prominence, and long-term average spectrum, significantly influenced trustworthiness ratings across both human and synthesised voices. Synthesised voices were rated as sounding more trustworthy than human voices with no explicit intent behind their tone of voice (i.e., neutral). However, synthesised voices were rated as sounding less trustworthy than human voices when human speakers intentionally attempted to sound trustworthy. Moreover, listener biases were measured using the Negative Attitudes toward Robots Scale (NARS), where a general scepticism toward robots lowered trustworthiness ratings overall. White speakers were consistently rated as more trustworthy than Black or south Asian speakers across all listener ethnic groups. The findings highlight the need to optimise acoustic properties of synthesised voices for trustworthiness while addressing biases related to speaker ethnicity and listener attitudes toward robots.

1. Introduction

Voice is central to social communication, enabling listeners to make rapid judgements about others, including whether they seem trustworthy (Kreiman & Sidtis, 2011; Maltezou-Papastylianou et al., 2025). While often used interchangeably, trust and trustworthiness are conceptually distinct (Castelfranchi & Falcone, 2010; Hardin, 2002). Trust refers to the willingness to rely on another, based on the expectation that they will not act against one's interests. This decision is context-dependent and shaped by risk or uncertainty. In contrast, perceived trustworthiness refers to the traits we attribute to others – such as honesty, warmth and competence – that influence whether we choose to trust them (Hardin, 2002; Oleszkiewicz et al., 2017; Syed et al., 2024; Tanis & Postmes, 2005). In other words, trustworthiness is the foundation upon which trust is built: if someone is deemed trustworthy, individuals are more inclined to trust them.

Beyond human interactions, individuals also instinctively attribute social traits – including trustworthiness – to pets and artificially intelligent agents (IAs), like voice assistants (Kepuska & Bohouta, 2018, pp. 99–103), humanoid robots (Kouravanas & Pavlopoulos, 2022; Radford et al., 2015; Shigemi et al., 2018), and virtual agents (Yuan et al., 2019). This tendency to anthropomorphise technology emphasises the potential of human-agent interactions (HAI) to simulate aspects of human-human communication (Nass & Brave, 2005; Seaborn et al., 2021). This tendency is highlighted in two key frameworks. The uncanny valley theory (Mori, 1970; Mori et al., 2012) warns that when IAs appear or sound almost – but not quite – human, subtle mismatches (e.g., unnatural pitch, timing, or facial expressions) can evoke a sense of uneasiness (Muralidharan et al., 2014). Meanwhile, the Computers as Social Actors (CASA) theory (Nass et al., 1994, pp. 72–78) suggests that humans attribute social characteristics to machines based on cues like voice, forming impressions similar to those made in human interactions (Asif, 2024; Aylett et al., 2017; Large et al., 2019; Nass & Brave, 2005). Since these mechanisms are deeply rooted in human social cognition, a meaningful exploration of trust in IAs must begin with how trust is formed in human-to-human interactions, where biases and expectations originate.

Building on this premise, the present study investigates how social biases related to speaker ethnicity, listener attitudes toward robots and vocal characteristics interact to shape trustworthiness perceptions

* Corresponding author. Department of Psychology and Centre for Brain Science, University of Essex, Colchester, CO4 3SQ, United Kingdom.
E-mail addresses: cm19066@essex.ac.uk (C. Maltezou-Papastylianou), r.scherer@essex.ac.uk (R. Scherer), paulmann@essex.ac.uk (S. Paulmann).

https://doi.org/10.1016/j.chbr.2025.100762
Received 2 June 2025; Received in revised form 15 July 2025; Accepted 25 July 2025
Available online 25 July 2025

across both human and synthesised (i.e., artificial) voices. The following sections review the relevant literature that informs this approach.

1.1. Human behaviour and biases

In human-human interactions, group affiliations such as ethnicity or profession, and broader societal norms, can further shape trustworthiness judgments and trust attitudes (Greenwald & Banaji, 1995; Schild et al., 2022; Tanis & Postmes, 2005). For instance, Geiger et al. (2023) found that in a U.S. job-hiring simulation, native English-speaking candidates were rated as more trustworthy than those with Mandarin Chinese accents – regardless of rater background – particularly in terms of perceived job-related abilities. These findings may reflect a general tendency to associate native English speakers with positive traits such as credibility and competence, particularly in the U.S., where English dominates professional and academic settings (Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017; Torre et al., 2024). Alternatively, they may reflect a similarity-attraction bias, whereby participants favour speakers who seem linguistically or culturally similar to themselves (Dahlbäck et al., 2007, pp. 1553–1556; Montoya & Horton, 2013). In a predominantly English-speaking American sample, native speakers may have been perceived as more culturally aligned with listeners, leading to more favourable evaluations.

Interestingly, contradictory findings challenge this pattern. A study conducted in Singapore revealed that Mainland Chinese speakers were trusted more by Singaporean Chinese listeners, exhibiting out-group favouritism – where listeners favour an ethnic group they are not affiliated with (Batsaikhan et al., 2021). These results were attributed to participants' cultural familiarity with traditional Chinese norms, such as the expectation that "a favour given must be returned" (Batsaikhan et al., 2021). The authors proposed that in the context of trust-related tasks, Mainland Chinese speakers were perceived as more aligned with reciprocity norms, which are highly valued in such interactions. Together, these studies suggest that societal norms and personal biases jointly shape how vocal trustworthiness is perceived. While such biases are evident in human-human interactions, they also manifest in HAI, particularly in trustworthiness evaluations of human versus synthesised voices.

1.2. Individual differences and biases toward IAs

In HAI, individuals have shown a preference toward IAs that reflect their own ethnicity or accent, often perceiving them as more personable, credible, and engaging (Bilal & Barfield, 2021; Liao & He, 2020, pp. 430–442; Schild et al., 2022; Yang et al., 2025). The similarity-attraction effect remains relevant, especially when evaluating out-group or unfamiliar speakers (Aylett et al., 2017; Dahlbäck et al., 2007, pp. 1553–1556; Zhang et al., 2025). Familiarity can mitigate such biases, but the artificial nature of voice-based IAs may reinforce perceptions of dissimilarity and reduce trust (Lima et al., 2019, pp. 533–538; Tanis & Postmes, 2005). Thereupon, one could raise the question of whether synthesised voices may be perceived as less trustworthy due to their association with non-human entities.

Moreover, listeners' predispositions (i.e., overall inclination to trust others) such as trust propensity toward IAs can further affect evaluations of trustworthiness (Nomura, Suzuki, Kanda, & Kato, 2006a, 2006b; Torre et al., 2024). Questionnaires like the Negative Attitudes to Robots Scale (NARS) (Nomura, Suzuki, Kanda, & Kato, 2006; Nomura et al., 2006a, 2006b) reveal how individual differences shape perceptions of IAs (Kühne et al., 2020; Lim et al., 2022, pp. 538–545). NARS measures attitudes across three subscales: interaction with robots, the social influence of robots, and emotional engagement with robots. Studies in Japanese samples found that NARS scores negatively correlated with measures of social acceptance of robots (Nomura et al., 2006a, 2006b). Similarly, other studies observed that listeners with higher NARS scores rated virtual robots with synthesised voices and physical robots lower on trust, reflecting a bias against robots (Krantz et al., 2022; Lim et al., 2022, pp. 538–545). However, it was also observed that NARS may reflect broader trust tendencies and predispositions in a robot rather than specific capabilities of the robot (Krantz et al., 2022), and older people may exhibit more negative attitudes toward technology than their younger counterparts (Matthews et al., 2019).

While these studies highlight preferences for- and acceptance of-robotic partners, few focus specifically on voice-based IAs, and even less so on ethnically diverse voice-based IAs. This opens an opportunity to examine how listeners' predispositions toward IAs shape trustworthiness evaluations of human versus synthesised voices, and whether vocal cues alone can offset biases linked to a voice's non-human origin.

1.3. Acoustic and contextual influences on trustworthiness perceptions

Past research has shown that listeners infer trustworthiness judgements from vocal cues such as pitch, intonation and speech rate (Belin et al., 2019; Ko et al., 2020, pp. 174–193; Lim et al., 2022, pp. 538–545). For example, in public communication and emergency scenarios, faster speech rates, higher pitch or varied intonation have been perceived as credible and engaging, leading to increased trustworthiness ratings in human and synthesised voices alike (Chan & Liberman, 2021; Kim et al., 2023, pp. 343–347; Rodero et al., 2014; Smith & Shaffer, 1995; Yokoyama & Daibo, 2012). Conversely, slower speech rates and lower pitch seem to be favoured in healthcare settings for their empathetic and calming tone (Maxim et al., 2023, pp. 1–8). Deliberate voice modulation with the intent of sounding trustworthy – such as using variable intonation patterns or sounding emotionally positive – has been suggested to further enhance perceptions of trustworthiness, rapport and learning (Belin et al., 2019; Cambre & Kulkarni, 2019; Torre et al., 2020; Zhang et al., 2025).

Voice quality features like harmonics-to-noise ratio (HNR), which can reflect a speaker's age and health condition, can be indicative of youthfulness and vocal smoothness with higher values, and aging with lower values (Ferrand, 2002). Some studies suggest older-sounding voices may be trusted more in certain contexts, due to perceived experience or wisdom (McAleer et al., 2014; Montepare et al., 2014). Higher-pitched voices are argued to increase perceptions of trustworthiness potentially due to increased association with a sense of friendliness and approachability (Ohala, 1983, 1995, pp. 325–347). Analogously, a halo effect (i.e., a person's overall positive impression influencing judgments about specific traits) extends to perceived trustworthiness of machines (Gabrieli et al., 2021; Huang et al., 2024); researchers found that displaying images of trustworthy-looking human faces on automated teller machines (ATMs) increased the perceived trustworthiness of the ATMs compared to those with less trustworthy-looking faces (Gabrieli et al., 2021). Overall, these findings highlight the multifaceted nature of trustworthiness perceptions, shaped by both vocal features and situational demands (Bachorowski & Owren, 1995).

