Communication of trustworthiness in human and synthesised speech

Author

Constantina Maltezou


0000-0002-9443-4226

A thesis submitted for the degree of

Doctor of Philosophy

Faculty of Science and Health

(cross-departmental PhD, Department of Psychology and

School of Computer Science and Electronic Engineering)

University of Essex

September, 2025

Abstract

This thesis investigates how trustworthiness is communicated in human and synthesised speech. Prior studies, largely based on young, white, Western participants, have overlooked how vocal trustworthiness perceptions may vary across age and ethnicity. This work addresses that gap by including under-represented groups such as older black and south Asian speakers and listeners.

The thesis adopts a multi-stage design, beginning with a systematic review that maps conceptual and methodological gaps in the literature. The following studies aim to address these limitations. Study 1 introduces an open-access dataset of 1,152 speech samples from 96 demographically diverse speakers (by age, sex, and ethnicity), examining how trustworthy vs neutral vocal intent is expressed. Drawing on this dataset, Study 2 explores how perceived trustworthiness aligns with core social perceptions of warmth and competence, as shaped by vocal cues. Multi-part Study 3 extends the analysis to the role of cognitive biases, demographic variation, and trust predispositions. Finally, Study 4 compares human and real-world synthesised voices, bridging trustworthiness perception across speaker identities.

Collectively, the findings converge on a set of acoustic features — perceived pitch, speech rate, HNR, shimmer and LTAS — that reliably shape trustworthiness impressions across both human and synthesised voices. Trustworthiness impressions were boosted with explicit vocal intent, especially where speaker group membership was uncertain or ambiguous. Listener predispositions shaped evaluations in distinct ways, highlighting that trustworthiness impressions are not only about how a voice sounds, but also who is

listening and what they bring to this process. By integrating speaker intent, listener bias, and speaker nature (human vs synthesised), the thesis offers invaluable cross-validation of trust-relevant acoustic cues across production and perception replication of trustworthy human and synthesised voices. It advances theoretical models of vocal impression formation and offers practical guidance for designing socially attuned, trust-enhancing voice technologies for diverse user groups.

Acknowledgments

These past three years have been profoundly transformative – not only in terms of academic growth, but also through deep personal change, loss, and healing. Behind the work of this thesis is a journey shaped as much by life experience as by research, and it would not have been possible without the support, presence, and care of more people than I can actually mention here.

To my supervisors, **Prof. Silke Paulmann** and **Prof. Reinhold Scherer** – I'm sincerely grateful to both of you for walking this journey with me and supporting me, even amidst your demanding roles leading the psychology and computer science departments. Silke, thank you for your thoughtfulness, encouragement, and psycholinguistic insight, which often brought clarity and comfort when it mattered most. Reini, thank you for bringing technical rigour and encouraging me to explore the machine learning side of things. I'm grateful for the complementary perspectives you brought to this interdisciplinary work. I'll always remember our chats with a smile.

To my panel chair, **Helge** – your upbeat demeanour and constructive approach was always appreciated. **To Trudi** – thank you for your kindness, care, and calm during uncertain moments. You, **Rick**, **the library staff**, and **my fellow doctoral researchers** brought fun times and compassion into the corners of academic life – it meant more than I can say.

To my husband, **Tasos** – for our shared love of learning, and the small everyday moments that added colour to these years. From feeding and photographing the park animals to baking and spontaneous day trips – these moments were gentle reminders of the parts of life that

exist beyond work.

To my mum, Elli – your love and strength have always been quiet, but never small. You held our family through storms with unwavering grace, and your faith and presence have grounded me in more ways than you'll ever know. Thank you for teaching me resilience and compassion without ever naming it. **To my dad, Antonis** – for your vivid spirit and enduring presence, even in absence. You're missed more than I can say. **To my grandparents** – thank you for the warmth and comfort that has always felt like home.

To my friends – old and new – thank you for the light you brought into my life. Your presence helped me hold onto joy and perspective when I needed it.

To Lambros – thank you for offering a steady ear and a space to reflect, heal, and grow during some of the most difficult moments of my life.

To Fr Michalis and Fr Ian – your guidance and prayers were anchors of calm and hope. Thank you for walking beside me in both grief and growth. **To the congregation of St Helen and St Martin** – thank you for welcoming me with open arms, and helping me rediscover the joy of belonging. Making prosphoro – and chanting for Christ – became a gentle balm for the soul.

Above all, I'm thankful **to God** – for carrying me through grief, growth, and hope, in all aspects of my life.

With gratitude,

Constantina Maltezou

Contents

List of Figures

List of Tables

List of Abbreviations

Table 1: List of all common abbreviations used in this thesis and their corresponding full terms

Abbreviation	Full term
AI	Artificial intelligence
APQ3	Amplitude perturbation quotient 3
AUC	Area under the curve
CPP	Cepstral peak prominence
HAI	Human-agent interaction
HNR	Harmonics-to-noise ratio
IA	Intelligent agent
LMM	Linear mixed-effects model
LOSO CV	Leave-one-speaker-out cross-validation
LTAS	Long-term average spectrum
MFCCs	Mel-frequency cepstral coefficients
NARS	Negative attitudes toward robots scale
OSF	Open science framework

Continued on next page

Table 1: List of all common abbreviations used in this thesis and their corresponding full terms (Continued)

Abbreviation	**Full term**
PRISMA	Preferred reporting items for systematic reviews and meta-analyses
RAP	Relative average perturbation
ROC curve	Receiver operating characteristic curve
SCM	Stereotype content model
TAM	Technology acceptance model
TTS	Text-to-speech
WWIB	White western individualist bias

Introduction

Trust is a multidimensional and dynamic socio-cognitive construct that influences human relationships across all levels of society —— from interpersonal and organisational to political and institutional. The ability to judge another's trustworthiness is essential to social capital, and these judgements are often shaped not only by observable behaviour, but by subtle cues —— including how someone speaks. The human voice, a rich channel of social and emotional information, plays a central role in shaping first impressions of core social traits such as trustworthiness, warmth, and competence (Cuddy, Fiske, & Glick, 2008; McAleer, Todorov, & Belin, 2014; Oleszkiewicz, Pisanski, Lachowicz-Tabaczek, & Sorokowska, 2017). These, often implicit first impressions drawn from "thin-slice" interactions (Gheorghiu, Callan, & Skylark, 2020), are critical for navigating increasingly diverse social environments, be they casual or high-stakes.

Humans are evolutionarily attuned to make rapid social inferences from vocal cues —— often within milliseconds of hearing a voice (Lavan, 2023). As voice-based technologies have become increasingly and seamlessly embedded in daily life, these rapid judgements now extend beyond human interactions to include artificially generated speech (i.e., synthesised voices). From voice assistants and customer service bots, to voice-enabled navigation systems in autonomous vehicles, and humanoid robots, synthesised voices now mediate many routine decisions and interactions. As a result, questions of trust in voice —— both human and synthesised —— are no longer theoretical but practically urgent: How do we judge whether a voice is trustworthy? What cues shape this judgement? And how are these cues

interpreted across speakers and listeners, diverse in nature (human vs artificial intelligence) and demographically?

This thesis investigates how trustworthiness is communicated (i.e., expressed and perceived) in both human and synthesised voices, with a focus on the acoustic, psychological, and identity (nature, demographics) factors that shape these impressions. The investigation follows a logically sequenced progression that begins with speaker-side vocal production —– specifically, how speakers express trustworthiness through intentional vocal modulation. It then shifts to the listener's perspective on trustworthiness impressions, investigating how these cues are perceived, interpreted, and shaped by social biases and trust predispositions. Finally, it integrates these insights into a comparative analysis of human and synthesised voices, highlighting the acoustic and perceptual relationship that influences trust-related evaluations in voice-based interactions, whether natural or synthesised. By linking production, perception, and cross-domain comparison, the thesis offers a rare, multi-level perspective on the acoustic foundations of vocal trust —– laying the groundwork for future models of human-robot communication.

Chapter 1 begins by mapping the existing literature through a systematic review of voice-based trustworthiness. Organised by speaker nature (human vs synthesised) and situational context (e.g., generic first impressions, public communication, telehealth, etc), this review identifies recurring acoustic features, methodological variations, and conceptual gaps. While the perceived pitch of speakers has dominated the literature, the review reveals that combinations of different acoustic features —– including prosodic and voice quality measures —– offer more consistent explanatory power. It also highlights the need for broader demographic representation and for models that incorporate listener predispositions (e.g., trust propensity, social biases). To the best of my knowledge, this constitutes the first structured synthesis of the voice-based trustworthiness literature to date, and has been peer-reviewed and published as an academic journal article (Maltezou-Papastylianou, Scherer, & Paulmann, 2025). The review serves not only as a field overview, but as a conceptual blueprint that guides the thesis's empirical investigations and research priorities (see Table 1.7).

Chapter 2 addresses a key limitation identified in the review: the lack of standardised and openly accessible datasets on voice trustworthiness that include speakers from diverse ethnic and age backgrounds. This chapter introduces a novel open-access dataset of 1,152 speech audio samples from 96 speakers spanning ethnicity (white, black, south Asian), sex, and age-groups (younger and older than 60 years). Each speaker contributed both "neutral" (i.e., natural tone of voice) and trust-intended utterances, enabling systematic comparison of vocal intent across demographic profiles. Acoustic classification analyses revealed that key features —– such as perceived pitch, harmonics-to-noise ratio (HNR), shimmer and long-term average spectrum (LTAS) —– were modulated in consistent, interpretable ways when speakers intended to sound trustworthy. While not hypothesis-driven in structure, this chapter offers both a methodological data resource to the field and foundational empirical evidence of how trust-signalling is vocally produced across speaker demographics. Published as a peer-reviewed data descriptor alongside its reusable audio dataset (Maltezou-Papastylianou, Scherer, & Paulmann, 2024b), it also provides the acoustic and demographic scaffolding for the experimental studies that follow.

Chapter 3 shifts to the listener's perspective by examining how speakers' vocal cues shape trustworthiness impressions, and whether vocal trustworthiness is perceived as a multidimensional construct. Drawing on theoretical models that frame trustworthiness as rooted in perceived ability and benevolence (e.g., Mayer, Davis, & Schoorman, 1995), the study evaluates whether impressions of trustworthiness align more closely with those of warmth and competence —– two universal dimensions in social cognition (Fiske, Cuddy, & Glick, 2007). Using the standardised dataset introduced in Chapter 2, this study investigates how speakers' intent to sound trustworthy influences listeners' impressions of trustworthiness, warmth and competence. Meanwhile, it also identifies which acoustic and voice quality features shape those impressions. In doing so, the chapter offers conceptual and empirical clarity on how trust is perceived from voice alone, and contributes to a more integrated model of vocal first impressions.

Chapter 4 extends the findings from Chapter 3 by introducing speaker-listener ethnic and age group membership, and listener predispositions, into the trustworthiness perception

framework. It examines (1) whether in-group speakers (matched by age and ethnicity) are judged as more trustworthy, (2) whether vocal intent can mitigate such bias, and (3) how listeners' propensity toward generalised and particularised trust affects their judgements. Results reveal complex interactions between vocal intent, speaker-listener demographics, and listeners' predispositions. For example, there were cases where intentional expressiveness improved out-group ratings and where trust predispositions influenced judgments independently of speaker characteristics. This chapter contributes to both voice perception and social cognition research by showing how these factors co-determine trustworthiness impressions in voice-only settings.

Chapter 5 serves as the thesis' final empirical step, integrating the findings from human speech production and perception into a comparative analysis of human and synthesised voices. Building on earlier chapters, it examines whether acoustic cues shown to influence trustworthiness in human speech similarly shape evaluations of commercially available, real-world synthesised voices. It also introduces listener-level variability via the Negative Attitudes toward Robots Scale (NARS), assessing how attitudes toward robots influence trust in artificially-intelligent agents. Synthesised voices were rated as more trustworthy than human-neutral voices but less so than human voices modulated with trust intent —– a pattern that aligns with their intermediate acoustic profile. These voices tended to exhibit acoustic values that fell between neutral and trust-modulated human speech. Thus, suggesting a "perceptual middle ground" where moderate expressiveness enhances perceived trustworthiness in casual, social impressions. This chapter completes the thesis' progression from production, to perception, to human–agent interaction —– offering a multi-level perspective on how vocal trustworthiness is constructed, interpreted, and potentially engineered across natural and synthesised modalities.

In sum, this thesis bridges inter-disciplinary behavioural research on human and synthesised voices to offer a multi-layered understanding of how trustworthiness is signalled and perceived through speech. By integrating speaker intent, acoustic cues, listener predispositions, and cross-domain comparison, it provides an empirically grounded and conceptually coherent framework for advancing voice research on trustworthiness. The

findings inform both theoretical models of vocal impression formation and the practical design of inclusive, context-sensitive voice technologies in an increasingly digital society.

Chapter 1

Systematic review of the literature

1.1. Introduction

Digitisation is changing the way modern societies interact and communicate. The use of artificial intelligence (AI) and speech synthesis has entered many domains of our daily life, such as autonomous vehicles, automated customer support, telehealth and companion robots, and smart home assistants. Considering that trust is a key factor in the acceptance of technology (Bryant, Borenstein, & Howard, 2020; Large et al., 2019; Seaborn, Miyake, Pennefather, & Otake-Matsuura, 2021) as well as the healthy functioning of a flourishing society, it makes the multi-disciplinary research area of trustworthy voice acoustics of growing importance and relevance. Overall, existing literature suggests that speech acoustics influence first impressions of speakers' perceived trustworthiness (C. Nass & Lee, 2000; Oleszkiewicz et al., 2017; Stewart & Ryan, 1982; Tsantani, Belin, Paterson, & McAleer, 2016). Nonetheless, when biological, demographic, cultural, and situational factors are not adequately considered, the overall findings often remain inconclusive. To the best of my knowledge, this is the first systematic review that aims to understand the relationship between voice acoustics and attributions of trustworthiness in humans and machines.