Building on past work, this study explores how voice quality and acoustic features interact with speaker nature, ethnicity and intent in shaping trustworthiness perceptions. Prior work (Maltezou-Papastylianou et al., 2023, 2025) has begun to address the role of ethnicity in voice evaluation, particularly in human speech; however, less is known about how these features unfold in synthesised voices and cross-ethnic speaker–listener pairings.

1.4. Research motivation and aims

Given the centrality of trust to societal well-being and technology acceptance, it is crucial to examine how voice-based IAs are perceived across diverse demographics (Ghorayeb et al., 2021; Jessup et al., 2019). With real-world applications of voice-based IAs becoming more ubiquitous and human-like, understanding how voice, ethnicity, and listener bias intersect is essential for building trustworthy, inclusive

technologies (Bilal & Barfield, 2021; Gluszek & Dovidio, 2010; Visser & El Fakiri, 2016).

To address these factors, the present study focuses on three key dimensions: speaker nature (human vs synthesised), speaker-listener ethnicity (White, Black, south Asian), intentional vocal modulation (neutral vs trustworthy) and listener attitudes toward robots, measured using the Negative Attitudes toward Robots Scale (NARS) (Nomura, Suzuki, Kanda, & Kato, 2006; Nomura et al., 2006a, 2006b). Firstly, we hypothesised that listeners with more negative attitudes toward robots (higher NARS scores) would rate synthesised voices as less trustworthy than human voices, regardless of speaker intent or demographics (H1). We further hypothesised that synthesised voices will differ in trustworthiness ratings compared to human voices with a neutral (non-trust-building) intent (H2). This non-directional hypothesis serves as a baseline in our study, to identify fundamental differences in trustworthiness perceptions between human and real-world, commercially available synthesised voices, in the absence of any deliberate trust-enhancing cues. Building on H2, we expected that human voices intentionally modulated to sound trustworthy would be rated as more trustworthy than synthesised voices (H3), reflecting the effectiveness of deliberate vocal strategies when conveyed by humans.

Beyond these confirmatory analyses, we also conduct an exploratory analysis to investigate how specific acoustic features – fundamental frequency (f_0), speech rate, HNR, jitter, shimmer, CPP and LTAS – relate to trustworthiness ratings. This analysis seeks to identify consistent acoustic patterns across speaker nature that may serve as perceptual cues of trustworthiness and offer practical guidance for future synthesised voice design. By integrating both confirmatory and exploratory approaches, this study aims to offer a comprehensive perspective on the relationship between social biases and vocal attributes in trust-related judgements – contributing evidence for more inclusive and psychologically grounded voice-based IAs.

This study has been pre-registered on the Open Science Framework (OSF) platform (https://osf.io/v7fam). Although speaker-listener sex were initially intended as variables alongside ethnicity, these were excluded to reduce analytical complexity and sharpen our study's focus. By narrowing the scope, we aimed to ensure clearer and better motivated hypotheses. The role of speaker-listener sex can be explored separately in future work.

2. Methods

2.1. Ethics declaration

All procedures performed in this study were approved by the Ethics Subcommittee 2 of the University of Essex (ETH2324-1869) and were carried out in accordance with the Declaration of Helsinki. All participants provided informed consent prior to participation, where they were also briefed that their anonymised data could be (1) shared in publicly accessible archives and (2) used in future research studies.

2.2. Stimuli

12 speakers from three ethnicities (White, Black and south Asian) spoke three sentences each ("Hi, the shops are still open."; "You may bring a friend with you."; "I will direct you on this."). The sentences were constructed to minimise bias towards any particular emotional interpretation, and they were standardised in length, each consisting of seven syllables. Six speakers were human (recruited in the UK; White female = 36 years old; White male = 25 years old; Black female = 26 years old; Black male = 36 years old; south Asian female = 22 years old; south Asian male = 31 years old) and six were IAs, balanced between ethnicities and sex.

Speakers were recorded in a quiet room, from the comfort of their own home. They used their personal computers and microphones to access a project-specific, online recording website, which records files in wav format. Human speakers were asked to speak the materials once with no specific social intent (i.e. neutral – using their natural tone of voice) and a second time while aiming to sound trustworthy. To mitigate experimenter bias, no examples were provided on how they should sound. A researcher was present during each recording to answer any queries, observe whether the instructions had been followed appropriately and assess the quality of the recordings to mark completion.

All audio recordings were processed using Audacity (version 2.3.3), where they were standardised to a mono channel with a sampling rate of 48.0 kHz, 16-bit depth, and a bit rate of 768 kb/s, and then normalised to 67 dB using Praat software (version 6.2.16) (Boersma, 2001). The final files were saved in an uncompressed wav format. For more information on the human stimuli and recording procedure see (Maltezou-Papastylianou et al., 2024, 2025).

The IA voices were generated using Narakeet text-to-speech web tool (https://www.narakeet.com/) from their pre-existing list of accented voices, and exported the audio files in wav format (British White male – Edward; British White female – Helen; Nigerian accent, Black male – Obinna; Nigerian accent, Black female – Thandiwe; Indian accent, south Asian male – Dilip; Indian accent, south Asian female – Pooja); Narakeet's default configuration was used (standard volume, normal speed, single audio file), and no particular intent was specified. Narakeet was selected for this study because, at the time of stimulus creation, it was the only text-to-speech tool the authors could identify that offered English-language voices representing speakers from all three ethnic groups and both sexes.

Acoustic and spectral features were extracted using the VoiceLab software (Feinberg & Cook, 2020; Feinberg D. , 2022). The features included mean f_0, standard deviation of f_0 to assess pitch variability, and voice duration as an indicator of speech rate. Several voice quality features were also extracted – HNR, jitter, shimmer, CPP, mean LTAS, LTAS standard deviation, and LTAS slope, which as noted in Table 1, they are commonly associated with signal clarity and noise levels, and are often linked to perceptions of vocal breathiness, roughness, or hoarseness. Following past research (Baus et al., 2019; McAleer et al., 2014), jitter was measured using the relative average perturbation (RAP) approach, which measures how much the duration of each vocal cycle (i.e., glottal periods) varies compared to the average length of its neighbouring cycles. Shimmer was measured using the amplitude perturbation quotient 3 (APQ3) method, which assesses how much the loudness of each vocal cycle fluctuates relative to surrounding cycles. f_0 was measured using VoiceLab's autocorrelation approach. For further description of each acoustic feature see Table 1. Summary descriptives of each feature per demographic group can be found in Tables 2 and 3 for human voices, and Table 4 for synthesised voices.

3. Participants/listeners

180 English-speaking adults (60 participants x 3 ethnicities) from the UK were recruited through Prolific (Prolific, 2014) to rate the audio stimuli. See Table 5 for more details on listener demographics. We followed guidance from past research that has indicated that a sample size of at least 28 participants per condition for trustworthiness research tends to yield a high Cronbach's alpha (McAleer et al., 2014). Throughout this paper, the terms participants/listeners may be used interchangeably.

3.1. Rating procedure

During the study, which took place online on a PHP-based, project-specific website, participants were required to firstly answer the 14-item NARS questionnaire, which is concerned with three themes classified under three subscales: negative attitudes toward situations of interaction with robots (S1), negative attitudes toward the social influence of robots (S2) and negative attitudes toward emotions in interaction with robots (S3) (Nomura et al., 2006a, 2006b). Higher score on the NARS or

Table 1
Summary characteristics of speech acoustics examined.

Acoustic signal	Unit of measurement	Key characteristics
Fundamental frequency (f_0)	Hertz (Hz)	f_0 refers to the base rate at which the vocal folds vibrate, and it is perceived as pitch. Variations in $f0$ across an utterance reflect changes in vocal intonation.
Amplitude	Decibels (dB)	Amplitude refers to the magnitude of air pressure variations in a sound wave and is perceived by listeners as loudness. Fluctuations in amplitude across an utterance contribute to the perceived intensity and emphasis of speech.
Harmonics-to-noise ratio (HNR)	dB	Noise in a voice signal refers to any factor that interferes with the clarity or quality of speech and is typically independent of the voice's fundamental frequency. It can arise from various sources such as changes in vocal fold function, muscle strain, breathing patterns, background noise, or technical distortions (Ferrand, 2002). A lower HNR reflects a higher level of noise present in the signal (Fernandes et al., 2018; Ferrand, 2002).
Jitter	%	Jitter measures tiny, rapid variations in pitch that occur due to uneven vocal fold vibrations. A lower jitter percentage reflects more stable pitch and smoother vocal production (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).
Shimmer	dB	Shimmer reflects small, rapid changes in amplitude, providing an indication of variability in the loudness or intensity of the voice. Jitter and shimmer are often examined together because they both reflect micro-instabilities in the voice (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).
Cepstral peak prominence (CPP)	dB	CPP measures the amplitude difference between the cepstral peak (harmonic structure) and the background noise in the cepstrum. A lower CPP indicates a breathy or dysphonic voice, while higher CPP values, are indicative of clearer, more resonant voices with stronger harmonic structure (Chan & Liberman, 2021; Hammarberg et al., 1980; Jalali-najafabadi et al., 2021).
Long-term average spectrum (LTAS)	dB	LTAS measures the average energy distribution of a sound signal across different frequencies over time, providing a spectral profile of a sound. A lower LTAS often reflects longer vocal tracts (Da Silva et al., 2011; Linville, 2002; Löfqvist, 1986), and it's associated with deeper, more resonant voices linked to dominance, particularly in males (Gussenhoven, 2002; Puts et al., 2007).

its subscales suggests a less favourable evaluation of the interaction. Subsequently, using their own computers in a quiet room, each participant listened to all speakers, where the audio stimuli were randomised using the Fisher-Yates Shuffle algorithm (Eberl, 2016). After each audio recording, they were asked to respond to the statement "This speaker sounds trustworthy" on a Likert scale ranging from 1 (strongly disagree) – 7 (strongly agree).

4. Results

All statistical analyses were conducted using JASP (version 0.19) for ANOVA and post-hoc tests, and Python (version 3.11.0) for linear mixed-effects models (via the statsmodels and pingouin libraries). In terms of reporting the ANOVA results, omega-squared (ω^2) was used as an indicator of effect sizes. Even though effect sizes are context-dependent, an $\omega^2 = 0.01$ (i.e. 1 % of variance explained) is typically considered a small effect in the literature, an $\omega^2 = 0.06$ a medium effect, and $\omega^2 = 0.14$ a large effect (Field, 2018; Kirk, 1996). When sphericity assumptions were violated, p-values for within-subjects comparisons were adjusted using the Greenhouse-Geisser correction. For significant main effects and interactions in ANOVAs, post-hoc comparisons were performed using Holm–Bonferroni corrected t-tests to control for Type I error inflation due to multiple comparisons (Abdi, 2010). For more information on the mean trustworthiness ratings per listener-speaker ethnicity, and speaker intent and nature, see Fig. 1.

4.1. Exploring acoustic features in classifying trustworthy human and synthesised voices

The exploratory analysis sought to investigate the role of acoustic features in terms of classifying trustworthy human and synthesised voices. Specifically, a mixed-effects model was used to determine which acoustic features are common across the two speaker natures (i.e. IA vs human speakers) in terms of listeners' perceived trustworthiness. The acoustic features acted as the fixed effects, trustworthiness ratings as the target and listeners as the random effect. Results revealed that while voice duration (measuring speech rate here), HNR, jitter, shimmer and CPP had a significant negative relationship with trustworthiness ratings, mean f_0 and mean LTAS exhibited a significant positive relationship with trustworthiness ratings. See Table 6 for further details.

H1. Higher NARS scores predict lower trust ratings for synthesised voices than human voices, regardless of intent or demographics.

A mixed-effects model was employed to examine H1 as to how NARS scores on each NARS subscale (S1, S2 and S3) have influenced trustworthiness ratings (dependent variable) of synthesised voices (i.e., IAs) compared to human voices. The model included fixed effects for NARS scores, speaker nature (human vs IA), speaker intent and ethnicity, and participant ethnicity, while accounting for interrater reliability with random intercepts by participant ID (grouping variable).

Results revealed that trustworthiness ratings were higher for speakers with a trustworthy intent and for speakers of White ethnicity compared to other groups. Conversely, participant ethnicity (south Asian and White) was associated with lower trustworthiness ratings. Significant interaction effects between speaker nature (IA) and NARS S3 scores suggest that participants' attitudes toward robots influenced their trustworthiness ratings of synthesised voices differently compared to human voices. The grouping variable, $\sigma^2 = 0.24$, reflects the amount of inter-individual variability in trustworthiness ratings attributable to differences between participants' baseline trustworthiness ratings (i.e., some participants consistently gave higher or lower ratings than others), independent of the fixed effects. Full results are presented in Table 7 and Fig. 1.

H2. Trustworthiness ratings differ between synthesised and human voices, influenced by speaker ethnicity.

To answer H2, data relating to human voices with trustworthy intent were excluded from the analysis. The goal was to ascertain whether synthesised voices would significantly differ in trustworthiness ratings compared to human voices that are not intentionally expressed to sound trustworthy (i.e. neutral), and that these ratings would be influenced by the speaker's ethnicity and sex. A 2 (Speaker Nature: Synthesised, Human) x 3 (Speaker Ethnicity: White, Black, south Asian) x 3 (Listener Ethnicity: White, Black, south Asian) mixed ANOVA was employed.

The main effect of speaker nature was significant, $F(1, 177) = 158.07$, $p < .001$, $\omega^2 = 0.27$, showing higher trustworthiness ratings for synthesised voices compared to human voices. The main effect of

Table 2
Human speakers with trustworthy intent: Descriptive statistics of acoustic features per demographic.

Acoustic features	Mean acoustic values [Standard deviation]					
	White		Black		South Asian	
	Male	Female	Male	Female	Male	Female
Duration (s)	**1.35** [0.20]	**1.64** [0.34]	**1.46** [0.22]	**1.93** [0.26]	**1.49** [0.25]	**1.43** [0.14]
f_0, mean pitch (Hz)	**153.46** [41.17]	**240.53** [16.30]	**171.29** [7.69]	**191.01** [21.12]	**115.99** [6.48]	**226.23** [13.86]
f_0, SD pitch (Hz)	**63.32** [39.63]	**73.22** [1.89]	**52.39** [20.11]	**33.68** [10.97]	**21.29** [8.37]	**44.29** [7.59]
HNR (dB)	**1.74** [1.27]	**8.9** [1.70]	**8.71** [0.85]	**9.12** [1.32]	**4.31** [1.62]	**11.98** [2.84]
Jitter (RAP)	**0.02** [0.005]	**0.01** [0.003]	**0.01** [0.001]	**0.01** [0.001]	**0.01** [0.004]	**0.02** [0.01]
Shimmer (APQ3)	**0.06** [0.01]	**0.02** [0.01]	**0.04** [0.01]	**0.02** [0.002]	**0.06** [0.02]	**0.03** [0.01]
CPP (dB)	**26.9** [2.26]	**30.31** [0.18]	**27** [2.81]	**26.16** [1.93]	**24.09** [2.61]	**28.68** [3.11]
LTAS, mean (dB)	**−5.83** [3.98]	**4.84** [3.67]	**−23.21** [1.52]	**5.82** [2.44]	**−0.97** [3.59]	**2.97** [3.87]
LTAS, SD (dB)	**18.78** [0.87]	**16.3** [0.60]	**21.7** [0.78]	**16.55** [1.05]	**17** [0.47]	**18.09** [0.92]
LTAS slope, (dB/octave)	**−6.27** [2.55]	**−8.66** [1.70]	**−15.73** [0.79]	**−10.49** [1.71]	**−12.67** [1.92]	**−14.41** [1.04]

Table 3
Human speakers with neutral intent: Descriptive statistics of acoustic features per demographic.

Acoustic features	Mean acoustic values [Standard deviation]					
	White		Black		South Asian	
	Male	Female	Male	Female	Male	Female
Duration (s)	**1.47** [0.17]	**1.44** [0.07]	**2.68** [0.35]	**1.98** [0.42]	**1.59** [0.01]	**1.63** [0.22]
f_0, mean pitch (Hz)	**98.53** [5.65]	195.51 [2.77]	**152.46** [1.26]	**175.38** [4.90]	**102.89** [2.37]	**156.39** [22.63]
f_0, SD pitch (Hz)	**35.63** [19.63]	58.48 [11.14]	**26.62** [3.11]	**18** [1.20]	**8.22** [1.93]	**38.18** [20.01]
HNR (dB)	**2.25** [0.54]	9.99 [1.86]	**8.47** [2.00]	**12.02** [1.94]	**5.08** [2.50]	**9.94** [2.59]
Jitter (RAP)	**0.01** [0.001]	**0.01** [0.00]	**0.01** [0.002]	**0.01** [0.00]	**0.01** [0.001]	**0.01** [0.005]
Shimmer (APQ3)	**0.05** [0.01]	**0.03** [0.001]	**0.03** [0.002]	**0.02** [0.004]	**0.05** [0.002]	**0.04** [0.01]
CPP (dB)	**26.16** [0.99]	**27.55** [1.89]	**25.55** [0.62]	**28.62** [2.25]	**26.09** [3.08]	**26.09** [0.92]
LTAS, mean (dB)	**−6.84** [4.80]	**3.65** [3.22]	**−21.79** [0.63]	**1.61** [4.16]	**−0.66** [2.40]	**4.03** [3.26]
LTAS, SD (dB)	**18.69** [1.12]	**16.6** [0.47]	**21.51** [1.06]	**16.64** [1.98]	**15.83** [1.12]	**18.56** [1.21]
LTAS slope, (dB/octave)	**−8.72** [0.76]	**−7.87** [2.70]	**−14.68** [0.79]	**−13.42** [2.27]	**−14.31** [2.36]	**−14.99** [2.88]

Table 4
IA speakers: Descriptive statistics of acoustic features per demographic.

Acoustic features	Mean acoustic values [Standard deviation]					
	White		Black		South Asian	
	Male	Female	Male	Female	Male	Female
Duration (s)	**1.41** [0.17]	**1.46** [0.20]	**1.49** [0.26]	**1.58** [0.33]	**1.5** [0.14]	**1.62** [0.23]
f_0, mean pitch (Hz)	**119.61** [13.69]	**180.75** [2.74]	**115.25** [7.85]	**189** [5.92]	**160.28** [6.38]	**238.95** [4.25]
f_0, SD pitch (Hz)	**37.41** [8.22]	**44.1** [5.45]	**20.04** [4.16]	**24.51** [7.40]	**39.3** [5.09]	**43.05** [1.48]
HNR (dB)	**3.95** [2.11]	**10.72** [1.76]	**4.23** [1.92]	**9.86** [1.75]	**6.18** [1.94]	**15.06** [1.97]
Jitter (RAP)	**0.01** [0.00]	**0.01** [0.002]	**0.01** [0.002]	**0.01** [0.002]	**0.01** [0.002]	**0.01** [0.00]
Shimmer (APQ3)	**0.03** [0.003]	**0.02** [0.01]	**0.03** [0.01]	**0.03** [0.01]	**0.02** [0.002]	**0.02** [0.002]
CPP (dB)	**23.9** [1.21]	**24.45** [0.94]	**25.72** [1.46]	**25.4** [1.47]	**22.37** [2.58]	**26.27** [1.70]
LTAS, mean (dB)	**0.25** [2.65]	**−18.05** [1.98]	**1.79** [2.21]	**3.6** [3.65]	**−16.63** [1.62]	**−17.04** [3.76]
LTAS, SD (dB)	**15.28** [0.91]	**26.92** [1.39]	**15.63** [1.22]	**15.24** [1.30]	**27.38** [1.33]	**25.13** [2.57]
LTAS slope, (dB/octave)	**−16.38** [1.28]	**−11.4** [1.09]	**−14.13** [1.67]	**−12.73** [1.65]	**−14.82** [0.91]	**−18.05** [1.66]

Table 5
Descriptive statistics of participant demographics.