1.1.1. The physiology of voice perception and speech acoustics

By merely hearing a stranger's voice, such as a telemarketer, we tend to form instant impressions of their identity, discerning cues like gender, age, accent, emotional state, personality traits (e.g. perceived trustworthiness), and even hints about their health condition (cf. Kreiman & Sidtis, 2011; C. I. Nass & Brave, 2005). Voice, the carrier of speech, allows us to perceive human traits through auditory signals generated during speech production. Physiologically, during speech production, airflow from the lungs is transformed into sound waves by vocal fold vibrations in the larynx, and these waves are shaped by the vocal tract's articulators, producing the diverse sounds of speech (cf. source-filter theory Kamiloğlu & Sauter, 2021; Lieberman, Laitman, Reidenberg, & Gannon, 1992).

Table 1.1 exhibits certain acoustic features and how speech acoustics shape first impressions during social interactions (Bachorowski & Owren, 1995; Cascio Rizzo & Berger, 2023; Maltezou-Papastylianou, Russo, Wallace, Harmsworth, & Paulmann, 2022; Shen, Elibol, & Chong, 2020; Weinstein, Zougkou, & Paulmann, 2018). Voice quality features such as Harmonics-to-noise ratio (HNR), jitter, shimmer, cepstral peak prominence (CPP) and long-term average spectrum (LTAS) tend to be indicative of the perceived roughness, breathiness or hoarseness of a voice, often seen in vocal aging and pathologies research (Chan & Liberman, 2021; Da Silva, Master, Andreoni, Pontes, & Ramos, 2011; Farrús, Hernando, & Ejarque, 2007; Jalali-najafabadi, Gadepalli, Jarchi, & Cheetham, 2021; S. E. Linville, 2002). Moreover, past studies seem to suggest that each attributed speaker trait may follow a different time course in terms of stimulus duration (Lavan, 2023; Mahrholz, Belin, & McAleer, 2018; McAleer et al., 2014). For instance, dominance attributions seem to develop as early as 25 milliseconds (ms), while trustworthiness and attractiveness attributions are strengthened gradually over exposure periods ranging from 25 ms to 800 ms (Lavan, 2023).

Table 1.1: Summary characteristics of speech acoustics

Acoustic features	**Typically measured in...**	**Key characteristics**
Fundamental frequency (f_0); perceived as pitch	Hertz (Hz)	- f_0 is the lowest rate of vocal fold vibrations, and f_0 variability is usually captured by vocal intonation within an utterance. - "Size or frequency code" theory (Ohala, 1983, 1995): Men's lower pitch due to longer, thicker folds; women's higher pitch due to shorter folds (Frühholz & Schweinberger, 2021; Latinus & Belin, 2011; Lavan et al., 2019). - Average speaking frequencies: Men, 100-200 Hz; Women, 200-240 Hz; Children, 300 Hz (Gelfand, 2017; Mahendru, 2014; Schweinberger et al., 2014).
Amplitude; perceived as loudness	Decibels (dB)	Indicative of air pressure variations.
Speech rate	Syllables per second (syll/s)	- Typically estimated at about 4-6 syllables per second in English (Reetz & Jongman, 2020). "Effort code" theory (Gussenhoven, 2002): Faster speech rate shown to increase speakers' perceived competence, credibility, trustworthiness and willingness to help (Rodero et al., 2014; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012).

Continued on next page

Table 1.1: Summary characteristics of speech acoustics (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
HNR	dB	- Lower HNR signifies more noise in a voice signal (Fernandes et al., 2018; Ferrand, 2002). Noise in terms of voice, encompasses any component of the signal that interferes with the clarity, purity and overall quality of the intended speech signal. Typically, this noise is not harmonically related to the fundamental frequency of the voice, such as alterations in vocal fold tissue, muscle tension, respiratory patterns, or even ambient sounds and electronic interference (Ferrand, 2002). - Older adults typically show slower speech rates, lower HNR, and differences in pitch and voice quality compared to younger adults (Baus et al., 2019; Ferrand, 2002; Heffernan, 2004; Lavan et al., 2019; McAleer et al., 2014; Rojas et al., 2020).
Jitter	%	Reveals micro-fluctuations in pitch caused by irregular vocal fold vibrations (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).
Shimmer	dB	Measures micro-fluctuations in amplitude, reflecting variations in voice intensity (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).

Continued on next page

Table 1.1: Summary characteristics of speech acoustics (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
CPP	dB	A lower CPP is indicative of a breathy voice (Da Silva et al., 2011; Hammarberg et al., 1980; S. E. Linville, 2002; Löfqvist, 1986).
LTAS	dB	A lower LTAS often indicates longer vocal tract sizes (Da Silva et al., 2011; Hammarberg et al., 1980; S. E. Linville, 2002; Löfqvist, 1986), which are linked to deeper, more resonant voices associated with dominance, particularly observed in males (Gussenhoven, 2002; Puts et al., 2007).
Alpha-ratio	dB	- Provides information about the distribution of energy across different frequency ranges (i.e., the ratio between low-frequency and high-frequency energy within a voice signal; McAleer et al., 2014; Sundberg et al., 2011). - It is often related to voice quality measures, such as the perceptual attributes of vocal effort, breathiness and vocal timbre (Chan & Liberman, 2021).

Continued on next page

Table 1.1: Summary characteristics of speech acoustics (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
Mel-frequency cepstral coefficients (MFCCs)	Unitless	- MFCCs are not voice signals themselves but derived from a multi-step process, including Fourier transformation, that provides a compact representation of the spectral properties of the voice signal (Zheng et al., 2001). They capture important information about the speech sounds while reducing the amount of data. - MFCCs are widely used in various applications such as speech recognition systems, speaker identification, and emotion detection. They are also used in machine learning models to distinguish between high-quality and low-quality voice recordings, or to detect specific voice disorders when combined with other acoustic features (Deng et al., 2024; Rehman et al., 2024).

1.1.2. Definitions of trust and perceived trustworthiness

Trust has been shown to influence perceptions of first impressions (Freitag & Bauer, 2016), personal relationships (Ter Kuile, Kluwer, Finkenauer, & Van der Lippe, 2017), work performance (Brion, Lount Jr, & Doyle, 2015; Lau, Lam, & Wen, 2014), cooperation and sense of safety within communities (Castelfranchi & Falcone, 2010; Krueger, 2021). While extensive literature discusses trust models, most are theoretical (Harrison McKnight & Chervany, 2001; Mayer et al., 1995), offering varying definitions encompassing expected actions (Gambetta, 2000), task delegation (Mayer et al., 1995), cooperativeness (Deutsch,

1960; Yamagishi, 2003; Yamagishi & Yamagishi, 1994), reciprocity (Ostrom & Walker, 2003), and "encapsulated interest" (Baier, 2014; Maloy, 2009). Current research tends to explore trust as either a single-scale or multi-dimensional concept, often focusing on the three-part relation of "A trusts B to do X," within specific contexts (cf. Bauer & Freitag, 2018). Intrinsically, trustee B's perceived trustworthiness to do X is shaped by trustor A's dispositional, learned and situational trust factors, risk assessment and beliefs towards the trustee, such as gender stereotyping in relation to different occupations and contexts (Castelfranchi & Falcone, 2010; Freitag & Bauer, 2016; Seligman, 2000; S. S. Smith, 2010; Tschannen-Moran & Hoy, 2000). Furthermore, social trust formation tends to lean towards a dichotomised view, namely generalised and particularised trust (cf. Freitag & Traunmüller, 2009; Schilke, Reimann, & Cook, 2021; Uslaner, 2002). Overall, trusting someone or perceiving them as trustworthy can be expressed as the trustor's reliance on a trustee (e.g. an individual, a community, an organisation or institution), with the belief or expectation of behaving in a manner that contributes to the trustor's welfare (e.g. by assisting in the completion of a task) or at least not against it (e.g. sharing a secret). In turn, this helps support or induce a sense of mutual benefit between them, all the while, taking into account the situational context and the trustor's predispositions.

Throughout this review, the terms trustor / listener / participant, and trustee / speaker may be used interchangeably.

1.1.3. Measuring trust propensity and perceived trustworthiness

Although there are a series of multi-disciplinary variations in past research aimed to capture the true essence of trust, it all boils down to two methods: (a) explicit measures of trust attitudes and behaviours through self-assessments using rating scales. These scales can be dichotomous (e.g. yes/no answers), probabilistic (i.e., ratings from 0% to 100%) or following a Likert scale format (Knack & Keefer, 1997; Rotter, 1967; Soroka, Helliwell, & Johnston, 2003); (b) implicit behavioural measures through the use of the prisoner's dilemma game and the trust game experiment (also known as the investment game) derived from behavioural

economics and games theory (Berg, Dickhaut, & McCabe, 1995; Deutsch, 1960). Explicit measures of trust have also become a standardised practise in assessing one's propensity to trust and perceived trustworthiness (Bauer & Freitag, 2018; Glaeser, Laibson, Scheinkman, & Soutter, 2000; H. H.-S. Kim, 2018; Naef & Schupp, 2009).

Previous behavioural and cognitive research, including studies on voice perception and production, has emphasized the significance of sample sizes and research environments. Samples of 24-36 participants per condition tend to reliably yield high agreement between participant ratings (Lavan, 2023; McAleer et al., 2014; Mileva, Tompkinson, Watt, & Burton, 2020), while both online and lab-based experiments have provided comparable data quality (Del Popolo Cristaldi, Granziol, Bariletti, & Mento, 2022; Germine et al., 2012; Honing & Reips, 2008; Uittenhove, Jeanneret, & Vergauwe, 2023).

1.1.4. Voice technology and the rise of intelligent agents

Humans naturally attribute social traits to others, including animals and even artificially intelligent entities (i.e., agents) like humanoid robots, virtual assistants, and chatbots. Consequently, research on human-agent interaction (HAI) emphasizes studying human behaviour for designing interactive intelligent agents (IAs), with voice playing a crucial role in attributing social traits, as seen in the "Computers as Social Actors" (CASA) paradigm (Lee & Nass, 2010; C. Nass, Steuer, & Tauber, 1994; Seaborn et al., 2021). The "uncanny valley" phenomenon further illustrates this, describing the uneasiness felt when an IA looks or sounds almost human but not quite (Mori, 1970; Mori, MacDorman, & Kageki, 2012).

Speech production in technological settings tends to refer to either canned speech (i.e., unchangeable pre-recorded speech samples) or synthesised speech, both seen in voice research (Cambre & Kulkarni, 2019; Clark et al., 2019; G. S. Kang & Heide, 1992; Kaur & Singh, 2023; C. I. Nass & Brave, 2005; Weinschenk & Barker, 2000). Past studies in HAI have revealed a positive relationship between perceptions of trustworthiness, rapport, learning and vocal entrainment (i.e., adapting one's vocal features to sound more similar to the person they are talking to; Cambre & Kulkarni, 2019). Further studies supporting the

effects of voice acoustics in IAs and trustworthiness have observed (1) a connection between vocal pitch and trustworthiness (Elkins & Derrick, 2013), (2) a preference towards more "natural" humanlike IA voices (Seaborn et al., 2021), and (3) the influence of the similarity-attraction effect. The similarity-attraction effect exhibits a preference and more positive attitudes towards speakers that are perceived to be more similar to the participant (Clark et al., 2019; Dahlbäck, Wang, Nass, & Alwin, 2007; C. Nass & Lee, 2000; C. I. Nass & Brave, 2005). For instance, Dahlbäck et al. (2007) observed a preference towards voice-based IAs that matched the listeners' own accent regardless of the IA's actual level of expertise, strengthening the case of people assigning human traits and predispositions to IAs.

Therefore, trustworthiness perceptions in voice-based IAs mirror those in human voices. Accordingly, trustors' dispositional, learned, and situational trust towards IAs, alongside IAs' perceived competence and ease of use should also be taken into account. Additional factors affecting trustworthiness attributions like perceived risk, especially regarding security, privacy, and transparency, also hold significance (Razin & Feigh, 2023), often examined through models such as the Technology Acceptance Model (TAM) and its variations (cf. Nam & Lyons, 2020; Riener, Jeon, & Alvarez, 2022).

Finally, trust propensity in HAI is often measured using scales like the Negative Attitudes to Robots (NARS; Jessup, Schneider, Alarcon, Ryan, & Capiola, 2019; Nam & Lyons, 2020). Overall, measurements of trustworthiness perceptions in HAI tend to follow the same methods laid out in the previous section with some alterations to match the technological aspect. For instance, sometimes a Wizard of Oz experiment is conducted for implicit measures, where during HAI the researcher either partly or fully operates the agent, while the participant is unaware, thinking the agent acts autonomously (Dahlbäck, Jönsson, & Ahrenberg, 1993; Riek, 2012).

1.1.5. Motivation

Given the above, this systematic review attempts to consolidate the existing multi-disciplinary literature on voice trustworthiness in both human and synthesised voices. Specifically, this

review aims to address the question of “how do acoustic features affect the perceived trustworthiness of a speaker?”, while also reviewing participant demographics, voice stimuli characteristics and task(s) involved.

1.2. Methods and Analysis

This systematic review followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) checklist (Page, McKenzie, et al., 2021; Page, Moher, et al., 2021). The search was performed on the 31st of October 2022, and all studies were initially identified by electronic search. Searches were repeated on the 18th January 2024 to identify any additional publications. A pre-registration protocol has been created for this review on Open Science Framework (`https://osf.io/v3kyg`) under the CC-by Attribution 4.0 International license.

This review adopted a narrative synthesis approach, to consolidate findings across studies investigating vocal trustworthiness in human speakers and voice-based IAs. The decision to use narrative synthesis was informed by the research objective, which focused on identifying and summarising acoustic features, demographic characteristics, and task paradigms across studies, rather than deriving effect sizes or pooled estimates. This approach allowed for a comprehensive examination and categorisation of findings into themes to identify trends, gaps, and contextual nuances in the literature, and inform future research directions.

1.2.1. Search strategy

Five bibliographic databases (Scopus, PsycInfo, ACM, ProQuest, PubMed) were searched using tailored search syntax detailed in Table 1.2, guided by the question: "How do acoustic features affect the perceived trustworthiness of a speaker?". Queries, developed collaboratively by all authors, have focused on English-language records published until January 18, 2024, using Boolean operators and wildcards for optimal search. Additional

records were identified through manual searches, citation chaining, and exploration of Scholar database, books, and conference proceedings.

Table 1.2: Search query syntax used in bibliographic databases.