Ethnicity	Sex	N	Mean age (years)	Age range	SD
White	Female	30	34.57	19–45	7.41
	Male	30	31.20	18–43	6.84
Black	Female	30	27.77	18–42	6.38
	Male	30	30.40	20–44	6.02
South Asian	Female	30	26.20	18–42	6.74
	Male	30	28.83	19–43	7.47

speaker ethnicity was also significant, $F(1.94, 342.50) = 25.89$, $p < .001$, $\omega^2 = 0.05$, and so was listener ethnicity, $F(2, 177) = 11.003$, $p < .001$, $\omega^2 = 0.04$. Post-hoc comparisons for speaker ethnicity showed higher trustworthiness ratings for White speakers over Black ($M_{diff} = 0.37$, $SE = 0.06$, $p < .001$) and south Asian ($M_{diff} = 0.27$, $SE = 0.06$, $p < .001$). Trustworthiness ratings for south Asian speakers were also significantly higher than for Black speakers ($M_{diff} = 0.11$, $SE = 0.05$, $p = .03$). Post-hoc comparisons for listener ethnicity showed higher trustworthiness ratings from Black listeners over White ($M_{diff} = 0.46$, $SE = 0.11$, $p < .001$) and south Asian ($M_{diff} = 0.39$, $SE = 0.11$, $p < .001$), but no significant difference between White and south Asian listeners ($p = .54$).

Speaker nature x speaker ethnicity was the only significant interaction, $F(1.68, 296.59) = 31.85$, $p < .001$, $\omega^2 = 0.05$. See Tables 8 and 9 for all results of the ANOVA. Post-hoc comparisons showed that White human speakers were rated as significantly more trustworthy than both Black ($M_{diff} = 0.77$, $SE = 0.06$, $p < .001$) and south Asian ($M_{diff} = 0.53$, $SE = 0.06$, $p < .001$) human speakers, but significantly less trustworthy than IA speakers across all ethnicities ($p < .001$). Black human speakers were rated lower than south Asian human speakers ($M_{diff} = -0.24$, $SE = 0.07$, $p = .002$), and Black ($M_{diff} = -1.27$, $SE = 0.09$, $p < .001$) and south

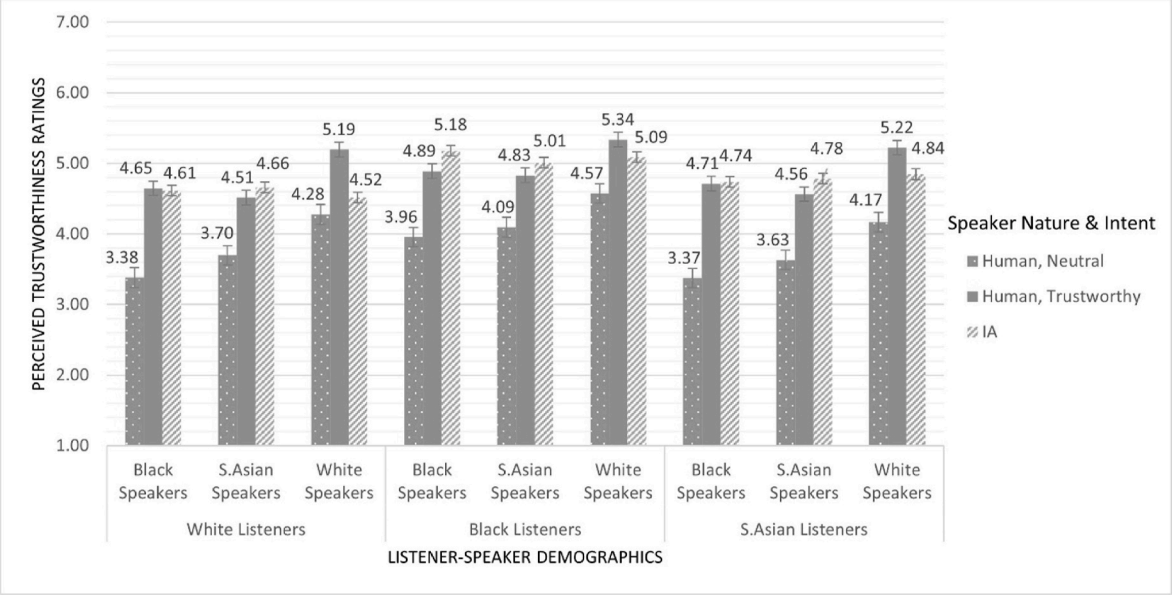

Fig. 1. Listeners' mean trustworthiness ratings (1-strongly disagree to 7-strongly agree) per speaker nature, intent and demographic group.

Table 6
Exploratory mixed-effects model results summary table.

	CI 95 %					
	Coefficient	Std.Err.	z	*p*-value	[0.025	0.975]
Intercept	**6.23**	**0.23**	**27.62**	**0.00**	**5.79**	**6.67**
Voice duration	**−0.59**	**0.04**	**−15.60**	**0.00**	**−0.66**	**−0.52**
f_0, mean	**0.02**	**0.001**	**20.88**	**0.00**	**0.02**	**0.02**
f_0, SD	0.001	0.001	−1.05	0.30	0.003	0.001
HNR	**−0.15**	**0.01**	**−21.00**	**0.00**	**−0.16**	**−0.14**
Jitter, RAP	**−17.49**	**3.75**	**−4.67**	**0.00**	**−24.83**	**−10.15**
Shimmer, apq3	**−8.16**	**1.43**	**−5.70**	**0.00**	**−10.97**	**−5.35**
CPP	**−0.06**	**0.01**	**−9.38**	**0.00**	**−0.07**	**−0.05**
LTAS, mean	**0.01**	**0.002**	**3.13**	**0.002**	**0.003**	**0.01**
LTAS, SD	−0.01	0.01	−1.91	0.056	−0.02	0.00
LTAS, slope	−0.01	0.01	−0.99	0.33	−0.02	0.01
Grouping variable	0.28	0.03				

Table 7
Mixed-effects model results summary table.

	CI 95 %					
	Coefficient	Std.Err.	z	*p*-value	[0.025	0.975]
Intercept	**3.69**	**0.36**	**10.15**	**0.00**	**2.98**	**4.40**
Speaker nature [IA]	0.30	0.27	1.10	0.27	−0.23	0.82
Speaker intent [Trustworthy]	**0.97**	**0.03**	**28.31**	**0.00**	**0.91**	**1.04**
Speaker ethnicity [South Asian]	0.03	0.03	0.94	0.35	−0.04	0.10
Speaker ethnicity [White]	**0.42**	**0.03**	**12.06**	**0.00**	**0.35**	**0.48**
Participant ethnicity [South Asian]	**−0.31**	**0.10**	**−3.14**	**0.002**	**−0.50**	**−0.12**
Participant ethnicity [White]	**−0.34**	**0.10**	**−3.51**	**0.00**	**−0.53**	**−0.15**
NARS S1 total score	**−0.03**	**0.01**	**−2.19**	**0.03**	**−0.05**	**−0.003**
Speaker nature [IA] x NARS S1 total score	0.004	0.008	0.44	0.66	−0.01	0.02
NARS S2 total score	0.02	0.01	1.59	0.11	−0.01	0.05
Speaker nature [IA] x NARS S2 total score	0.02	0.01	1.70	0.09	−0.003	0.04
NARS S3 total score	0.03	0.02	1.79	0.07	−0.003	0.07
Speaker nature [IA] x NARS S3 total score	**0.03**	**0.01**	**1.97**	**0.049**	**0**	**0.06**
Grouping variable	0.24	0.02				

Asian (M_{diff} = *-1.25, SE = 0.10, p < .001*) IA speakers. South Asian human speakers were rated lower than south Asian IA speakers (M_{diff} = *-1.01, SE = 0.09, p < .001*).

White IA speakers were rated as significantly more trustworthy than Black ($M_{diff} = 1.25$, *SE = 0.10, p < .001*) and south Asian ($M_{diff} = 1.01$, *SE = 0.10, p < .001*) human speakers but no significance found with Black and south Asian IA speakers (*p = 1.00*). Black IA speakers were perceived as more trustworthy than south Asian, human speakers (M_{diff} *= 1.04, SE = 0.08, p < .001*), albeit no significance found with south Asian IA speakers (*p = 1.00*).

To summarise, H2 results revealed that synthesised voices were rated significantly higher on perceived trustworthiness than human voices

Table 8
Repeated measures ANOVA results of H2 for within subjects effects.

	Sum of squares	df	Mean square	F	*p*-value	ω^2
Speaker nature	228.41	1	228.41	158.07	<0.001	0.27
Speaker nature x Participant ethnicity	3.09	2	1.54	1.07	0.35	0
Residuals	255.77	177	1.45			
Speaker ethnicity	26.43	1.94	13.66	25.89	<0.001	0.05
Speaker ethnicity x Participant ethnicity	1.78	3.87	0.46	0.87	0.48	0
Residuals	180.67	342.50	0.53			
Speaker nature x Speaker ethnicity	29.53	1.68	17.62	31.85	<0.001	0.05
Speaker nature x Speaker ethnicity x Participant ethnicity	1.20	3.35	0.36	0.65	0.60	0
Residuals	164.10	296.59	0.55			

Table 9
Repeated measures ANOVA results of H2 for between subjects effects.

	Sum of squares	df	Mean square	F	*p*-value	ω^2
Participant ethnicity	44.45	2	22.22	11.003	<0.001	0.04
Residuals	357.50	177	2.02			

with a neutral intent. Trustworthiness ratings were also influenced by speaker and listener ethnicity, with White speakers rated higher than Black and south Asian speakers, and Black listeners providing higher ratings than other groups. A significant interaction showed that synthesised voices were consistently rated more trustworthy than human voices across all ethnicities, with White human speakers rated higher than Black and south Asian human speakers.

H3. Synthesised voices receive lower trust ratings than human voices with trustworthy intent, influenced by speaker ethnicity.

The same factorial ANOVA as in H2 was employed to answer H3, except that this time the data relating to human voices with neutral intent were replaced with those with trustworthy intent. The goal with H3 was to ascertain whether synthesised voices would receive lower trustworthiness ratings compared to human voices with a trustworthy intent, influenced by speaker ethnicity and sex. Thus, a 2 (Speaker Nature: Synthesised, Human) x 3 (Speaker Ethnicity: White, Black, south Asian) x 3 (Listener Ethnicity: White, Black, south Asian) mixed ANOVA was employed.

The main effect of speaker ethnicity was significant, $F(1.92, 338.95) = 19.56, p < .001, \omega^2 = 0.03$, and similarly listener ethnicity, $F(2, 177) = 6.29, p = .002, \omega^2 = 0.02$. Post-hoc comparisons for speaker ethnicity showed higher trustworthiness ratings for White speakers over Black ($M_{diff} = 0.24, SE = 0.05, p < .001$) and south Asian ($M_{diff} = 0.31, SE = 0.06, p < .001$). No significant difference was found between Black and south Asian speakers ($p = .13$). Post-hoc comparisons for listener ethnicity revealed significantly higher trustworthiness ratings from Black listeners than White ($M_{diff} = 0.37, SE = 0.11, p = .002$) and south Asian speakers ($M_{diff} = 0.24, SE = 0.11, p = .04$). No significant difference was found between White and south Asian listeners ($p = .25$).

Only the speaker nature x speaker ethnicity interaction was significant, $F(1.73, 306.65) = 21.93, p < .001, \omega^2 = 0.04$. See Tables 10 and 11 for all results of the ANOVA. Post-hoc comparisons showed that White human speakers were rated as significantly more trustworthy than both Black ($M_{diff} = 0.50, SE = 0.06, p < .001$) and south Asian ($M_{diff} = 0.62, SE = 0.06, p < .001$) human speakers, and significantly more trustworthy than IA speakers too, across all ethnicities ($p < .001$). There were no significant findings when comparing ratings for Black human speakers with ratings in response to south Asian human speakers ($p = .57$), nor with Black and south Asian IA speakers ($p = 1.00$). No significant difference was found between south Asian human speakers and south Asian IA speakers either ($p = .088$).

Table 10
Repeated measures ANOVA results of H3 for within subjects effects.

	Sum of squares	df	Mean square	F	*p*-value	ω^2
Speaker nature	0.79	1	0.79	0.88	0.35	0
Speaker nature x Participant ethnicity	3.12	2	1.56	1.74	0.18	0
Residuals	159.15	177	0.90			
Speaker ethnicity	18.74	1.92	9.78	19.56	<0.001	0.03
Speaker ethnicity x Participant ethnicity	0.63	3.83	0.17	0.33	0.85	0
Residuals	169.54	338.95	0.50			
Speaker nature x Speaker ethnicity	19.86	1.73	11.46	21.93	<0.001	0.04
Speaker nature x Speaker ethnicity x Participant ethnicity	1.67	3.47	0.48	0.92	0.44	0
Residuals	160.28	306.65	0.52			

Table 11
Repeated measures ANOVA results of H3 for between subjects effects.

	Sum of squares	df	Mean square	F	*p*-value	ω^2
Participant ethnicity	24.88	2	12.44	6.29	0.002	0.02
Residuals	350.32	177	1.98			

Trustworthiness ratings did not differ significantly when comparing White IA speakers with Black and south Asian IA and human speakers ($p = 1.00$), nor between Black IA speakers and south Asian IA speakers ($p = 1.00$). However, Black IA speakers were rated as significantly more trustworthy than south Asian human speakers ($M_{diff} = 0.21, SE = 0.07, p = .02$).

To summarise, H3 results revealed significant effects of speaker and listener ethnicity on trustworthiness ratings, with White speakers rated higher than Black and south Asian speakers, and Black listeners providing higher ratings than White and south Asian listeners. A significant interaction showed that White human speakers with trustworthy intent were rated more trustworthy than all other groups, including synthesised voices, while no significant differences were observed among synthesised voices of different ethnicities.

5. Discussion

The present research investigated how listener biases toward robots, speaker-listener ethnicity, and acoustic features influence trustworthiness ratings for human and synthesised voices. The findings provide insights into the perception of voice trustworthiness and highlight the complex interaction of ethnicity, vocal intent and social biases toward robots.

5.1. Acoustic features and trustworthiness

Our exploratory analysis identified key acoustic features that influenced trustworthiness ratings across both human and synthesised voices from White, Black, and south Asian speakers. Specifically, speech rate,

mean fundamental frequency (perceived as pitch), and the voice quality features of HNR, jitter, shimmer, CPP, and LTAS emerged as significant predictors of trustworthiness perceptions.

Faster speech rates were associated with higher trustworthiness ratings. This aligns with research showing that faster speech can convey engagement, credibility and persuasiveness (Rodero et al., 2014; Smith & Shaffer, 1995; Yokoyama & Daibo, 2012). When listeners hear faster-paced delivery, they may interpret it as a sign of effort and eagerness to help or invested in a conversation (Chan & Liberman, 2021; Gussenhoven, 2002; Kim et al., 2023, pp. 343–347). In contexts involving social first impressions, such as ours, these impressions may be well regarded in social settings (Maltezou-Papastylianou et al., 2025), which emphasises the effect of situational context.

Higher mean pitch was also associated with greater trustworthiness, supporting prior work that links higher pitch to emotional warmth and friendliness (Ohala, 1983; Torre et al., 2020). This association highlights the role of pitch in conveying warmth and approachability, traits closely tied to perceived trustworthiness (Belin et al., 2019; Hardin, 2002; McAleer et al., 2014; Ohala, 1995, pp. 325–347; Tanis & Postmes, 2005). The joint effect of faster speech rate with higher pitch may have consequently spilt over into a halo effect which boosted an overall sense of perceived benevolence and warmth in those speakers (Gabrieli et al., 2021; Huang et al., 2024; McAleer et al., 2014).

Conversely, measures of shimmer, jitter, and HNR, which tend to reflect vocal instability and aging, were negatively associated with trustworthiness (Ferrand, 2002; Schweinberger et al., 2014). However, it is worth noting that not all vocal "imperfections" are necessarily undesirable: some irregularity, when paired with warmth might convey vulnerability or emotional sincerity (Bachorowski & Owren, 1995). Future work might examine how these vocal markers are interpreted in different emotional or relational contexts. In contrast, features such as pitch variability, LTAS slope, and LTAS variability did not significantly predict trustworthiness. This may reflect the context-dependence of such cues: lower LTAS values, for instance, have been linked to deeper, more resonant voices associated with dominance and authority (Linville, 2002; Puts et al., 2007). While such traits may enhance perceived competence in knowledge-based or task-oriented interactions, they may be less aligned with social trustworthiness, which often hinges on warmth, empathy, and perceived likability (Maxim et al., 2023, pp. 1–8; Oleszkiewicz et al., 2017). In short, not all acoustic cues are equally salient in all situational contexts – a finding that future research should explore more systematically by varying situational contexts but keeping acoustics constant across contexts.

These results offer valuable guidance for synthesised voice design. While not all vocal parameters need to be optimised simultaneously, our findings suggest that targeting a specific cluster of traits – moderately fast speech, elevated pitch, and reduced vocal irregularities – may be most effective for enhancing perceived trustworthiness in everyday voice-based interactions. Rather than replicating the full complexity of human vocal dynamics, designers of voice-based IAs might focus on prominent acoustic markers that consistently shape positive impressions, adapting these to different usage scenarios (e.g., healthcare vs customer service).

5.2. Listener trust attitudes toward robots and trustworthiness perceptions

The current study partially supported the prediction that individuals with higher negative attitudes toward robots – as measured by the NARS scale – would rate synthesised voices lower than human voices. However, the pattern was not consistent across all subscales, suggesting a more differentiated relationship between listener predispositions and trustworthiness evaluations.

Negative attitudes toward interaction scenarios, as measured by NARS Subscale 1, were associated with lower trustworthiness ratings overall. This suggests that individuals who are generally sceptical about engaging with robots may extend this discomfort to social interactions more broadly within HAI contexts. Rather than responding to specific vocal cues, their judgments may reflect a more global reluctance to engage with artificial agents as social partners. This interpretation is supported by previous findings linking higher NARS scores to reduced trust in robots (Krantz et al., 2022; Lim et al., 2022, pp. 538–545; Nomura et al., 2006a, 2006b). These effects also align with CASA and uncanny valley frameworks, which propose that people automatically apply social characteristics to IAs, and may withdraw trust when the interaction within HAI contexts feels unnatural or dissonant (Matthews et al., 2019; Mori et al., 2012; Nass & Brave, 2005; Nass et al., 1994, pp. 72–78).

Unlike prior studies reporting broader effects of NARS scores (Krantz et al., 2022; Lim et al., 2022, pp. 538–545; Nomura et al., 2006a, 2006b), this study found no significant impact between negative attitudes toward social influence – as measured with NARS subscale 2 – and trustworthiness ratings. While there was a marginal trend indicating that individuals with greater negativity toward robots' societal influence rated synthesised voices more favourably, this result was not robust. This lack of influence suggests that concerns about robots' societal roles—like job displacement or loss of autonomy—may not directly shape how people evaluate trustworthiness in individual voices (Matthews et al., 2019; Seaborn et al., 2021). Such concerns may be more relevant in high-stakes, professional contexts where robots are seen as competitors or decision-makers. In contrast, our study involved socially casual, everyday impressions, where voice-based IAs were likely perceived as familiar, benign tools – especially in domestic settings like those involving Alexa or Google Assistant (Kepuska & Bohouta, 2018, pp. 99–103). Whether these perceptions shift in more consequential scenarios remains an open question for future work.