Database	**Search query syntax**
Scopus	(TITLE-ABS-KEY (trust*) AND TITLE-ABS-KEY (voice OR vocal* OR prosod* OR speech OR acoustic* OR utter* OR speaker$ OR praat OR pitch OR "fundamental frequency" OR hnr OR "harmonic$-to-noise" OR "voice quality" OR accent*) AND TITLE-ABS-KEY (adult$))
PsycInfo	AB trust* AND AB (voice OR vocal* OR prosod* OR speech OR acoustic* OR utter* OR speaker OR praat OR pitch OR "fundamental frequency" OR HNR OR "harmonics-to-noise" OR "voice quality" OR accent*) AND AB adult
ACM	[Abstract: trust*] AND [[Abstract: voice] OR [Abstract: vocal*] OR [Abstract: prosod*] OR [Abstract: speech] OR [Abstract: acoustic*] OR [Abstract: utter*] OR [Abstract: speaker?] OR [Abstract: praat] OR [Abstract: pitch] OR [Abstract: "fundamental frequency"] OR [Abstract: hnr] OR [Abstract: "harmonic?-to-noise"] OR [Abstract: "voice quality"] OR [Abstract: accent*]]
ProQuest	summary(trust*) AND summary(voice OR vocal* OR prosod* OR speech OR acoustic* OR utter* OR speaker$ OR praat OR pitch OR "fundamental frequency" OR HNR OR "harmonic$-to-noise" OR "voice quality" OR accent*) AND summary(adult$)

Continued on next page

Table 1.2: Search query syntax used in bibliographic databases. (Continued)

Database	Search query syntax
PubMed	(trust*[Title/Abstract]) AND (voice[Title/Abstract] OR vocal*[Title/Abstract] OR prosod*[Title/Abstract] OR speech[Title/Abstract] OR acoustic*[Title/Abstract] OR utter*[Title/Abstract] OR speaker[Title/Abstract] OR praat[Title/Abstract] OR pitch[Title/Abstract] OR "fundamental frequency"[Title/Abstract] OR HNR[Title/Abstract] OR "harmonics-to-noise"[Title/Abstract] OR "voice quality"[Title/Abstract] OR accent* [Title/Abstract]) AND (adult[Title/Abstract])

1.2.2. Eligibility criteria for screening and selection of studies

Full-text papers have been obtained for titles and abstracts deemed relevant, based on specified inclusion and exclusion criteria. Papers were independently screened by CMP and SP, and any discrepancies were resolved through discussion.

Studies were included if: (a) participants were adults, irrespective of ethnicity, nationality, age and gender; (b) the study design involved a quantitative or mixed-methods approach; and (c) examined variables and reported outcomes focused on the acoustic characteristics of a speaker, with respect to their perceived trustworthiness.

Studies were excluded if: (a) reported outcomes did not focus on acoustic cues in relation to perceptions of trustworthiness of a human or IA; (b) characteristics of participants, stimuli and tasks involved could not be obtained; (c) the study design followed a qualitative-only approach; and (d) only the abstract was written in English, while the main paper was written in a language other than English.

1.2.3. Data extraction

Extracted information was divided into three categories accompanied by the publication's title and a reference key: (a) study characteristics, containing data such as the author, publication year, country that the study has taken place, number of participants, the aim of the study, vocal cues examined, task(s) involved, analyses and outcome; (b) listener characteristics, relating to the demographics of participants; (c) stimuli characteristics, including details of the stimulus itself and speaker demographics.

1.2.4. Risk-of-bias assessment method

The methodological quality and risk of bias of the included studies were assessed using a tailored scoring rubric adapted from Leung, Oates, and Chan (2018). The assessment evaluated risk of bias across five domains: conceptual clarity, reliability, internal validity, external validity, and reproducibility. Each domain covers specific criteria, scored from 0 to 2 points (0 = high risk of bias, 1 = moderate risk of bias, 2 = low risk of bias), detailed in the supplementary material. The maximum possible score for a study was 18 points (9 criteria × 2 points). The findings from the risk of bias assessment can be found in the Results section. Note that, such risk-of-bias scales do not necessarily reflect the quality of the evidence collected and used in the respective studies per se, or the reliability or quality of the studies involved more generally. Rather, they reflect "risk" in terms of how and what appears presented in the final publications, as filtered through the present authors' ability to extract these points from the respective manuscripts in the structured manner dictated by the scoring tool.

1.3. Results

1.3.1. Quantity of research available

Electronic and hand searches have identified 2,467 citations, of which 2,000 unique ones have been screened via Rayyan software (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016). Following elimination of duplicates, 81 potentially relevant citations remained. After full-text review and application of inclusion criteria, 57 citations have been excluded, resulting in 24 eligible studies (see Figure 1.1).

The 24 studies have been published between 2012 and 2024 and were conducted across Europe, America and Asia – nine in the UK, six in the US, two in Poland and one study each in France, Canada, China, Japan and Singapore, while two remain unclear (see Table 1.4). Eight of those are conference proceedings (Elkins, Derrick, Burgoon, & Nunamaker Jr., 2012; J. Kim, Gonzalez-Pumariega, Park, & Fussell, 2023; Klofstad, Anderson, & Peters, 2012; Lim et al., 2022; Maxim, Zalake, & Lok, 2023; Muralidharan, de Visser, & Parasuraman, 2014; Tolmeijer et al., 2021; Torre, White, & Goslin, 2016) and the remaining 16 are journal publications. Among them, fourteen studies have focused on perceived trustworthiness in terms of human speakers and the remaining 10 in terms of voice-based IAs. Twenty-one studies have focused on the effects of vocal pitch or pitch-related features with 12 of them incorporating the additional properties of pitch range, intonation, glide, formant dispersion, harmonic differences, HNR, jitter, shimmer, MFCCs, alpha ratio, loudness, pause duration and speech rate (see Table 1.4). Four studies solely focused on either speech duration or speaking rate.

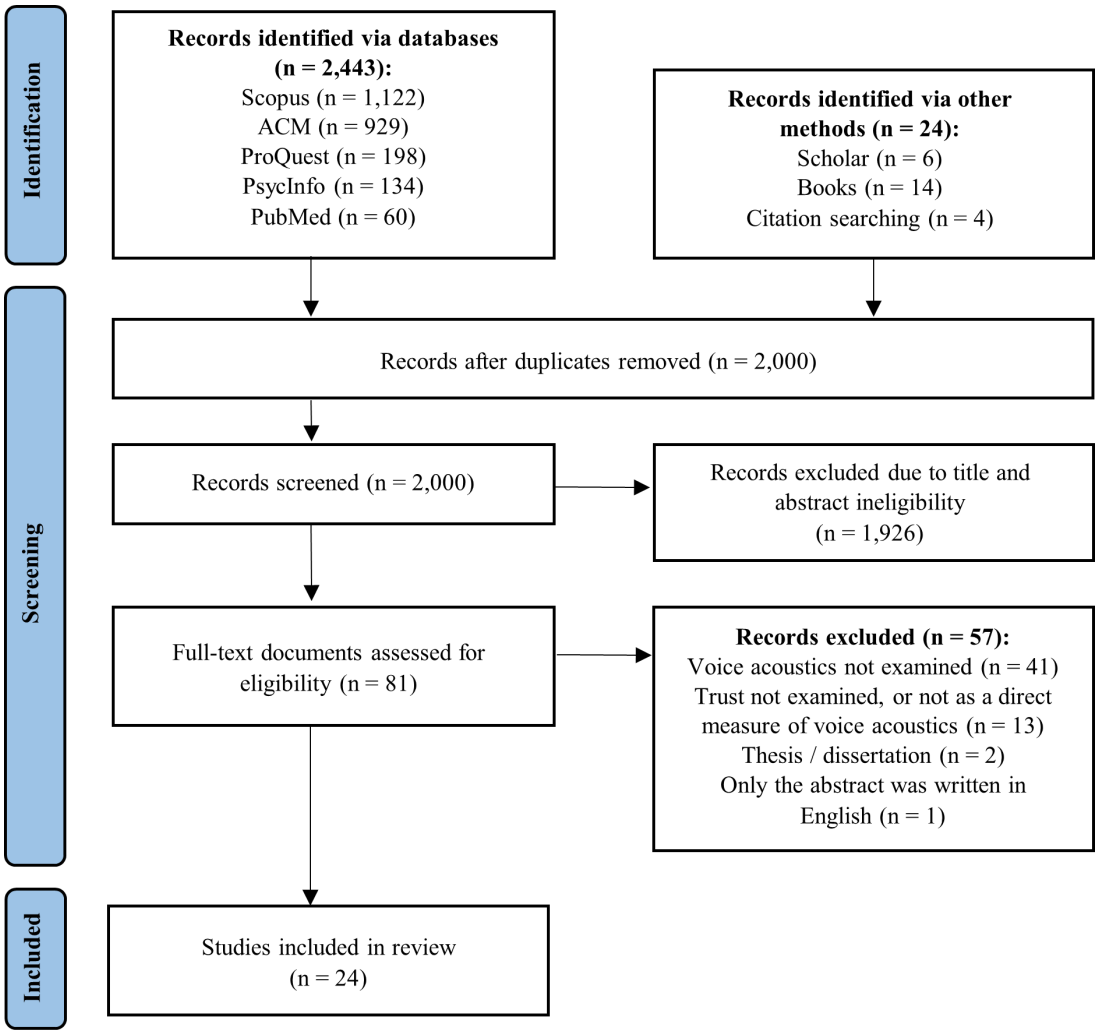

Figure 1.1: Identification of included studies in the systematic review, following the PRISMA flow diagram (Page, McKenzie, et al., 2021; Page, Moher, et al., 2021).

Most studies used Likert scales, typically in the rage of 1-7, to assess perceived trustworthiness (see Table 1.4). Some employed implicit decision tasks, while others combined explicit and implicit measures. Regression models, including linear mixed models and logistic regression, were common for exploring vocal acoustics and trustworthiness. Pearson's correlations assessed relationship strength. ANOVA, t-tests, and occasionally PCA or mixed methods were used for analysis.

Only one study examined age-group differences, i.e., adults older and younger than 60 years old (Schirmer, Chiu, Lo, Feng, & Penney, 2020). As seen in Table 1.5, eleven studies had fewer than 100 participants Deng et al. (2024); Elkins et al. (2012); Goodman and Mayhorn (2023); J. Kim et al. (2023); Mileva, Tompkinson, Watt, and Burton (2018); Mileva et al. (2020); Muralidharan et al. (2014); O'Connor and Barclay (2018); Oleszkiewicz et

al. (2017); Ponsot, Burred, Belin, and Aucouturier (2018); Schirmer et al. (2020), six of those with up to 50 (Goodman & Mayhorn, 2023; J. Kim et al., 2023; Mileva et al., 2018; Muralidharan et al., 2014; Oleszkiewicz et al., 2017; Ponsot et al., 2018). Ten studies had 100-550 participants (Baus et al., 2019; Belin, Boehme, & McAleer, 2019; Klofstad et al., 2012; Lim et al., 2022; Mahrholz et al., 2018; Maxim et al., 2023; McAleer et al., 2014; Tolmeijer et al., 2021; Torre, Goslin, & White, 2020; Yokoyama & Daibo, 2012), while one had over 2,000 (Groyecka-Bernard et al., 2022). For more information see Table 1.3 and Table 1.5. Most used audio-only stimuli, but 7 used audio-visual (Deng et al., 2024; Elkins et al., 2012; Lim et al., 2022; Maxim et al., 2023; Mileva et al., 2018, 2020; Yokoyama & Daibo, 2012). Five studies created over 100 usable stimuli (see Table 1.6; Groyecka-Bernard et al., 2022; Mahrholz et al., 2018; Ponsot et al., 2018; Schirmer et al., 2020; Torre et al., 2016).

Table 1.3: Descriptive statistics of the total sample size averaged between all included studies.

	Mean	Median	SD	Mode	Min	Max
Human speaker studies						
Listeners	346.3	181	625	85, 40	40	2,538
Speakers	42	25	51	64	1	208
Voice-based IA studies						
Listeners	108.2	86	69.3	None	30	234
Speakers	3	2	3.5	1	1	12

As indicated in the "Theme" column of Table 1.4, all 24 studies have been assigned a thematic (i.e., contextual) category based on shared situational attributes to provide more clarity and relevance during the discussion of their findings. Specifically, during the review stage, the situational factors of each study were examined. These factors were derived from either the study's inherent task (e.g. customer-barista interaction or fire warden simulation scenarios) or the meaning conveyed by the uttered stimuli (e.g. election speech,

or generic greeting). They played a key role in qualitatively grouping studies that shared similar situational contexts. For instance, the "public communication" theme has examined interactions involving public speaking in conferences (Yokoyama & Daibo, 2012), student elections (Mileva et al., 2020), or a political context (Klofstad et al., 2012; Schirmer et al., 2020). This iterative process was aimed to uncover consistent patterns and variations in how vocal acoustic features like pitch, amplitude, and intonation influence trustworthiness perceptions within specific, similar situational contexts.

Ultimately, seven distinct thematic categories were derived from this approach. These categories spanned a spectrum from generic first impressions, such as greetings and factual statements (Baus et al., 2019; Belin et al., 2019; Groyecka-Bernard et al., 2022; Mahrholz et al., 2018; McAleer et al., 2014; Mileva et al., 2018; Oleszkiewicz et al., 2017; Ponsot et al., 2018; Tsantani et al., 2016), to specific domains such as public communication (Klofstad et al., 2012; Mileva et al., 2020; Schirmer et al., 2020; Yokoyama & Daibo, 2012), social behaviour (O'Connor & Barclay, 2018), customer service (Lim et al., 2022; Muralidharan et al., 2014; Tolmeijer et al., 2021), financial services (Torre et al., 2020, 2016), telehealth advice (Goodman & Mayhorn, 2023; Maxim et al., 2023) and safety procedures (Deng et al., 2024; Elkins et al., 2012; J. Kim et al., 2023).

1.3.2. Risk-of-bias assessment findings

The total risk of bias scores for the 24 reviewed studies ranged from 8 to 16 out of a maximum of 18 points, with a mean, median and mode of 12 (SD = 2.5). Eight studies (33%) scored between 14 and 16 points, 12 studies (50%) scored between 9 and 13 points, and four studies (17%) scored 8 points (see Table 1.4).