Surprisingly, NARS Subscale 3 (negative attitudes toward emotional interactions with robots) showed a marginal trend in the opposite direction of our initial prediction: listeners with higher scepticism toward emotional interactions rated synthesised voices as more trustworthy than human voices. A potential interpretation could be that synthesised voices, which lack the emotional unpredictability and richness of human voices, might offer a sense of predictability or impartiality. As such, the current findings and interpretation seem to align with CASA and uncanny valley theories (Mori et al., 2012; Nass & Brave, 2005; Nass et al., 1994, pp. 72–78), alongside findings by Krantz et al. (2022), who argue that NARS may reflect broader psychological orientations, such as discomfort with affective ambiguity, rather than specific robot capabilities.

Collectively, these findings demonstrate that listener biases toward robots do not exert a uniform influence on trustworthiness evaluations. Instead, each NARS subscale captures distinct dimensions of robot-related attitudes, which appear to interact differently with voice-based trust judgments. For instance, while general discomfort with robot interactions (Subscale 1) may suppress trust across the board, attitudes toward robots' emotional or social influence (Subscales 2 and 3) seem more context-sensitive. These findings may also align with the similarity-attraction bias noted in the introduction (Dahlbäck et al., 2007, pp. 1553–1556; Yang et al., 2025; Zhang et al., 2025), suggesting that listeners may gravitate toward voices that align with their own preferences for neutrality or expressiveness.

By focusing on the relationship between listener predispositions and speaker characteristics, these findings deepen our understanding of how synthesised voices can be designed for different user demographics, preferences, and contexts. For practitioners, these findings emphasise the need to create synthesised voices tailored to diverse listener attitudes. For example, features that emphasise emotional neutrality while maintaining warmth and clarity may appeal to users who are sceptical of emotional expressiveness in voice-based IAs. Additionally, addressing general scepticism about robot interactions – more common among older populations and individuals with higher NARS scores (Ghorayeb et al., 2021; Jessup et al., 2019) – could enhance the inclusivity and acceptance of voice-based IAs in trust-dependent applications such as

legal consultations or threat detection applications.

5.3. Real-world synthesised voices outperform human voices with a neutral intent

Interestingly, the real-world synthesised voices we used as our stimuli were rated as more trustworthy than our stimuli of human speakers with a neutral intent. This could be revealing the unique positioning of the real-world synthesised voices used in this study, which may potentially possess acoustic properties engineered to achieve a balance between naturalness and consistency. As discussed in the introduction, deviations from natural human-like speech patterns, such as lower pitch ranges or increased speech time delays, tend to make voices sound more machine-like and less trustworthy (Muralidharan et al., 2014).

Hence, a possible explanation for this preference may lie in the acoustic characteristics of the synthesised voices. Interestingly, the present analysis revealed that the synthesised voices of this study occupy a middle ground between neutral human voices and those intentionally modulated to sound trustworthy for certain acoustic cues (see Tables 2–4). For example, synthesised voices from a White ethnic background had speech rate and mean pitch values between human neutral and human trustworthy intent voices from the same ethnic group, albeit with slightly higher HNR for synthesised voices. We characterise this pattern as occupying a "perceptual middle ground", a term introduced here to describe this balance point in vocal trustworthiness design. Unlike neutral human voices, which may lack distinct acoustic cues that signal trustworthiness, the synthesised voices seem to have been designed with features that balance with listener preferences, fostering positive trustworthiness perceptions. This interpretation builds on the "uncanny valley" phenomenon (Kühne et al., 2020; Mori et al., 2012), suggesting that synthesised voices perceived as clear, natural, and consistent can reduce unease and enhance trustworthiness evaluations. The slower speech rate, higher mean f_0 and range of voice quality features of the synthesised voices may have made them sound less machine-like, avoiding the discomfort and scepticism often associated with artificial agents (Muralidharan et al., 2014; Torre et al., 2018; Yuan et al., 2019). By avoiding the extremes of overly robotic or overly human-like qualities, these synthesised voices may achieve an optimal blend that mitigates negative listener reactions and promotes trustworthiness, strengthening the case that HAI is informed by human-to-human communication (Asif, 2024; Kühne et al., 2020; Lee & Nass, 2010; Nass et al., 1994, pp. 72–78). Future work could explore whether this balance is replicable across diverse synthesised voice designs or remains specific to the voice stimuli used in this study.

5.4. Human voices with a trustworthy intent outperform real-world synthesised voices

In contrast to neutral human voices, when human speakers modulated their voice with the intent to sound trustworthy, they outperformed synthesised voices in trustworthiness ratings. This finding shows the unique expressive advantage of human speakers, who can adjust vocal traits and convey emotional nuances that remain challenging for current voice-based IA systems to replicate (Nass & Brave, 2005).

One way to interpret this result is based on the previous discussion section where the real-world synthesised voices used in this study appear to have been engineered with middle-ground values in features such as speech rate and mean pitch when compared to human neutral and human trustworthy intent voices from this study. Another likely interpretation though, could lie in listeners' sensitivity to deliberate manipulations of vocal cues in human speakers, potentially due to increased familiarity with human voices rather than synthesised voices. Intentional adjustments in pitch, intonation, emotional tone and speech rate appear to enhance perceptions of positive qualities and emotions linked to trustworthiness (Belin et al., 2019; Torre et al., 2020; Torre et al., 2018; Yokoyama & Daibo, 2012; Zhang et al., 2025). By contrast, synthesised voices, while consistent, may lack the emotional depth required to evoke similar responses. These findings emphasise the need for voice synthesis technologies to move beyond consistency and explore methods for imbuing voices with greater emotional and contextual adaptability, particularly in applications requiring high levels of trust, such as healthcare or counselling services.

5.5. The role of speaker and listener ethnicity

Speaker and listener ethnicities emerged as critical factors shaping trustworthiness ratings, reinforcing the significant role of biases and social dynamics in voice perception. The finding that White speakers were consistently rated as more trustworthy than Black and south Asian speakers – regardless of speaker nature and intent – highlights how vocal trust evaluations may be shaped by both acoustic profiles and ingrained social biases. While our earlier analysis identified certain acoustic cues – such as faster speech rate, higher mean pitch and lower HNR – as predictive of trustworthiness, these features also tended to cluster in White speakers within our dataset (see Tables 2 and 3). On the surface, this could suggest that acoustic properties alone explain trustworthiness ratings. However, such a view risks overlooking how listeners may map socially learned associations onto voice characteristics.

For instance, faster and clearer speech has been linked to competence in past research (Rodero et al., 2014; Yokoyama & Daibo, 2012), but these traits may also be more readily recognised and rewarded when they align with dominant cultural norms – such as standardised, native English speech patterns – particularly in native English countries like the UK and U.S. (Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017). Similarly, the presence of lower pitch variability in Black or south Asian speakers may have activated subtle stereotypes about warmth, competence, or credibility, regardless of their actual vocal performance (Bilal & Barfield, 2021; Gluszek & Dovidio, 2010; Yang et al., 2025). Interestingly, Black listeners gave higher trustworthiness ratings overall, potentially suggesting greater cultural flexibility or broader inclusivity in trustworthiness heuristics. This aligns with literature on familiarity and intergroup trust, which suggests that exposure to diverse voices can mitigate stereotyping in social evaluations (Batsaikhan et al., 2021; Belin et al., 2019; Dahlbäck et al., 2007, pp. 1553–1556; Montoya & Horton, 2013; Zhang et al., 2025). In this way, what appears to be an "acoustic" effect may, in practice, reflect a bias in what counts as trustworthy sounding speech (Lima et al., 2019, pp. 533–538).

Our findings that White, native English speakers were rated as more trustworthy overall, are reinforced by past work showing that non-native or accented speakers are often rated less favourably on social impressions, compared to native speakers, even when content is controlled (Cambre & Kulkarni, 2019; Dahlbäck et al., 2007, pp. 1553–1556; Geiger et al., 2023; Torre et al., 2024). Such judgments are not only culturally constructed but also deeply entangled with racialised and linguistic differences in society (Bilal & Barfield, 2021; Gluszek & Dovidio, 2010; Visser & El Fakiri, 2016). That these biases persist even in relatively controlled experimental conditions signals the need for caution in how voice is operationalised in voice-based IA design.

Taken together, these findings suggest that while acoustic features contribute meaningfully to trustworthiness evaluations, they likely interact with social identity cues and listener expectations. This intertwined relationship draws attention to the importance of considering both vocal and sociocultural factors when designing voice-based IAs (Aylett et al., 2017; Greenwald & Banaji, 1995). Rather than viewing acoustic optimisation in isolation, developers may benefit from a more holistic approach – one that also reflects on how voice design can accommodate diverse listener backgrounds and reduce potential bias in trust-related judgments. Our current and past research, such as (Maltezou-Papastylianou et al., 2023, 2024, 2025a,b), have led us to create a set of guidelines (see Table 12), aiming to help researchers and

Table 12
Evidence-based design recommendations for enhancing vocal trustworthiness in human and synthesised voice applications.

Guidelines	Design insights
G1. Consider the situational relevance of warmth and competence	Trustworthiness aligned more closely with warmth than competence in socially-framed, low-stakes settings. However, competence may dominate in intellectual, task-based or evaluative situational contexts (e.g., legal or academic advice). Vocal tone should reflect contextual priorities.
G2. Expressiveness can enhance trust - but must fit the context and user	Intentional modulation increases trust ratings, especially for human voices. However, expressiveness may backfire for sceptical users (e.g., high-NARS individuals) or where emotional neutrality may be expected (e.g., security alerts).
G3. Calibrate human-likeness in synthetic voice design	The synthesised voices in the current study exhibited acoustic values in-between those of neutral and intentionally trustworthy human speech. This may have helped them avoid sounding too robotic or too human – balancing familiarity with predictability. Overly human-like voices risk triggering the "uncanny valley" or inflating user expectations. For example, a highly realistic voice in a basic customer service assistant may signal higher competence and inflate users' social expectations, which, if unmet, may lead to frustration or mistrust. Designers should not only calibrate vocal realism, but also proactively manage user expectations through onboarding, disclosure of capabilities, and situational framing.
G4. Anticipate and accommodate user predispositions	Listener biases influenced trust outcomes in the current study. High generalised trust improved ratings across the board, while robot-related scepticism (NARS) reduced ratings for synthetic voices. Tailoring delivery styles to audience characteristics may improve engagement, e.g., more emotionally neutral tones for high-NARS users.
G5. Be aware of voice-based social bias – and do not neutralise by default	White voices were consistently rated as more trustworthy than Black or south Asian voices. Attempts to "neutralise" voice identity may obscure rather than correct bias. Instead, evaluation processes should be inclusive and bias-aware, especially in high-impact settings.
G6. Time and pitch are powerful cues – but require restraint	Faster speech rate and higher pitch were associated with more favourable ratings across traits. These features can convey energy, sociability, credibility and engagement. However, excessive modulation may sound unnatural or inappropriate depending on the context. Optimisation must balance clarity, tone, and task demands.
G7. Intentional trust-building works – but never rely on it alone	Expressing vocal intent can boost trustworthiness impressions, especially in early-stage or low-stakes interactions. But its effect depends on situational context, listener expectations, and how strongly the speaker's identity is perceptually categorised. Combine vocal modulation with personalised content, credibility or warmth cues, and expectation management for best results. Design should avoid assuming universal cue interpretation in intentional vocal modulation.
G8. Cultural norms and familiarity shape how vocal cues are interpreted	Features such as pitch variability were expressed and received differently across ethnic groups in our studies – perceived positively among White speakers, but less so or negatively for south Asian voices. These patterns highlight the importance of culturally adaptive voice design, especially in multi-ethnic or global deployments.
G9. When in doubt, design for a "perceptual middle ground" in acoustic expressiveness	In the current study, trustworthy perceptions of synthesised voices tended to occupy acoustic values between neutral and trustworthy-intended human speech. This "perceptual middle ground" as we call it here, may serve as a practical design default when demands on situational context are unclear, or when the product team has yet to determine the appropriate tone of voice. It offers a balance between sounding engaging and avoiding inflated user expectations – particularly useful in early-stage system development or broad public deployment.
G10. First impressions are rapid – design accordingly	Strong trait impressions were formed from brief utterances with a duration of approximately 2 s. For both humans and voice-based IAs, early speech cues (e.g., pitch, pacing) significantly shaped perceived trustworthiness. This is especially relevant in onboarding scenarios, help requests, or cold calls.

industry professionals design more trustworthy, appealing and user friendly voice-based systems. Table 12 encapsulates all our research to show that designing for vocal trustworthiness requires more than replicating human-like features or optimising for clarity. It demands an adaptive, context-sensitive approach that accounts for who is speaking, who is listening, and the social function of the interaction. Whether in public speaking, healthcare communication, or voice interface design, the evidence presented here advocates for a shift away from universal design rules toward a more modular, data-driven understanding of what builds (or breaks) trust in vocal communication. Our guidelines are intended not as fixed prescriptions, but as a flexible framework to support inclusive, informed, and psychologically grounded voice design for human speakers and voice-based IAs.

6. Limitations and future directions

This study focused on English-speaking voices across three ethnic groups, offering insights into vocal trustworthiness across speaker-listener pairings. However, future research should extend this work by incorporating greater linguistic diversity – including multilingual and accented voices – to assess how cultural familiarity and linguistic variation interact with trust judgments, particularly in non-Western populations. Although our synthesised voice stimuli reflected real-world text-to-speech (TTS) technology, they were limited to pre-existing commercial systems with fixed prosodic styles. As voice synthesis continues to evolve, future studies should examine how more expressive or emotionally adaptive systems affect listener trust, especially in sensitive or high-stakes contexts like healthcare, finance and education.

In addition to quantitative analysis, future studies should consider incorporating mixed methods – such as follow-up interviews or trust calibration tasks – to help uncover the reasoning behind participants' ratings. This may clarify the role of implicit biases, expectations, or perceived speaker intent that underlie observed behaviours. Finally, trust is context-sensitive. Our controlled, perception-based design cannot fully capture the dynamics of real-time interaction. Testing voice

trustworthiness in applied settings – e.g., virtual customer service, AI tutoring, or medical triage simulations – will help validate whether the effects observed here generalise to practical use cases.

Lastly, while our study did not measure human speaker effort or vocal strain, some research suggests that faster speech may reflect increased cognitive or muscular load on the speaker (Anikin, 2023; Nudelman et al., 2024). Future work might explore whether such physiological features of vocal production influence listener perceptions of trustworthiness, particularly in more interactive or longer duration speech contexts.

7. Conclusion

This study advances our understanding of how trustworthiness is evaluated in ethnically diverse human and synthesised voices by highlighting the joint influence of acoustic features (speech rate, mean fundamental frequency, HNR, jitter, shimmer, CPP, and LTAS), speaker intent, and listener attitudes toward robots. Real-world synthesised voices – demonstrated balanced acoustic properties that were positioned between perceived neutral and trustworthy human voices, a term that we call here as a "perceptual middle ground" – and were rated as more trustworthy than human voices with neutral intent (see Tables 2–4 for the acoustic values). However, modulated human voices intended to convey trustworthiness still outperformed voice-based IAs, reaffirming the enduring advantage of expressive control and emotional nuance in human communication.

Trust-related impressions were not purely acoustic-based. Listener attitudes, particularly scepticism toward interacting with robots, also influenced ratings, drawing attention to the role of cognitive predispositions in HAI. Moreover, consistent patterns of higher ratings for White speakers across listener groups point to the influence of broader sociocultural expectations, highlighting the importance of further investigating how implicit biases may shape voice evaluations in both human and voice-based IA contexts.

These findings highlight a key implication: optimising trust in voice-based IAs requires more than refining acoustic signal properties – it requires culturally sensitive, psychologically informed design choices that reflect the diversity of real-world users (see Table 12 for our research-informed guidelines). As voice technologies become increasingly embedded in education, healthcare, finance and public services, their ability to inspire trust across social groups will be central to their success. Future work should continue to examine how contextual factors, user expectations, and social dynamics converge to shape trust in both human and artificial speakers.

CRediT authorship contribution statement

Constantina Maltezou-Papastylianou: Writing – review & editing, Writing – original draft, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Reinhold Scherer:** Writing – review & editing, Supervision. **Silke Paulmann:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Declaration of the use of AI assisted technologies

During the preparation of this work the authors used Grammarly for proof reading and spell checking. The authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets of the current study are available on the OSF (https://osf.io/sfb3g/).