Conceptual clarity was a consistent domain of weakness, with only six studies providing a clear and explicit definition of trust or trustworthiness (Deng et al., 2024; Elkins et al., 2012; Goodman & Mayhorn, 2023; J. Kim et al., 2023; Lim et al., 2022; Muralidharan et al., 2014). The majority relied on implicit or vague conceptualisations, potentially limiting the interpretability and comparability of findings across studies. Reliability demonstrated

notable variation, with only nine studies (38%) achieving the maximum score of 4 for using validated tools for measuring acoustic features and reporting intra- or inter-rater reliability (Baus et al., 2019; Elkins et al., 2012; Goodman & Mayhorn, 2023; Klofstad et al., 2012; Mahrholz et al., 2018; McAleer et al., 2014; Mileva et al., 2018, 2020; Schirmer et al., 2020).

Majority of studies scored highly on internal validity due to clear randomisation or pseudo-randomisation procedures, stimuli quality and justified sample sizes. External validity emerged as a widespread limitation, with only three studies (13%) scoring highly for diverse speaker and listener samples (Baus et al., 2019; Oleszkiewicz et al., 2017; Schirmer et al., 2020). Most studies were restricted to narrow demographic groups. Reproducibility was a strength, with 19 studies (75%) earning maximum scores due to detailed methodological descriptions.

Overall, the assessment highlighted strengths in the reproducibility domain and weaknesses in the domains of conceptual clarity and external validity. Greater attention to defining trust and trustworthiness, diversifying speakers and listeners, and improving methodological transparency is needed to strengthen the robustness and applicability of future research. For more information, see Tables 1.4 – 1.6, while the full scoring criteria and explanations for individual study scores are available in the supplementary material.

Table 1.4: Summary of all included studies

Study	Country	Study design			Vocal cues examined and outcome (i.e., more trustworthy when...)									Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
Studies: Perceived trustworthiness of human speakers.														
(Groyecka-Bernard et al., 2022)	Poland	Explicit 7-point	Generic	Regression	+	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Gender-irrelevant outcome.	14
(Schirmer et al., 2020)	Singapore	Explicit 7-point	Public comms	Regression	N/A	-	N/A	+	-	+	+	+	Younger and female speakers. Amplitude = intensity range.	16
(Mileva et al., 2020)	UK	Explicit 9-point	Public comms	Correlation	N/A	NSR	N/A	NSR	NSR	NSR	NSR	NSR	NSR for formant dispersion too.	12

Continued on next page

Table 1.4: Summary of all included studies (Continued)

Study	Country	Study design				Vocal cues examined and outcome (i.e., more trustworthy when...)								Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
(Baus et al., 2019)		Explicit 9-point	Generic	PCA Regression	N/A	NSR	NSR	N/A	+	NSR	NSR	N/A	NSR for formant dispersion, glide and alpha ratio too. + HNR for Scottish speakers only.	14
(Belin et al., 2019)	UK	Explicit	Generic	t-test	N/A	NSR	+	N/A	N/A	N/A	N/A	N/A	+ for intonation pattern of higher pitch at the start and end of an utterance and lower in the middle.	13
(Ponsot et al., 2018)	France	Explicit	Generic	Regression ANOVA	N/A	+	+	N/A	N/A	N/A	N/A	N/A	Gender-irrelevant outcome.	8
(Mahrholz et al., 2018)	UK	Explicit	Generic	Correlation Regression	+	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Stronger gender correlations (M > F).	15

Continued on next page

Table 1.4: Summary of all included studies (Continued)

Study	Country	Study design			Vocal cues examined and outcome (i.e., more trustworthy when...)									Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
(O'Connor & Barclay, 2018)	Canada	Explicit 2AFC & 7-point	Social behav.	ANOVA	- (P) + (A)	N/A	N/A	N/A	N/A	N/A	N/A	N/A	Male speakers only. P = prosocial, A = antisocial.	11
(Mileva et al., 2018)	UK	Explicit 9-point	Generic	ANOVA	N/A	NSR	N/A	N/A	N/A	N/A	N/A	N/A		14
(Oleszkiewicz et al., 2017)	Poland	Explicit 7-point	Generic	Regression	N/A	-	N/A	N/A	N/A	N/A	N/A	N/A	Gender-irrelevant outcome.	12
(Tsantani et al., 2016)	UK	Explicit 2AFC	Generic	t-test ANOVA	N/A	-	N/A	N/A	N/A	N/A	N/A	N/A	Gender-irrelevant outcome.	11
(McAleer et al., 2014)	UK	Explicit 9-point	Generic	PCA Regression	N/A	+ (M)	+ (F)	N/A	- both sexes	NSR	NSR	N/A	M / F = gender. - glide for females. NSR for formant dispersion & alpha ratio too.	13

Continued on next page

Table 1.4: Summary of all included studies (Continued)

Study	Country	Study design			Vocal cues examined and outcome (i.e., more trustworthy when...)									Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
(Klofstad et al., 2012)	US	Explicit 2AFC	Public comms	t-test	N/A	-	N/A	N/A	N/A	N/A	N/A	N/A	Gender-irrelevant outcome.	14
(Yokoyama & Daibo, 2012)	Japan	Explicit	Public comms	ANCOVA	N/A	N/A	N/A	N/A	N/A	N/A	N/A	+		8
Studies: Perceived trustworthiness of voice-based IAs.														
(Deng et al., 2024)	China	Mixed 7-point	Safety proced.	ANOVA Regression	NSR	-	-	NSR	N/A	N/A	N/A	NSR	Listener vocal response measured. NSR for formants. + MFCC.	13
(Maxim et al., 2023)	US	Explicit 7-point	Tele-health	ANOVA	N/A	NSR	N/A	-	N/A	N/A	N/A	-	Trend towards lower pitch.	12

Continued on next page

Table 1.4: Summary of all included studies (Continued)

Study	Country	Study design				Vocal cues examined and outcome (i.e., more trustworthy when...)								Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
(J. Kim et al., 2023)	US	Mixed 7-point	Safety proced.	MANCOVA	N/A	+	+	N/A	N/A	N/A	N/A	+		8
(Goodman & Mayhorn, 2023)	US	Mixed 7-point	Tele-health	Correlation t-test	N/A	NSR	N/A	N/A	N/A	N/A	N/A	N/A	Female IA preference.	16
(Lim et al., 2022)	UK	Explicit 7-point	Cust. service	Binomial Correlation Qualitative	N/A	+	N/A	+	N/A	N/A	N/A	+	Trust-propensity was measured too.	12
(Tolmeijer et al., 2021)	US	Mixed 5-point	Cust. service	Non-parametric t-test and ANOVA	N/A	NSR	N/A	N/A	N/A	N/A	N/A	N/A		8
(Torre et al., 2020)	UK	Mixed 7-point	Finance services	Regression	N/A	+	N/A	N/A	N/A	N/A	N/A	N/A		12

Continued on next page

Table 1.4: Summary of all included studies (Continued)

Study	Country	Study design				Vocal cues examined and outcome (i.e., more trustworthy when...)								Risk score
		Task	Theme	Analyses	Duration	Pitch	Intonation pattern	Amp.	HNR	Jitter	Shimmer	Speech rate	Additional notes	
(Torre et al., 2016)	UK	Implicit	Finance services	Regression	N/A	+	N/A	N/A	N/A	N/A	N/A	+	Speech rate = articulation rate.	12
(Muralidharan et al., 2014)		Mixed	Cust. service	ANOVA	N/A	+	N/A	N/A	N/A	N/A	N/A	N/A	Significance for lower time delay (flanging). Pitch = pitch range.	9
(Elkins et al., 2012)	US	Implicit	Safety proced.	Regression	N/A	+	N/A	N/A	N/A	N/A	N/A	N/A	Listener vocal response measured. + response time. Less prominent effects over time.	15

Risk score: Higher scores denote lower "risk" (out of a maximum possible score of 18 points) — see the relevant Methods section for an explanation of what this measures.

Note 1: NSR, no statistical-significance reported; +/−, higher/lower.

Note 2: Green cells highlight statistically significant correlations and associated directionality (+/−) between acoustic features and perceived trustworthiness, whereas red cells highlight non-significant or inconclusive results (NSR).

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
Studies: Perceived trustworthiness of human speakers.				
(Groyecka-Bernard et al., 2022)	2,538	46% males; 54% females.	32.51	N/A
(Schirmer et al., 2020)	80	25% younger males; 25% younger females; 25% older males; 25% older females.	23.7 [20 – 32 years] (younger males); 21.1 [19 - 27 years] (younger females); 67.9 [60 - 91 years] (older males); 68 [60 - 77 years] (older females).	Older adults: 2 with normal hearing (<= 25 dB); 28 with slight hearing impairment (26 – 40 dB); 9 with moderate impairment (41 – 60 dB); 1 with severe impairment (61 – 80 dB) that was corrected with a hearing aid.
(Mileva et al., 2020)	99	7% males; 93% females.	19 [18 – 50 years]	N/A

Continued on next page

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses (Continued)

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
(Baus et al., 2019)	279 (study 1); 258 (study 2).	33% males (S1); 67% females (S1); 50% males (S2); 50% females (S2).	20.2 (S1); 22.03 (S2).	Spanish nationality.
(Belin et al., 2019)	500	29% males; 71% females.	Median age = 24 [19 - 65 years]	N/A
(Ponsot et al., 2018)	44 (study 1; trust task = 23); 40 (study 2; trust task = 19).	52% males (S1 trust task); 48% females (S1 trust task); 47% males (S2 trust task); 53% females (S2 trust task).	22 (S1 trust task); 21 (S2 trust task).	N/A
(Mahrholz et al., 2018)	181	24% males; 76% females.	21.3 [18 – 27 years] (males); 20.1 [18 – 30 years] (females).	Scottish nationality.
(O'Connor & Barclay, 2018)	85 (study 1); 63 (study 2).	100% females (S1 & S2)	18.21 (S1); 18.9 (S2).	N/A

Continued on next page

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses (Continued)

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
(Mileva et al., 2018)	40	20% males; 80% females.	20.1 [18 - 30 years]	N/A
(Oleszkiewicz et al., 2017)	50	20% blind males; 34% blind females; 16% sighted males; 30% sighted females.	37.9 [24 – 64 years] (healthy blind adults); 38.7 [24 – 65 years] (sighted adults).	N/A
(Tsantani et al., 2016)	40 (study 1); 240 (study 2).	33% males (S1); 67% females (S1); 24% males (S2); 76% females (S2).	24 (S1); 20 (S2).	N/A
(McAleer et al., 2014)	320	37% males; 63% females.	28.5	N/A
(Klofstad et al., 2012)	210	50% males; 50% females.	Undergraduate students.	N/A

Continued on next page

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses (Continued)

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
(Yokoyama & Daibo, 2012)	466	53% males; 47% females.	19.6	N/A
Studies: Perceived trustworthiness of voice-based IAs.				
(Deng et al., 2024)	75	23% males (group 1 & 2); 25% females (group 1); 29% females (group 2).	22.69 [19 – 27 years] (group 1); 22.15 [19 – 26 years] (group 2).	N/A
(Maxim et al., 2023)	165	56% males; 43% females; 1% non-binary.	43.35 [24 – 68 years]	144 white; 9 Asian; 5 black; 7 mixed-race.
(J. Kim et al., 2023)	30	50% males; 50% females.	21 [18 – 38 years]	N/A

Continued on next page

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses (Continued)

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
(Goodman & Mayhorn, 2023)	47	55% males; 38% females; 2% non-binary; 5% undisclosed.	19.5	N/A
(Lim et al., 2022)	202	60% males; 38% females; 2% non-binary.	28.11 [18 – 60 years]	N/A
(Tolmeijer et al., 2021)	234	41% males.	33 [19 – 74 years]	US nationality.
(Torre et al., 2020)	108	22% males; 78% females.	19 [18 – 48 years]	British nationality.
(Torre et al., 2016)	83	38% males; 62% females.	Median age = 21 [18 - 67 years]	British nationality = 5 from Wales and the rest from across England.
(Muralidharan et al., 2014)	50 (study 1); 23 (study 2).	39% males (S2); 61% females (S2).	[18 - 28 years]	N/A

Continued on next page

Table 1.5: Participant characteristics of all included studies. The "adjusted sample size" column notes the total number of participants after having excluded any individuals from the analyses (Continued)

Study	Adjusted sample size	Gender	Mean age [range]	Additional participant details
(Elkins et al., 2012)	88	60% males; 40% females.	25.45	N/A

Table 1.6: Stimuli characteristics of all included studies

Study	Stimuli	Speaker demographics
Studies: Perceived trustworthiness of human speakers.		
(Groyecka-Bernard et al., 2022)	1,248 audio-only stimuli; 60 Polish-language WAV files per listener; Sampling rate = 96 kHz; Resolution = 16-bit.	208 Polish speakers; 52% males, 48% females; Mean age = 32.83.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Schirmer et al., 2020)	520 audio-only stimuli; 2 sentences x 13 expressions x 20 speakers.	20 Singaporean native English speakers with acting experience; Younger adults: 25% males, 25% females; Mean age = 23.8 (males), 22.2 (females); Older adults: 25% males, 25% females; Mean age = 63 (males), 69.2 (females).
(Mileva et al., 2020)	22 audio-visual stimuli; 7 stimuli from females; Mean duration = 3.41 seconds.	22 speakers; 32% females.
(Baus et al., 2019)	Audio-only stimuli; Study 1: 64 Spanish recordings of the word "Hola"; mean duration: males = 319 ms; females = 338 ms; normalised; Study 2: 64 recordings, re-used from McAleer et al. (2014).	Study 1: 64 Spanish; 50% males; Mean age = 22.1; Study 2: 64 Scottish voices, re-used from McAleer et al. (2014).

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Belin et al., 2019)	Audio-only stimuli; Re-synthesised and manipulated pre-existing Scottish voice stimuli of the word "hello" from McAleer et al. (2014); Split between low and high trustworthiness as per the rating results obtained by McAleer et al. (2014).	Subset of Scottish male and female voices, re-used from McAleer et al. (2014).
(Ponsot et al., 2018)	Audio-only stimuli; Study 1: ∼700 trials x 2 genders of the French word "bonjour"; Study 2: 420 stimuli (20 French words, including "bonjour" x 7 pitch contour filters x 3 repetitions); For all stimuli: Sampling rate = 44.1 kHz; Resolution = 16-bit mono; Normalisation range = 75 - 80 dB.	Study 1: 2 French speakers; 1 male (aged 28); 1 female (aged 29); Study 2: 12 French speakers; 50% males, 50% females; Mean age = 33.33 [21 - 57 years].

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Mahrholz et al., 2018)	120 audio-only stimuli; Lab-based WAV recordings; 2 durations (word/sentence) x 2 contexts (with/without context); Sampling rate = 44.1 kHz; Resolution = 16-bit mono; Normalised; Average duration: males = 411.1 - 3,019.6 ms; females = 394.6 - 3,172.8 ms.	60 Scottish; 50% males, 50% females; Mean age = 23.2 (males), 20.2 (females).
(O'Connor & Barclay, 2018)	Audio-only stimuli; Paired words x 2 contexts (prosocial/antisocial) x 2 genders (feminised = higher pitch; masculinised = lower pitch).	4 speakers; 100% males; Mean age = 18.
(Mileva et al., 2018)	40 audio-visual stimuli; 2 genders x 2 pitch conditions (higher/lower); males: Higher-pitch = 140 Hz, lower-pitch = 90 Hz; females: higher-pitch = 250 Hz, lower-pitch = 170 Hz.	20 speakers; 50% males; Mean age = 23.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Oleszkiewicz et al., 2017)	Audio-only stimuli; WAV format with higher/lower pitch manipulation; Sampling rate = 96 kHz; Resolution = 32-bit; Normalisation = 70 dB.	8 speakers; 50% males, 50% females.
(Tsantani et al., 2016)	66 audio-only stimuli per study, Re-used from McAleer et al. (2014); 2 pitch conditions (higher/lower, 20 Hz shift) x 2 contexts/studies (backward/forward speech manipulation); Average duration = 400 ms.	33 Scottish voices, re-used from McAleer et al. (2014); 55% males, 45% females.
(McAleer et al., 2014)	64 audio-only stimuli; WAV format with neutral tone of voice of the word "Hello"; Sampling rate = 44.1 kHz; Resolution = 16-bit mono; Average duration = 319 ms (males), 390 ms (females).	64 Scottish; 50% males; Mean age = 28.2.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Klofstad et al., 2012)	54 audio-only stimuli; 2 genders x 2 pitch conditions (higher / lower); Sampling rate = 44.1 kHz; Amplitude normalised; Mean pitch = 187 Hz females, 107 Hz males.	27 speakers; 37% males, 63% females;Mean age = 33 [20 – 55 years] (males), 31 [21 – 60 years] (females).
(Yokoyama & Daibo, 2012)	4 audio-visual stimuli; 2 gaze states (high = 8% looking at the camera; low = 83%) x 2 speech rates (faster = 510 syllables per minute; slower = 330).	1 Janapanese, female speaker; 23 years old.
Studies: Perceived trustworthiness of voice-based IAs.		
(Deng et al., 2024)	Audio-visual stimuli. Participant responses were recorded and stored for speech analysis in relation to perceived trustworthiness in HAI.	Automated-vehicle system with audio-visual interaction features and voice recognition features.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Maxim et al., 2023)	2 audio-visual stimuli; 1 agent x 1 scenario x 2 voice characteristics (1 extroverted and 1 introverted); Extroverted agent: Speech rate = 216 words per minute; Base pitch = 140 Hz; Introverted agent: Speech rate = 184 words per minute; Base pitch = 84 Hz; Volume = 15% less (-1.41 dB) than the extroverted voice.	A female embodied conversational agent.
(J. Kim et al., 2023)	2 audio-only stimuli; "Urgent" vs "calm" voice; Urgent voice = faster speech rate, higher pitch, variable intonation; Calm voice = slow speech rate, static intonation.	Recorded human voices.
(Goodman & Mayhorn, 2023)	6 audio-only stimuli; 3 x pitch conditions (high/intermediate/low).	2 synthesised voices; 1 male, 1 female.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Lim et al., 2022)	2 audio-visual stimuli; 2 x personalities (extroversion = higher pitch, speech rate, volume; introversion = lower pitch, speech rate, volume).	An embodied conversational agent.
(Tolmeijer et al., 2021)	5 audio-only stimuli; 2 genders x 2 pitch conditions (higher/lower); 1 gender ambiguous voice = pitch shifted towards the average of high-pitch female and low-pitch male voices.	A voice assistant using a US accent; 1 male, 1 female and 1 gender-ambiguous voice.
(Torre et al., 2020)	40 audio-only stimuli; 2 intents (neutral/amused); Sentence length = 16.6 syllables.	4 British females in their 20s; Birmingham accent = 50% speakers; SSBE accent = 50% speakers.
(Torre et al., 2016)	240 audio-only stimuli; 4 blocks of 20 sentences per speaker; Mean number of syllables per sentence = 16.95.	12 British females in their 20s; Plymouth accent = 25% speakers; Birmingham accent = 25% speakers; London accent = 25% speakers; SSBE accent = 25% speakers.

Continued on next page

Table 1.6: Stimuli characteristics of all included studies (Continued)

Study	Stimuli	Speaker demographics
(Muralidharan et al., 2014)	Audio-only stimuli; 5x pitch range conditions = 525 Hz (humanlike), 395 Hz, 195 Hz, 125 Hz, 1 Hz (machine-like).	2 synthesised voices; 1 male, 1 female.
(Elkins et al., 2012)	Audio-visual stimuli; 4 questions x 2 genders x 2 demeanors (neutral / smiling). Participant responses were recorded and stored for f_0 analysis, resulting to a total of 866 WAV files with a final sampling rate of 11.025 kHz.	1 embodied conversational agent, portraying both male and female audio-visual aspects independently.

1.4. Discussion

In this review, vocal pitch has emerged as a predominant focus across all 24 included studies, followed by investigations into amplitude, intonation, HNR, jitter, shimmer, speech duration, and/or speech rate. To facilitate a comprehensive discussion, findings have been categorised into sections on human speakers and voice-based IAs, grouping relevant studies accordingly.

The interpretation of study outcomes has been significantly shaped by contextual factors, leading to the qualitative grouping of studies into thematic (i.e., contextual) categories. Each thematic category summarises findings on acoustic features and their implications for perceptions of trustworthiness within specific contexts or situations, as detailed further in the discussion. For instance, studies within the "telehealth advice" theme have examined trustworthy voice acoustics in scenarios involving medication guidance and mental wellness practices. This thematic approach has facilitated the identification of consistent patterns and variations in how vocal acoustic features contribute to communication dynamics and shape perceptions of trustworthiness within specific contexts. Without these situational considerations, the overall findings across studies seemed to be inconclusive.

In total, seven contextual themes have been identified (also see Table 1.4). Three of these themes are evident in human speaker studies: "generic first impressions" (e.g. from greetings to factual statements), "public communication", and "social behaviour". The remaining four themes are identified in voice-based IA studies: "customer service", "financial services", "telehealth advice", and "safety procedures". For a summary of findings see Table 1.7.

1.4.1. The role of acoustic cues in the perceived trustworthiness of human speakers

Thirteen of the 24 studies have focused on perceived trustworthiness of adult human voices. Six have solely assessed pitch-related measures (Belin et al., 2019; Mileva et al., 2018;

O'Connor & Barclay, 2018; Oleszkiewicz et al., 2017; Ponsot et al., 2018; Tsantani et al., 2016), four have combined pitch with HNR, jitter, shimmer, loudness, formant dispersion, or speech rate (Baus et al., 2019; McAleer et al., 2014; Mileva et al., 2020; Schirmer et al., 2020), two have focused solely on speech duration (Groyecka-Bernard et al., 2022; Mahrholz et al., 2018), and one on speaking rate (Yokoyama & Daibo, 2012).

All studies have used explicit measures like rating scales, with 7-point (Groyecka-Bernard et al., 2022; O'Connor & Barclay, 2018; Oleszkiewicz et al., 2017; Schirmer et al., 2020) and 9-point (Baus et al., 2019; McAleer et al., 2014; Mileva et al., 2018, 2020) Likert scales being common. Analyses have included correlational, inferential, and regression models (details in Table 1.4). While some studies have linked trustworthiness to lower or higher pitch independent of gender, others have noted gender's influence. Building on the premise of situational factors, the following part of this subsection presents a discussion on study findings, categorised thematically according to contextual similarities.

1.4.1.1. "Generic first impressions" theme

Nine of the studies on human voice trustworthiness have focused on generic first impression scenarios, using a variety of audio stimuli (e.g. greetings such as the word "hello", or snippets from The Rainbow Passage; Fairbanks, 1960). The main aspects that have been studied under this theme include pitch and related features like intonation and glide (Baus et al., 2019; McAleer et al., 2014; Oleszkiewicz et al., 2017), and some have also considered voice quality features (Baus et al., 2019; Belin et al., 2019; McAleer et al., 2014; Mileva et al., 2018; Ponsot et al., 2018; Tsantani et al., 2016). Two studies specifically, have only analysed speech duration (e.g. comparison between shorter and longer sentences or words; Groyecka-Bernard et al., 2022; Mahrholz et al., 2018).

Vocal pitch and related features: Current findings have primarily suggested a positive link between pitch, rising intonation at both ends of a stimulus and trustworthiness attributions in English-speaking contexts (Belin et al., 2019; McAleer et al., 2014). Nevertheless, cultural differences seem to be prevalent, as mixed findings for pitch have been identified for non-English speaking studies (Baus et al., 2019; Oleszkiewicz et al., 2017; Ponsot et

al., 2018). Multimodal research (i.e., faces and voices) has also yielded inconclusive results regarding pitch's impact, noting that there may be a stronger influence of faces in such cases (Mileva et al., 2018). Moreover, methodological differences seem to have played a role in the current findings: English-speaking studies using Likert scales have favoured higher pitch for trustworthiness, whereas research utilising a 2AFC task (Tsantani et al., 2016) has deemed lower pitch as more trustworthy. Further research comparing these methodologies is necessary for a clearer understanding.

Voice quality features: Significant findings have centered on HNR, revealing cultural disparities based on English-speaking stimuli: native listeners seem to favour lower HNR for trustworthiness (McAleer et al., 2014), whereas non-native listeners seem to prefer higher HNR (Baus et al., 2019), regardless of the speaker's gender. Voice quality features tend to be sensitive in respect to voice quality pathologies and physiological changes that occur in aging (Farrús et al., 2007; Felippe et al., 2006; Ferrand, 2002; Jalali-najafabadi et al., 2021; Rojas et al., 2020), which may account for these preferences. For instance, native listeners may gravitate more towards youthful-sounding voices, which may promote more positive or upbeat impressions. In contrast, non-native listeners, may prioritise vocal clarity and precision in foreign speech that usually comes with a higher HNR. Considering that cross-cultural vocal trustworthiness studies seem to be scarce, further investigations are warranted for a more comprehensive understanding.

Temporal features: Both studies examining speech duration have indicated that longer stimuli, around 2-3 seconds, tend to be perceived as more trustworthy than shorter ones, e.g. a vowel or a word (Groyecka-Bernard et al., 2022; Mahrholz et al., 2018). However, Mahrholz et al. (2018) has added that even stimuli as short as 0.5 seconds can convey trustworthiness, consistent with previous research (Lavan, 2023; McAleer et al., 2014). Moreover, these perceptions appear to be consistent across cultures, such as Polish (Groyecka-Bernard et al., 2022) and Scottish (Mahrholz et al., 2018) speakers. A potential explanation for these findings may relate to longer speech duration potentially allowing for more thorough processing, thus influencing trust perceptions, as well as introducing more opportunities for response variability among listeners (Groyecka-Bernard et al., 2022). Having said that,

further cross-cultural studies are still needed for definitive conclusions.

1.4.1.2. "Public communication" theme

Four studies seem to fall under this theme category, which either tackle trustworthiness judgments in terms of public speaking in conferences (Yokoyama & Daibo, 2012) and student elections (Mileva et al., 2020), or in terms of stimuli with a political context (Klofstad et al., 2012; Schirmer et al., 2020).

Temporal features: Yokoyama and Daibo (2012) has assessed trustworthiness perceptions based solely on the speech rate of a female speaker in Japan, finding a preference for faster speech. Another study, despite using Singaporean English speakers and listeners, has reached similar conclusions (Schirmer et al., 2020). In support of these findings, past research, including the "effort code" theory, suggest that faster speech rates tend to convey greater knowledge and expertise (Gussenhoven, 2002; Rodero et al., 2014; S. M. Smith & Shaffer, 1995). Consequently, boosting speakers' perceived confidence, credibility, and persuasiveness, particularly in public speaking contexts. Additionally, these findings may also be indicative of listeners' preference towards younger speakers, considering that slower speech rate tends to be more associated with aging (Schirmer et al., 2020).

Voice quality features: The aforementioned Singaporean study (Schirmer et al., 2020) has also shown a preference for voices with lower pitch and HNR, but higher jitter, shimmer, and intensity range. This is the only study that has explicitly explored age differences, revealing a preference for younger speakers and a general preference for female speakers across ages. The contradictory lower HNR, higher jitter and shimmer preferences though, may stem from perceived expressiveness or individual and cultural influences on vocal aesthetic preferences. Conversely, a UK study under this theme (Mileva et al., 2020) has yielded inconclusive results, potentially due to their multimodal design (faces and voices). Their multimodality makes it more difficult for a direct comparison with the previous, unimodal (i.e., voice-only) studies, and to interpret their findings.

Vocal pitch and related features: Lastly, two studies (Klofstad et al., 2012; Schirmer et al., 2020) have exhibited a preference for lower-pitched voices regardless of gender, which

may potentially be influenced by individual and cultural norms of vocal aesthetic appeal. An alternative interpretation for lower-pitched female voices may be that they sound more dominant and thus, perceived as more authoritative, confident, and competent (Klofstad et al., 2012; Ohala, 1983).

1.4.1.3. "Social behaviour" theme

Vocal pitch and related features: The only study under this theme has explored male voices in pro-social and anti-social scenarios (O'Connor & Barclay, 2018). Lower-pitched voices have been noted as more trustworthy in positive contexts and higher-pitched voices in negative contexts. These observations were partly explained in terms of higher pitch potentially mitigating the perceived intimidation of antisocial behaviour in men (O'Connor & Barclay, 2018). This seems to align with the "frequency code" theory, where higher-pitched voices tend to signal smaller body sizes, primarily seen in women and children; thus potentially conveying a friendlier or less threatening demeanour (Ohala, 1983, 1995).