References

Abdi, H. (2010). Holm's sequential Bonferroni procedure. *Encyclopedia of research design, 1*, 1–8.
Anikin, A. (2023). The honest sound of physical effort. *PeerJ, 11*, Article e14944.
Asif, M. S. (2024). Voicing trust: An acoustic analysis of trustworthiness in automated customer service interfaces. *Voicing trust: An acoustic analysis of trustworthiness in automated customer service interfaces*. University of Twente.
Aylett, M. P., Vinciarelli, A., & Wester, M. (2017). Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing, 11*, 361–372.
Bachorowski, J.-A., & Owren, M. J. (1995). Vocal expression of emotion: Acoustic properties of speech are associated with emotional intensity and context. *Psychological science, 6*, 219–224.
Batsaikhan, M., He, T.-S., & Li, Y. (2021). Accents, group identity, and trust behaviors: Evidence from Singapore. *China Economic Review, 70*, Article 101702.
Baus, C., McAleer, P., Marcoux, K., Belin, P., & Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Scientific Reports, 9*. https://doi.org/10.1038/s41598-018-36518-6
Belin, P., Boehme, B., & McAleer, P. (2019). Correction: The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *PLoS One, 14*, Article e0211282.
Bilal, D., & Barfield, J. (2021). *Increasing racial and ethnic diversity in the design and use of voice digital assistants*.
Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.
Cambre, J., & Kulkarni, C. (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on human-computer interaction, 3*, 1–19.
Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.
Chan, M. P., & Liberman, M. (2021). *An acoustic analysis of vocal effort and speaking style* (Vol. 45). AIP Publishing.
Da Silva, P. T., Master, S., Andreoni, S., Pontes, P., & Ramos, L. R. (2011). Acoustic and long-term average spectrum measures to detect vocal aging in women. *Journal of Voice, 25*, 411–419.
Dahlbäck, N., Wang, Q., Nass, C., & Alwin, J. (2007). *Similarity is more important than expertise: Accent effects in speech interfaces*.
Eberl, M. (2016). Fisher-yates shuffle. *Arch. Formal Proofs, 2016*, 19.
Feinberg, D. (2022). *VoiceLab: Software for Fully Reproducible Automated Voice Analysis*, 351–355.
Feinberg, D. R., & Cook, O. (2020). *VoiceLab: Automated reproducible acoustic analysis*.
Felippe, A. C., Grillo, M. H., & Grechi, T. H. (2006). Standardization of acoustic measures for normal voice patterns. *Revista Brasileira de Otorrinolaringologia, 72*, 659–664.
Fernandes, J., Teixeira, F., Guedes, V., Junior, A., & Teixeira, J. P. (2018). Harmonic to noise ratio measurement-selection of window and length. *Procedia computer science, 138*, 280–285.
Ferrand, C. T. (2002). Harmonics-to-noise ratio: An index of vocal aging. *Journal of Voice, 16*, 480–487.
Field, A. (2018). *Discovering statistics using IBM SPSS statistics*. Sage publications limited.
Gabrieli, G., Ng, S., & Esposito, G. (2021). Hacking trust: The presence of faces on automated teller machines (ATMs) affects trustworthiness. *Behavioral Sciences, 11*, 91.
Geiger, M. K., Langlinais, L. A., & Geiger, M. (2023). Accent speaks louder than ability: Elucidating the effect of nonnative accent on trust. *Group & Organization Management, 48*, 953–965.
Ghorayeb, A., Comber, R., & Gooberman-Hill, R. (2021). Older adults' perspectives of smart home technology: Are we developing the technology that older people want? *International Journal of Human-Computer Studies, 147*, Article 102571.
Gluszek, A., & Dovidio, J. F. (2010). The way they speak: A social psychological perspective on the stigma of nonnative accents in communication. *Personality and Social Psychology Review, 14*, 214–237.
Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: Attitudes, self-esteem, and stereotypes. *Psychological Review, 102*, 4.
Gussenhoven, C. (2002). *Intonation and interpretation: Phonetics and Phonology*.
Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta Oto-Laryngologica, 90*, 441–451.
Hanzlíková, D., & Skarnitzl, R. (2017). Credibility of native and non-native speakers of English revisited: Do non-native listeners feel the same? *Research in Language, 15*, 285–298.
Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Huang, D., Markovitch, D. G., & Stough, R. A. (2024). Can chatbot customer service match human service agents on customer satisfaction? An investigation in the role of trust. *Journal of Retailing and Consumer Services, 76*, Article 103600.
Jalali-najafabadi, F., Gadepalli, C., Jarchi, D., & Cheetham, B. M. (2021). Acoustic analysis and digital signal processing for the assessment of voice quality. *Biomedical Signal Processing and Control, 70*, Article 103018.
Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust technology. *The measurement of the propensity to trust technology*. Springer International Publishing.
Kepuska, V., & Bohouta, G. (2018). *Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home)*. IEEE.
Kim, J., Gonzalez-Pumariega, G., Park, S., & Fussell, S. R. (2023). *Urgency builds trust: A voice agent's emotional expression in an emergency*.
Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement, 56*, 746–759.
Ko, S., Liu, X., Mamros, J., Lawson, E., Swaim, H., Yao, C., & Jeon, M. (2020). The effects of robot appearances, voice types, and emotions on emotion perception accuracy and subjective perception on robots - HCI international 2020 - late breaking papers. *Multimodality and intelligence: 22nd HCI international conference, HCII 2020, Copenhagen, Denmark, july 19–24, 2020, proceedings*. https://doi.org/10.1007/978-3-030-60117-1_13
Kouravanas, N., & Pavlopoulos, A. (2022). Social robots: The case of robot sophia. *Homo Virtualis, 5*, 136–165.
Krantz, A., Balkenius, C., & Johansson, B. (2022). Using speech to reduce loss of trust in humanoid social robots. *arXiv preprint arXiv:2208.13688*.
Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.
Kühne, K., Fischer, M. H., & Zhou, Y. (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*. https://doi.org/10.3389/fnbot.2020.593732
Large, D. R., Harrington, K., Burnett, G., Luton, J., Thomas, P., & Bennett, P. (2019). To please in a pod: Employing an anthropomorphic agent-interlocutor to enhance trust and user experience. In *An autonomous, self-driving vehicle* (pp. 49–59).
Lee, J.-E. R., & Nass, C. I. (2010). Trust in computers: The computers-are-social-actors (CASA) paradigm and trustworthiness perception in human-computer communication. *IGI Global*.
Liao, Y., & He, J. (2020). *Racial mirroring effects on human-agent interaction in psychotherapeutic conversations*.
Lim, M. Y., Lopes, J. D., Robb, D. A., Wilson, B. W., Moujahid, M., De Pellegrin, E., & Hastie, H. (2022). We are all individuals: The role of robot personality and human traits in trustworthy interaction - 2022 31st. *IEEE international conference on robot and human interactive communication (RO-MAN)*. https://doi.org/10.1109/RO-MAN53752.2022.9900772
Lima, L., Furtado, V., Furtado, E., & Almeida, V. (2019). *Empirical analysis of bias in voice-based personal assistants*.
Linville, S. E. (2002). Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice, 16*, 472–479.
Löfqvist, A. (1986). The long-time-average spectrum as a tool in voice research. *Journal of Phonetics, 14*, 471–475.
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S.. Can I trust you? Discovering how trustworthiness is communicated through vocal cues. https://doi.org/10.17605/OSF.IO/SFB3G.
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S.. Trustworthy Intent in Speech (TIS) Corpora Dataset. https://doi.org/10.17605/OSF.IO/45D8J.
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2024). Acoustic classification of speech with trustworthy intent. In. *Proc. SpeechProsody, 2024*, 961–964.
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2025). How do voice acoustics affect the perceived trustworthiness of a speaker? A systematic review. *Frontiers in Psychology, 16*. https://doi.org/10.3389/fpsyg.2025.1495456
Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2025). Human voices communicating trustworthy intent: A demographically diverse speech audio dataset. *Scientific Data, 12*(1), 921.
Matthews, G., Lin, J., Panganiban, A. R., & Long, M. D. (2019). Individual differences in trust in autonomous robots: Implications for transparency. *IEEE Transactions on Human-Machine Systems, 50*, 234–244.
Maxim, A., Zalake, M., & Lok, B. (2023). *The impact of virtual human vocal personality on establishing rapport: A study on promoting mental wellness through extroversion and vocalics*.
McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? Personality impressions from brief novel voices. *PLoS One, 9*. https://doi.org/10.1371/journal.pone.0090779
Montepare, J. M., Kempler, D., & McLaughlin-Volpe, T. (2014). The voice of wisdom: New insights on social impressions of aging voices. *Journal of Language and Social Psychology, 33*, 241–259.
Montoya, R. M., & Horton, R. S. (2013). A meta-analytic investigation of the processes underlying the similarity-attraction effect. *Journal of Social and Personal Relationships, 30*, 64–94.
Mori, M. (1970). Bukimi no tani [The uncanny valley]. *Energy, 7*, 33.
Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics and Automation Magazine, 19*, 98–100.
Muralidharan, L., de Visser, E. J., & Parasuraman, R. (2014). The effects of pitch contour and flanging on trust in speaking cognitive agents - CHI '14 extended abstracts on. *Human Factors in Computing Systems*, 2167–2172. https://doi.org/10.1145/2559206.2581231
Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.
Nass, C., Steuer, J., & Tauber, E. R. (1994). *Computers are social actors*.
Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006a). *Altered attitudes of people toward robots: Investigation through the negative attitudes toward robots scale, 2006* pp. 29–35).
Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006b). Measurement of negative attitudes toward robots. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems, 7*, 437–454.
Nudelman, C., Webster, J., & Bottalico, P. (2024). The effects of reading speed on acoustic voice parameters and self-reported vocal fatigue in students. *Journal of Voice, 38*, 243.
Ohala, J. J. (1983). Cross-language use of pitch: An ethological view. *Phonetica, 40*, 1–18.
Ohala, J. J. (1995). *The frequency code underlies the sound-symbolic use of voice pitch*. Sound symbolism.
Oleszkiewicz, A., Pisanski, K., Lachowicz-Tabaczek, K., & Sorokowska, A. (2017). Voice-based assessments of trustworthiness, competence, and warmth in blind and sighted adults. *Psychonomic Bulletin & Review, 24*, 856–862. https://doi.org/10.3758/s13423-016-1146-y
Prolific. (2014). *Prolific: Quickly find research participants you can trust. Prolific*. Quickly find research participants you can trust. Retrieved from https://www.prolific.com/.
Puts, D. A., Hodges, C. R., Cárdenas, R. A., & Gaulin, S. J. (2007). Men's voices as dominance signals: Vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior, 28*, 340–344.
Radford, N. A., Strawser, P., Hambuchen, K., Mehling, J. S., Verdeyen, W. K., Donnan, A. S., … Bridgwater, L. (2015). Valkyrie: Nasa's first bipedal humanoid robot. *Journal of Field Robotics, 32*, 397–419.
Rodero, E., Mas, L., & Blanco, M. (2014). The influence of prosody on politicians' credibility. *Journal of Applied Linguistics and Professional Practice, 11*.
Schild, C., Braunsdorf, E., Steffens, K., Pott, F., & Stern, J. (2022). Gender and context-specific effects of vocal dominance and trustworthiness on leadership decisions. *Adaptive Human Behavior and Physiology, 8*, 538–556.
Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*, 15–25.
Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). *Voice in human-agent interaction: A survey* (Vol. 54, pp. 1–43). ACM Computing Surveys (CSUR).
Shigemi, S., Goswami, A., & Vadakkepat, P. (2018). ASIMO and humanoid robot research at Honda. *Humanoid robotics: A reference, 55*, 90.
Smith, S. M., & Shaffer, D. R. (1995). Speed of speech and persuasion: Evidence for multiple effects. *Personality and Social Psychology Bulletin, 21*, 1051–1060.
Syed, I., Baart, M., & Vroomen, J. (2024). The multimodal trust effects of face, voice, and sentence content. *Multisensory Research, 37*, 125–141.
Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European Journal of Social Psychology, 35*, 413–424.
Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior, 105*. https://doi.org/10.1016/j.chb.2019.106215
Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). *Trust in artificial voices: A" congruency effect" of first impressions and behavioural experience*.
Torre, I., White, L., Goslin, J., & Knight, S. (2024). The irrepressible influence of vocal stereotypes on trust. *Quarterly Journal of Experimental Psychology, 77*, 1957–1966.
Visser, M. A., & El Fakiri, F. (2016). The prevalence and impact of risk factors for ethnic differences in loneliness. *The European Journal of Public Health, 26*, 977–983.
Yang, H., Song, H., Wang, Y.-C., & Ma, E. (2025). Similarity-attraction theory perspective on service employees and service robots' interactions. *Service Industries Journal*, 1–24.
Yokoyama, H., & Daibo, I. (2012). Effects of gaze and speech rate on receivers' evaluations of persuasive speech. *Psychological Reports, 110*, 663–676. https://doi.org/10.2466/07.11.21.28.PR0.110.2.663-676
Yuan, L., Dennis, A., & Riemer, K. (2019). *Crossing the uncanny valley? Understanding affinity, trustworthiness, and preference for more realistic virtual humans in immersive environments*.
Zhang, Q., Yang, X. J., & Robert, L. (2025). *Voice Similarity and Its Impact on Cognitive and Affective Trust in Automated Vehicles*.