Altogether, vocal cues in human voices seem to play a significant role in trustworthiness attributions, albeit influenced by contextual factors. It is further suggested that vocal cues may have stronger effects when voice acts as the sole or primary modality for drawing trustworthiness inferences.

1.4.2. The role of acoustic cues in the perceived trustworthiness of voice-based IAs

The remaining 11 studies in this review focused on assessing the perceived trustworthiness of voice-based Intelligent Agents (IAs), whether using synthesized or pre-recorded human voices. Similar to human speakers, voice-based IAs are often evaluated with human behaviour in mind, with context also playing a significant role. Contextual themes and associated acoustic features for trustworthy speech are discussed further.

1.4.2.1. "Customer service" theme

Three voice-based IA studies examining trustworthiness attributions fall under this theme category. Contexts vary from barista scenarios (Lim et al., 2022) to task-assistance scenarios (Muralidharan et al., 2014; Tolmeijer et al., 2021).

Vocal pitch and related features: Findings on pitch have been inconclusive, which may partly stem from differences in study designs; one study used audio-visual stimuli with correlational analyses (Lim et al., 2022), while the other two employed audio-only stimuli with inferential models (Muralidharan et al., 2014; Tolmeijer et al., 2021). Tolmeijer et al. (2021) has also focused extensively on gender-stereotyping, manipulating synthetic voices to sound more masculine, feminine, or gender-ambiguous. The lack of pitch significance in trustworthiness perceptions in these studies, suggests that listeners may not rely solely on pitch for voice-based IAs in assistive roles. These findings challenge the importance of vocal pitch in shaping trustworthiness perceptions of IAs.

Vocal pitch in combination with other acoustic features: Past research (Muralidharan et al., 2014) has suggested that combining pitch and flanging (i.e., speech time delay manipulation) influences trustworthiness perceptions. They have found that a lower pitch range with greater time delay tends to be perceived as more machine-like and less trustworthy compared to natural human speech. They added that human speech typically has a natural time delay of about 0.01 seconds, and increasing this delay can make it sound less natural. This deviation, along with a less animated voice, may lead to uneasiness in listeners, supporting theories on social inferences from HAI (Mori, 1970; Mori et al., 2012; Muralidharan et al., 2014; C. Nass et al., 1994).

Furthermore, a louder voice with a faster speech rate and higher pitch tends to be perceived as more trustworthy, supporting theories linking trust formation with positive traits (Lim et al., 2022). Faster speech rate tends to portray speakers' deeper understanding and passion for the subject. In combination with higher pitch it is usually associated with extroversion and openness (Lim et al., 2022; Maxim et al., 2023; Ohala, 1995), further portraying speakers as competent, persuasive, and credible (Gussenhoven, 2002; Rodero et al., 2014; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012). Only one study

has examined listeners' trust propensity, revealing positive and negative associations with trustworthiness attributions dependent on the scales used (Lim et al., 2022). Overall findings under this theme seem to be appropriate if they are interpreted in terms of listeners being more accepting and trusting of speakers' assistance on a task. Nonetheless, more extensive research is needed in this area before these findings can be deemed as generalisable.

1.4.2.2. "Financial services" theme

Both studies (Torre et al., 2020, 2016) in this theme employed implicit investment tasks, with one also using a 7-point Likert scale (Torre et al., 2020). Both have assessed female-only voices with various British accents and used regression models for analysis.

Vocal pitch in combination with other acoustic features: Findings have indicated that higher pitch and faster articulation rate seem to be associated with more trustworthiness. Additionally, they have linked higher pitch to positive emotions such as happiness. These findings seem to align with past research linking greater articulatory effort to higher perceptions of knowledge, confidence, and helpfulness (Gussenhoven, 2002). The preference for higher-pitched voices in female IAs strengthens the case of attributing human traits to IAs, as women typically have higher-pitched voices due to physiological factors. Past research has also exhibited a preference for higher-pitched women, linking them with positive traits like attractiveness and trustworthiness (Lavan, 2023; McAleer et al., 2014). The current findings may also strengthen the case for humans assigning gender roles to assistive occupations, even in HAI (Tolmeijer et al., 2021).

1.4.2.3. "Telehealth advice" theme

Two studies have explored trustworthiness judgments in receiving advice for medication (Goodman & Mayhorn, 2023) and mental wellness (Maxim et al., 2023) contexts.

Vocal pitch in combination with other acoustic features: While one has focused on vocal pitch of male and female IA using audio-only, the other has examined pitch, speech rate, and loudness of a female IA with audio-visual stimulus. Despite no reported acoustic significance for trustworthiness, a trend towards lower pitch, speech rate, and volume in

female voices is observed. Additionally, extroverted listeners have offered higher ratings overall, irrespective of speakers' perceived traits (Maxim et al., 2023).

Authors seem to have attributed these observations to voice similarity with mental health professionals, suggesting softer, empathetic, and confident perceptions (Maxim et al., 2023). Moreover, slower speech rate and lower volume, which are often associated with physiological changes occurring in aging (Baus et al., 2019; Ferrand, 2002; Heffernan, 2004; Lavan et al., 2019; McAleer et al., 2014; Rojas et al., 2020). As such, speakers may have also been perceived as older and probably more knowledge. These findings further highlight HAI drawing inferences from human-human interactions and linking trustworthiness to positive traits. Nonetheless, limited stimuli and differing methodologies between the two studies may affect their generalisability. For instance, Maxim et al. (2023) examined the similarity-attraction effect among other aspects and employed a multi-modal design (i.e., faces and voices), which makes it more difficult for a direct comparison with the second, unimodal (i.e., voice-only) study (Goodman & Mayhorn, 2023), and to interpret their findings.

1.4.2.4. "Safety procedures" theme

The last three studies on voice-based IAs explored attributions of trustworthiness employing scenarios such as security screening (Elkins et al., 2012), fire warden simulation (J. Kim et al., 2023) and voice assistance during driving simulation (Deng et al., 2024).

Vocal pitch in combination with other acoustic features: All three studies have associated higher vocal pitch with increased trustworthiness in voice-based IAs, albeit varying in their methodology. Two of them have assessed trustworthiness through participants' verbal responses during HAI (Deng et al., 2024; Elkins et al., 2012). They have reported that higher-pitched responses with greater pitch and MFCC variability, higher intensity, and longer response time may correspond to higher trustworthiness ratings. These findings may relate to participants developing more positive perceptions of the IA, in terms of dominance, authoritativeness and competence, and feeling more invested during HAIs as per the "effort code" theory (Gussenhoven, 2002; Klofstad et al., 2012; Ohala, 1983). However, these effects seem to diminish with prolonged HAI, possibly due to the accumulation of

information and the opportunity to make further inferences over time (Elkins et al., 2012). While these studies provide valuable insights, pre-assessing participants' trust propensity and personality traits could enhance conclusions. The final study (J. Kim et al., 2023), which examined the acoustics of voice-based IAs instead, has similarly reported that higher pitch with faster speech rate and variable intonation has prompted higher trustworthiness ratings, labelling that combination of acoustics as an "urgent voice".

Granted that these three studies have offered limited stimuli, which like previously mentioned, might not be sufficient to draw generalised conclusions to the broader population. Nevertheless, despite methodological variances, all of them have consistently reported similar results. This consistency may be attributed to the heightened vocal urgency observed in speakers during emergency situations, which could also be perceived as more authoritative, eager to assist, and concerned with everyone's safety (Gussenhoven, 2002; Rodero et al., 2014; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012).

All things considered, vocal cues of voice-based IAs seem to be playing a significant role in attributions of trustworthiness. However, contextual and situational factors are equally prevalent in this section as in research on human voices, enhancing the interpretability of findings. It is further highlighted the influence of human-human interactions and social inferences from human behaviour when studying HAIs. Finally, majority of the HAI studies had less than a hundred participants (Deng et al., 2024; Elkins et al., 2012; Goodman & Mayhorn, 2023; J. Kim et al., 2023; Muralidharan et al., 2014; Torre et al., 2016), and only one study had more than 5 speakers (Torre et al., 2016) making their findings potentially more difficult to generalise to the wider population, even though they were reported to be well-powered.

Table 1.7: Summary of trust-related acoustic features in human and IA studies: Actionable insights for practitioners and recommendations for future research

Theme	Trustworthy acoustic features	Limitations	Recommendations and insights
Studies: Perceived trustworthiness of human speakers.			
Generic first impressions	**Pitch:** In English contexts, a higher pitch or rising intonation often seems to boost trustworthiness perceptions, albeit mixed findings in non-English settings. **Voice quality:** Native English listeners often favour lower HNR, while non-native listeners may prefer higher HNR for vocal clarity or precise enunciation. **Speech duration:** Longer segments (∼2–3s) allow more processing time, enhancing trustworthiness.	Primarily English-speaking samples; limited cross-cultural research. Some multimodal designs (face + voice) complicate pure acoustic findings. Conflicting pitch results can arise from different task types (Likert vs forced-choice).	**For researchers:** Compare short vs long utterances in diverse languages and speaker demographics. **For practitioners (e.g., marketers, voice coaches):** In English contexts, use slightly longer greetings plus moderate/higher pitch for a friendly first impression, checking cultural fit in non-English contexts.
Public communication	**Pitch:** A lower pitch can convey authority or dominance in both male and female speakers, depending on cultural norms. **Voice quality:** Younger or more expressive voices (e.g., increased jitter/shimmer) can be favoured, but cultural preferences vary. **Speech rate:** A faster rate suggests competence/expertise ("effort code" theory).	Highly varied contexts (political speeches, conferences, elections) limit universal generalisation, since each environment has its own norms, audience expectations, and stakes. Biases based on demographic diversity (e.g., age, ethnicity, gender) remain under-explored (e.g., preference for younger/female voices). Some studies combine vocal with facial cues.	**For researchers:** Conduct single-modality (voice-only) tests to isolate acoustic influences, and then compare with multimodal tasks (audio-visual). Investigate for different speaker-listener demographics, cultures and languages. **For practitioners (e.g., speakers and trainers):** Use a moderately faster rate to project competence and a slightly lower pitch for authority—mindful of local and cultural norms, and audience preferences (e.g., age, gender).

Continued on next page

Table 1.7: Summary of trust-related acoustic features in human and IA studies: Actionable insights for practitioners and recommendations for future research (Continued)

Theme	Trustworthy acoustic features	Limitations	Recommendations and insights
Social behaviour	**Pitch:** In pro-social contexts, lower-pitched male voices are deemed more trustworthy; in antisocial contexts, a higher pitch can reduce perceived aggression or intimidation. Aligns with "frequency code" theory: lower pitch = dominance, higher pitch = submission/non-threat.	Only one study specifically contrasting pro- vs antisocial male voices. Cultural, age and gender nuances beyond male speakers remain under-explored. Other acoustic features (loudness, speech rate, voice quality) rarely examined here.	**For researchers:** Replicate with broader demographics (e.g., female, non-Western speakers-listeners) and varied social contexts. Examine pitch synergy with other acoustic and voice quality features. **For practitioners (e.g., campaign strategists):** In altruistic messaging, lower-pitched male voices may be deemed as trustworthiness. However, in negative or conflict scenarios, a slightly higher pitch may soften intimidation.
Studies: Perceived trustworthiness of voice-based IAs.			

Continued on next page

Table 1.7: Summary of trust-related acoustic features in human and IA studies: Actionable insights for practitioners and recommendations for future research (Continued)

Theme	Trustworthy acoustic features	Limitations	Recommendations and insights
Customer service	**Pitch:** Mixed or inconclusive; some data suggest higher pitch helps, others find no effect. **Speech rate & loudness:** Faster, louder voices often project competence and extroversion. **Time delay (flanging):** A delay beyond ∼0.01 seconds yields a "machine-like" sound, reducing trust. **Synergy:** Higher pitch + faster rate + louder volume can signal enthusiasm, while lower pitch + longer delay appears unnatural.	Different methods (audio vs audio-visual) produce varied pitch outcomes. Small samples or limited speaker diversity reduce generalisability. Gender stereotyping manipulations not always generalisable.	**For researchers:** Investigate how different acoustic cues interact globally (e.g., Western vs. Asian markets) to capture global variations and IA personas. Conduct A/B tests to see how minor pitch/rate tweaks affect warmth, competence, and trust. **For practitioners (e.g., chatbot / voice-tech scientists and product managers):** Use a moderately faster speech rate and louder tone for high-stakes support scenarios (e.g., billing disputes or quick issue resolutions) to convey urgency and competence. For a personalised, friendly brand, adopt moderately higher pitch and faster speech for an enthusiastic tone – or personalise based on users' mood and personality. Avoid lower pitch with steady cadence, as it risks sounding mechanical or impersonal. Limit flanging (i.e., avoid speech delays > 0.01s) and robotic intonations to ensure the voice sounds human and engaging. Track user metrics (satisfaction, conversation duration, etc). If distrust arises, tweak acoustics gradually and retest.

Continued on next page

Table 1.7: Summary of trust-related acoustic features in human and IA studies: Actionable insights for practitioners and recommendations for future research (Continued)

Theme	Trustworthy acoustic features	Limitations	Recommendations and insights
Financial services	**Pitch & speech rate:** Higher pitch + faster articulation in female-sounding IAs often associated with perceived happiness, helpfulness, competence (humans ascribe personality traits to the voice).	Mainly female British accents; potential cultural and demographic biases. Predominantly investment tasks; unsure if findings extend to other financial contexts (insurance, loans, etc).	**For researchers:** Examine if pitch and speech rate preferences hold for male voices too. Assess if a higher pitch and a faster speech rate is effective beyond investment contexts (e.g., insurance, banking, etc). **For practitioners (e.g., robot-advisor scientists and developers):** For virtual advisors, consider using a slightly higher pitch with faster articulation for competence and positive traits – be aware of accent preferences. Track conversation outcomes through real-time analytics (e.g., abandonment rates, user satisfaction). If trust declines, tweak pitch or speed gradually, then retest with A/B experiments.
Telehealth advice	**Pitch, speech rate, loudness:** A lower pitch, slower rate, and softer volume often convey empathy, especially in female voices. **Listener traits:** Extroverted listeners may trust IAs more regardless of acoustic settings, indicating individual differences may override vocal features.	Typically, small samples and varied methodologies; some purely audio, others multimodal.	**For researchers:** Develop consistent trust metrics for telehealth IAs. Investigate user personality traits (e.g., extroversion vs. introversion). **For practitioners (e.g., mental health app and companion robot designers):** For a remote triage and guidance service, providers could adopt a gentler profile (lower pitch, slower rate, softer volume) to foster a caring, professional vibe —- mindful of individual differences. Similarly, for personal therapy session, consider adaptive voice settings (e.g., pitch level, speech rate) that can be fine-tuned to patient demographics or preferences (e.g., older adults, mental health patients).

Continued on next page

Table 1.7: Summary of trust-related acoustic features in human and IA studies: Actionable insights for practitioners and recommendations for future research (Continued)

Theme	Trustworthy acoustic features	Limitations	Recommendations and insights
Safety procedures	**Synergy:** Higher pitch + faster rate + varied intonation + varied MFCC + higher intensity in listeners' responses often linked to boosted immediate trust in emergencies (fire alarms, driving instructions). Associated to feeling more invested in HAI. Trust may fade over time as urgency subsides or listeners gain more information.	Limited speakers/scenarios (often short stimuli). Long-term trust or repeated exposure seldom explored. IAs' acoustic features not examined.	**For researchers:** Examine if an "urgent voice" remains effective over prolonged or repeated alerts. Include speaker diversity (age, gender, ethnicity) for broader applicability. **For practitioners (e.g., emergency system designers):** For immediate hazard warnings (e.g., earthquake, road hazards), adopt higher pitch with a faster speech rate to convey urgenc —- then reduce intensity once people start following instructions. Alternatively, offer tiered voice prompts, where the first alert is highly urgent, followed by calmer updates to sustain trust without alarm fatigue.

1.4.3. Limitations and the future of research on trustworthy voices

The 24 papers identified in this review, represent the body of existing research in relation to speech acoustics and perceptions of trustworthiness. The current conclusions are drawn from a comprehensive synthesis of all available evidence.

Studies varied in participant numbers, with 13 involving less than 100 participants and 6 of those having less than 50 (see Table 1.5). Regarding speakers, most studies had 5 or fewer speakers, with 8 having 60 or fewer; see Table 1.6 for a summary of the stimuli and Table 1.3 for the descriptive statistics of participants and speakers across all reviewed studies. While participant sample sizes may appear limited, past research supports sample sizes of 24-36 per condition (Lavan, 2023; McAleer et al., 2014; Mileva et al., 2020). Most studies have used explicit, self-reported tasks, with some attempting real-life scenario recreation for additional behavioural data. More effort may be needed for capturing a wider range of contexts.

Most studies have relied on convenience sampling from student populations, raising concerns about demographic diversity and external validity. This sampling approach may not represent the broader population, potentially impacting the generalisability of findings. Consequently, variations in sample size and recruitment methods could have contributed to the polarised research outcomes identified, with a potential bias towards younger white generations. Moreover, online experiments have been proposed as viable alternatives to lab-based studies, offering comparable data quality and potentially better generalisability and ecological validity depending on the research question and recruitment characteristics (Del Popolo Cristaldi et al., 2022; Germine et al., 2012; Honing & Reips, 2008; Uittenhove et al., 2023).

Future research should address limitations in sample characteristics of both speakers and listeners to enhance demographic diversity and generalisability. Methodological limitations of existing studies should be acknowledged and addressed to improve the reliability of reported outcomes. Additionally, future research should explore the relationship between perceived trustworthiness based on listeners' voice ratings and their trust propensity, as well as individual differences in listeners and speakers. Cross-examinations should be expanded to include a wider range of demographic factors such as age, accents, ethnicity, and nationality, while also considering their disposition towards trust. Rigorous mixed-methods study designs should be employed to provide comprehensive insights into the effects of past and current behaviours on trustworthiness perceptions from voice acoustics, ensuring conclusive findings. Moreover, current research lacks studies examining speakers' own self-perceptions of producing trustworthy speech, which could complement existing literature on listeners' trustworthiness attributions.

Furthermore, the qualitative thematic categorisation has highlighted disparities in the depth of exploration on voice trustworthiness across different situational contexts. While themes like generic first impressions (Baus et al., 2019; Belin et al., 2019; Groyecka-Bernard et al., 2022; Mahrholz et al., 2018; McAleer et al., 2014; Mileva et al., 2018; Oleszkiewicz et al., 2017; Ponsot et al., 2018; Tsantani et al., 2016) seem to have received substantial attention, others such as telehealth advice (Goodman & Mayhorn, 2023; Maxim et al., 2023), financial

services (Torre et al., 2020, 2016) and customer service (Lim et al., 2022; Muralidharan et al., 2014; Tolmeijer et al., 2021) seem to be comparatively under-explored. This highlights the need for future research to address these gaps and expand our understanding of how vocal acoustic features influence trustworthiness perceptions across diverse contexts.

Overall, this systematic review highlights both shared and unique aspects of how trustworthiness is perceived in human voices and voice-based IAs. For human voices, judgements of trustworthiness emerge from a complex blend of acoustic features, social inferences, and interactional context. In contrast, voice-based IAs rely more on engineered acoustic profiles, yet they, too, are often evaluated along human-like social dimensions. As shown in Table 1.4 and Table 1.7, factors such as pitch, speech rate, loudness, and voice quality can be tuned to elicit or reduce trust, with different combinations proving more effective in specific scenarios (e.g., faster, louder delivery for customer service; slower, softer voices for telehealth). Moreover, Table 1.7 consolidates common acoustic features across both human and IA voices, demonstrating how certain cues, when appropriately balanced, can transcend medium or modality to influence trustworthiness perceptions.

Given these overlapping mechanisms, the need for comparative research on human and IA voices is more pressing than ever. Trust remains central to social cohesion and collaboration; thus, as voice-based IAs increasingly permeate telehealth (e.g., mental health triaging, companion robots or wellbeing apps), customer service (e.g., call centre chatbots, dispute resolution voice-based IAs), financial services (e.g., AI-driven robot advisors, voice-based personal budgeting IAs, automated insurance underwriting), and even self-driving vehicles (e.g., real-time hazard alerts and route guidance), there is a growing need to adapt these technologies so they inspire and sustain user trust – see Table 1.7 for actionable insights per industry. Moreover, since everyday tasks now blur the boundaries between human and machine interactions, understanding how we attribute trust to non-human voices is both academically significant and practically essential. A dual focus on human and synthesised voices can offer valuable insights into the cognitive processes guiding trust judgements, ultimately shaping the development of more effective, natural-sounding AI voices. By aligning voice design more closely with human-like trust cues, these systems will be better

equipped to function ethically and efficiently in an increasingly technological society.

1.4.4. Conclusion

This paper has systematically reviewed 24 studies to explore the impact of vocal acoustics on perceived trustworthiness in both human speakers and voice-based IAs, shedding light on human behaviour and attitudes toward vocal communication.

In summary, acoustic features appear to correlate with trustworthiness judgments in both human and IA voices, albeit they may exert more pronounced effects when the voice serves as the sole or predominant modality for inferring trustworthiness. Moreover, their effects are best understood within their intended contexts for enhanced interpretability. Overall, pitch seems to be influential when assessed in combination with other acoustic features, while as a sole factor it appears to be less reliable. Additionally, HAI seems to draw social inferences from human-human interactions, listeners' trust propensity and personality traits. Hence, highlighting the importance of studying these factors side by side.

To conclude, a comprehensive approach is needed to advance research on voice trustworthiness for more robust and well-rounded insights, as discussed in more detail in the limitations section of the discussion. Firstly, by considering dispositional and situational trust attitudes alongside current measures. Secondly, by cross-examining individual differences and demographic diversity in speaker-listener samples. Thirdly, there seems to be a gap in existing research regarding studies that explore speakers' self-perceptions of delivering speech with trustworthy intent, a facet that could complement the existing literature on listeners' attributions of trustworthiness. Lastly, by expanding the study of voice trustworthiness across diverse situational contexts, researchers can deepen insights into communication nuances and trustworthiness perceptions in contexts that have been less frequently investigated. See Table 1.7 for a more detailed summary of findings, paired with actionable insights for practitioners and recommendations for future research.

In closing, this review serves as a valuable reference for policymakers, researchers, and other interested parties. It offers insights into the current state of research while highlighting

existing gaps and suggesting directions for future multi-disciplinary investigations.

Transition to the next Chapter

The findings from the systematic review of Chapter 1 emphasise the significance of vocal acoustic features in shaping trustworthiness perceptions, especially when combined with demographic factors and listeners' trust predispositions. These findings point to the urgent need for a more inclusive, standardised speech dataset to enable empirical testing across diverse speaker characteristics.

Chapter 2 responds to this need by introducing a novel open-access dataset comprising demographically diverse human voices, recorded under both neutral and trustworthy intent conditions. This dataset serves as a methodological scaffold for the empirical studies that follow — supporting controlled, reproducible investigations into how vocal cues and speaker demographics jointly influence trustworthiness perception.

As each chapter in this thesis reflects stand-alone work (i.e., either published or under peer review), some background literature may be repeated. Nonetheless, the thesis progresses cumulatively — advancing from human speech production to listener-based perception, and ultimately to comparative evaluations involving synthesised voices. Chapter 2 will now present the dataset development and intent classification analyses in detail.

Chapter 2

Communicating trustworthy intent: A demographically diverse speech dataset

2.1. Introduction

The way we speak has been the subject of interdisciplinary research for decades, given its pivotal role in everyday interactions and its contribution to our survival and societal integration. Voice plays a vital role in human existence by facilitating expression, fostering connections, and conveying emotions and intentions (Kreiman & Sidtis, 2011). Moreover, it enables individuals to perceive and interpret the expressions of others, including personality traits like trustworthiness (Castelfranchi, Cesta, Conte, & Miceli, 1993).

In the area of voice acoustics, the use of recorded speech audio samples has become fundamental (Baus et al., 2019; McAleer et al., 2014; Ponsot et al., 2018). Different datasets enable scientists to examine the intricacies of voice perception and cognition, emotion recognition, and listener predispositions and personality perceptions of a speaker, among other factors (Baus et al., 2019; McAleer et al., 2014; Ponsot et al., 2018). By leveraging such voice samples, we can enhance our understanding of human communication as well as contribute to the advancement of speech technologies that have seamlessly become part of everyday life, (cf. Latinus & Belin, 2011; C. I. Nass & Brave, 2005). Table 2.1 provides a summary of the speech acoustics examined in this paper.

Re-using validated and standardised voice samples can assist researchers in conducting meaningful comparisons across studies. When the present work refers to "standardised" it means voice samples that adhere to consistent and predefined stimuli characteristics such as audio file formats, sampling rates and spoken content across speakers. This practice leads to more reliable insights and advancements in the field. However, current research on voice trustworthiness tends to rely on younger, white western populations (Baus et al., 2019; McAleer et al., 2014; Oleszkiewicz et al., 2017; Ponsot et al., 2018; Tsantani et al., 2016). Focusing primarily on white, western populations can affect the generalisability of such outcomes and miss out on additional insights that could be gained from ethnic cross-examination, sometimes referred to as white western individualist bias (WWIB), (cf. Taylor & Rommelfanger, 2022). Moreover, current research has predominantly focused on how listeners perceive speakers as trustworthy, rather than how speakers attempt to communicate trustworthy intent during speech production (Baus et al., 2019; McAleer et al., 2014; Ponsot et al., 2018; Tsantani et al., 2016). To enhance research opportunities and provide a broader, more diverse range of stimuli, a unique speech audio dataset has been created. This dataset embodies a diverse range of sentences, incorporating recordings from untrained speakers (i.e., not relying on actors) across various age-groups (i.e., age range of 18 – 90), sex, and ethnic (i.e., white, black, south Asian) backgrounds. Moreover, it encompasses both natural speech patterns and deliberate attempts to communicate trustworthiness within each spoken utterance, as perceived by the speakers.

Table 2.1: Summary characteristics of speech acoustics examined

Acoustic features	**Typically measured in...**	**Key characteristics**
Fundamental frequency (f_0); perceived as pitch	Hertz (Hz)	- f_0 is the lowest rate of vocal fold vibrations, and f_0 variability is usually captured by vocal intonation within an utterance. - "Size or frequency code" theory (Ohala, 1983, 1995): Men's lower pitch due to longer, thicker folds; women's higher pitch due to shorter folds (Frühholz & Schweinberger, 2021; Latinus & Belin, 2011; Lavan et al., 2019). - Average speaking frequencies: Men, 100-200 Hz; Women, 200-240 Hz; Children, 300 Hz (Gelfand, 2017; Mahendru, 2014; Schweinberger et al., 2014).
Amplitude; perceived as loudness	Decibels (dB)	Indicative of air pressure variations.

Continued on next page

Table 2.1: Summary characteristics of speech acoustics examined (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
HNR	dB	- Lower HNR signifies more noise in a voice signal (Fernandes et al., 2018; Ferrand, 2002). Noise in terms of voice, encompasses any component of the signal that interferes with the clarity, purity and overall quality of the intended speech signal. Typically, this noise is not harmonically related to the fundamental frequency of the voice, such as alterations in vocal fold tissue, muscle tension, respiratory patterns, or even ambient sounds and electronic interference (Ferrand, 2002). - Older adults typically show slower speech rates, lower HNR, and differences in pitch and voice quality compared to younger adults (Baus et al., 2019; Ferrand, 2002; Heffernan, 2004; Lavan et al., 2019; McAleer et al., 2014; Rojas et al., 2020).
Jitter	%	Reveals micro-fluctuations in pitch caused by irregular vocal fold vibrations (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).
Shimmer	dB	Measures micro-fluctuations in amplitude, reflecting variations in voice intensity (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).

Continued on next page

Table 2.1: Summary characteristics of speech acoustics examined (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
CPP	dB	A lower CPP is indicative of a breathy voice (Chan & Liberman, 2021; Hammarberg et al., 1980; Jalali-najafabadi et al., 2021).
LTAS	dB	A lower LTAS often indicates longer vocal tract sizes (Da Silva et al., 2011; Hammarberg et al., 1980; S. E. Linville, 2002; Löfqvist, 1986), which are linked to deeper, more resonant voices associated with dominance, particularly observed in males (Gussenhoven, 2002; Puts et al., 2007).

This paper describes this new speech audio dataset, focusing on speaker demographics in relation to their intent to sound trustworthy versus their natural speaking voice, termed "neutral" intent. The dataset is validated as to how well the acoustic features can classify trustworthy intent to understand how these speakers attempt to convey trust based on their subjective perceptions, addressing a gap in the existing literature.

2.2. Methods

2.2.1. Ethics declaration

All procedures performed in this study were approved by the Ethics Subcommittee 2 of the University of Essex (ETH2324-2113) 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 voice recordings, ratings and overall data could be (1) shared in publicly accessible archives and (2) used in future research studies.

2.2.2. Participants

Ninety-six untrained (i.e. not actors), English-speaking adults were recruited to record the audio stimuli. All younger adult speakers (all below 45 years of age), and older (all 60 years or older) white speakers were recruited online through Prolific (Prolific, 2014), an online participant recruitment panel. Most older black and older south Asian speakers were recruited through posters and word of mouth given the lack of responses on Prolific. See Table 2.2 for more details on speaker demographics. Younger adults were up to a maximum age of 45, in an attempt to have a wide-enough age gap between younger and older speakers. All speakers reported normal hearing and were given a monetary reward. Throughout this paper, the terms participants / speakers may be used interchangeably.

Table 2.2: Descriptive statistics of speaker demographics

Ethnicity	**Age-group**	**Sex**	**N**	**Mean age (years)**	**Age range (years)**	**SD**
White	Younger	Female	10	31.70	21 - 44	7.99
		Male	10	29.70	21 - 43	6.31
	Older	Female	10	70.00	60 - 87	9.51
		Male	10	67.00	60 - 76	5.85
Black	Younger	Female	11	27.64	22 - 42	5.70
		Male	9	29.22	20 - 37	6.36
	Older	Female	5	61.00	60 - 62	0.71
		Male	3	61.33	60 - 63	1.53
South Asian	Younger	Female	10	29.00	22 - 39	5.70
		Male	10	28.20	18 - 40	6.12
	Older	Female	4	66.75	60 - 77	7.63

Continued on next page

Table 2.2: Descriptive statistics of speaker demographics (Continued)

Ethnicity	Age-group	Sex	N	Mean age (years)	Age range (years)	SD
		Male	4	70.50	61 - 90	13.33

2.2.3. Materials

The materials were designed to not bias towards a specific emotional reading (e.g., You may call me anytime), as to not influence or bias the listener with loaded language or emotional tone. They were also controlled for sentence length, resulting in twenty 7-syllable sentences. A full list of the sentences created can be found in Table 2.3.

Table 2.3: All 20 sentences spoken in the speech audio dataset

Number / Code	Sentence
1	I can drive you if you want.
2	You may use my car later.
3	Hello, I arrived early.
4	I will give you a lift home.
5	You should visit more often.
6	I can remind you later.
7	You may bring a friend with you.
8	I will save a seat for you.
9	I will direct you on this.
10	Hi, the shops are still open.
11	Hi, I'm waiting for someone.

Continued on next page

Table 2.3: All 20 sentences spoken in the speech audio dataset (Continued)

Number / Code	Sentence
12	You should wear something warmer.
13	You may call me anytime.
14	I will call you a taxi.
15	You should call me tomorrow.
16	You should get to know the team.
17	I can send you a message.
18	I can give you some guidance.
19	Hello, welcome to the team.
20	You may borrow these two books.

2.2.4. Recording procedure

The recording process occurred online via a project-specific website, with participants primarily engaging remotely. However, one older adult was recorded in person due to a lack of computer access. Participants recorded their allocated materials using their personal computers and microphones. To mitigate the lack of control over the recording environment, speakers were instructed to record their voice in front of a computer that has a working microphone, in a quiet room with no background noise or other people talking or interfering, and to minimise interruptions (e.g., turn off phones). This approach follows past research from online versus lab-based studies (Germine et al., 2012; Horton, Rand, & Zeckhauser, 2011; McAleer et al., 2014).

Participants were asked to speak all sentences assigned to them twice: first, in their natural tone of voice (i.e., neutral intent), and then, with the intention of eliciting trust from the listener (i.e., trustworthy intent). To mitigate experimenter bias, no examples were provided

on how they should sound. A researcher was present remotely 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. Each participant submitted an audio file containing at least twelve utterances.

2.2.5. Audio pre-processing

2.2.5.1. Sampling rate and file format standardisation

Audacity audio editing and recording software (version 2.3.3) was used to standardise all recordings at a sampling rate of 48.0 kHz, 16-bits depth and 768 kb/s bit rate using a mono channel. The audio files were stored in an uncompressed WAV format.

2.2.5.2. Segmentation and intensity normalisation

Praat software (version 6.2.16) (Boersma, 2001) was used to segment all WAV files. Subsequently, each shorter sound file (i.e., sentence) was evaluated to eliminate any potential duplicates and normalised to 67 dB. Therefore, a total of 1,152 audio files (576 neutral and 576 with trustworthy intent) are accounted for in the final speech audio dataset.

2.2.5.3. Acoustic and spectral feature extraction

All acoustic and spectral features were extracted using VoiceLab software to analyse multiple audio files at once (D. Feinberg, 2022; D. R. Feinberg & Cook, 2020). The features used in the analyses to describe the materials are mean f_0 for perceived pitch, standard deviation of f_0 for perceived pitch variability, sentence duration, HNR, jitter, shimmer, CPP, LTAS, standard deviation of the LTAS and LTAS slope. For the analyses, VoiceLab's auto-correlation values were used for f_0, the relative average perturbation (RAP) value for jitter, and the amplitude perturbation quotient 3 (APQ3) value for shimmer, as seen in past research (Baus et al., 2019; McAleer et al., 2014). Summary descriptives of each feature per intent can be found

in Table 2.4 for white speakers, Table 2.5 for black speakers and Table 2.6 for south Asian speakers, while a definition of each acoustic can be found in Table 2.1.

Table 2.4: White speakers — Descriptive statistics of acoustic features per speaker intent, age-group and sex

Acoustic features	**Mean acoustic values [Standard deviation] for white speakers**							
	Neutral intent				**Trustworthy intent**			
	Younger female	**Younger male**	**Older female**	**Older male**	**Younger female**	**Younger male**	**Older female**	**Older male**
Voice duration	1.57 [0.31]	1.55 [0.36]	1.88 [0.36]	1.67 [0.29]	1.63 [0.39]	1.40 [0.30]	1.95 [0.48]	1.68 [0.39]
f_0**, mean (Hz)**	194.11 [18.55]	105.11 [14.24]	181.89 [24.94]	110.68 [20.72]	224.02 [24.38]	137.35 [31.63]	207.90 [27.29]	134.23 [30.31]
f_0**, SD (Hz)**	29.49 [15.84]	17.57 [12.00]	34.05 [16.17]	18.33 [14.92]	48.13 [19.10]	38.56 [24.65]	51.24 [18.15]	34.93 [18.55]
HNR (dB)	10.21 [2.62]	5.10 [2.27]	10.81 [2.63]	6.15 [1.43]	10.76 [2.74]	4.18 [2.09]	11.12 [2.28]	4.60 [1.87]
Jitter (RAP)	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]
Shimmer (APQ3)	0.04 [0.02]	0.04 [0.01]	0.04 [0.01]	0.06 [0.02]	0.04 [0.02]	0.05 [0.01]	0.04 [0.01]	0.06 [0.02]
CPP (dB)	28.20 [2.40]	25.57 [2.19]	27.69 [2.19]	24.99 [2.05]	28.64 [2.41]	25.16 [2.63]	27.93 [2.44]	24.70 [2.21]
LTAS, mean (dB)	-1.51 [6.37]	-5.42 [7.80]	-2.92 [4.92]	-7.78 [6.96]	-2.46 [6.38]	-5.75 [8.36]	-3.51 [4.99]	-7.95 [7.43]
LTAS, SD (dB)	17.27 [2.20]	18.79 [3.28]	16.67 [1.31]	18.05 [2.97]	17.53 [2.06]	18.82 [3.33]	16.95 [1.36]	18.55 [3.11]
LTAS, slope (dB/octave)	-13.16 [4.03]	-14.41 [4.54]	-15.98 [3.83]	-17.51 [4.01]	-12.58 [3.39]	-13.78 [4.53]	-16.58 [3.81]	-17.13 [4.28]

Table 2.5: Black speakers — Descriptive statistics of acoustic features per speaker intent, age-group and sex

Acoustic features	**Mean acoustic values [Standard deviation] for black speakers**							
	Neutral intent				**Trustworthy intent**			
	Younger female	**Younger male**	**Older female**	**Older male**	**Younger female**	**Younger male**	**Older female**	**Older male**
Voice duration	1.58 [0.26]	1.66 [0.46]	2.21 [0.55]	1.61 [0.25]	1.56 [0.29]	1.46 [0.36]	2.01 [0.51]	1.63 [0.31]
f_0**, mean (Hz)**	174.35 [23.94]	110.49 [22.11]	176.64 [32.06]	101.44 [13.79]	211.98 [23.32]	129.81 [21.96]	220.49 [46.00]	140.45 [48.16]
f_0**, SD (Hz)**	29.43 [16.15]	15.58 [9.80]	28.90 [14.15]	14.89 [12.12]	41.83 [16.02]	25.18 [14.13]	52.82 [21.14]	34.30 [36.34]
HNR (dB)	10.47 [2.91]	6.36 [2.19]	10.85 [3.22]	5.76 [3.29]	10.54 [2.78]	5.90 [2.52]	10.29 [3.01]	6.83 [2.67]
Jitter (RAP)	0.01 [0.00]	0.01 [0.01]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]
Shimmer (APQ3)	0.04 [0.01]	0.04 [0.01]	0.03 [0.01]	0.04 [0.01]	0.03 [0.01]	0.04 [0.01]	0.03 [0.01]	0.04 [0.01]
CPP (dB)	26.93 [2.18]	25.77 [2.18]	27.10 [2.76]	23.53 [1.83]	27.61 [2.10]	25.80 [2.44]	26.97 [3.29]	24.30 [1.68]
LTAS, mean (dB)	-3.15 [5.47]	-5.54 [7.63]	-12.72 [8.25]	-19.27 [7.03]	-2.96 [5.41]	-6.72 [8.15]	-14.05 [7.48]	-20.46 [4.02]
LTAS, SD (dB)	17.10 [1.76]	17.68 [2.60]	22.39 [5.13]	25.17 [2.31]	17.55 [1.74]	17.80 [2.73]	23.59 [5.21]	26.07 [1.97]
LTAS, slope (dB/octave)	-15.48 [4.09]	-15.47 [5.28]	-16.50 [3.39]	-15.21 [5.49]	-13.76 [4.01]	-15.21 [4.66]	-13.87 [4.47]	-15.85 [5.03]

Table 2.6: South Asian speakers — Descriptive statistics of acoustic features per speaker intent, age-group and sex

Acoustic features	Mean acoustic values [Standard deviation] for south Asian speakers							
	Neutral intent				Trustworthy intent			
	Younger female	Younger male	Older female	Older male	Younger female	Younger male	Older female	Older male
Voice duration	1.59 [0.28]	1.56 [0.27]	1.96 [0.41]	1.85 [0.46]	1.45 [0.25]	1.48 [0.27]	1.85 [0.48]	2.05 [0.72]
f_0, mean (Hz)	189.60 [25.73]	119.75 [14.40]	189.63 [12.66]	135.66 [43.21]	230.60 [35.62]	135.08 [22.65]	224.19 [40.35]	155.86 [36.61]
f_0, SD (Hz)	31.30 [15.05]	21.29 [12.15]	30.86 [10.66]	25.21 [29.67]	47.72 [18.98]	30.65 [20.10]	50.29 [14.05]	40.23 [23.73]
HNR (dB)	12.06 [3.22]	7.62 [3.29]	12.07 [1.98]	6.74 [3.37]	11.62 [2.98]	7.44 [2.98]	11.38 [3.05]	6.63 [3.92]
Jitter (RAP)	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.00]	0.01 [0.01]	0.01 [0.01]	0.01 [0.01]	0.01 [0.00]
Shimmer (APQ3)	0.04 [0.01]	0.05 [0.02]	0.04 [0.01]	0.05 [0.02]	0.03 [0.01]	0.05 [0.02]	0.04 [0.01]	0.05 [0.02]
CPP (dB)	27.38 [2.59]	25.49 [2.51]	27.95 [1.95]	26.07 [1.53]	27.37 [3.22]	24.76 [2.55]	28.06 [2.53]	27.42 [2.54]
LTAS, mean (dB)	-6.67 [10.32]	-8.62 [8.34]	-11.75 [10.01]	-9.09 [4.06]	-7.66 [10.44]	-8.06 [8.43]	-12.89 [9.50]	-8.72 [4.95]
LTAS, SD (dB)	16.96 [2.38]	16.30 [3.85]	17.26 [5.75]	18.74 [3.74]	16.97 [2.48]	16.93 [3.54]	17.55 [6.28]	18.60 [3.34]
LTAS, slope (dB/octave)	-18.22 [4.91]	-19.71 [6.59]	-19.69 [5.18]	-14.79 [4.65]	-17.69 [5.67]	-18.38 [6.87]	-18.37 [7.40]	-15.74 [3.19]

2.3. Data records

The speech audio dataset is publicly available on the Open Science Framework (OSF) repository (Maltezou-Papastylianou et al., 2024b) (DOI: 10.17605/OSF.IO/45D8J) under the CC-by Attribution 4.0 International license. All data are anonymous, and available in a folder named "Speaker Data". Inside this folder two CSV files can be found containing speaker demographics and extracted acoustic features per speech audio file. There is also a "README.md" file, which offers additional guidance on how to find and make use of the current dataset. There are also two sub-folders:

"Speech WAV Files": This sub-folder contains all 1,152 speech audio recordings of the current dataset in .wav format, normalised to 67 dB. The audio files are further split into sub-folders by speaker ethnicity and age group. The name of each audio file follows the sequence of "speaker ID"_"ethnicity""age-group""sex"_"intent""sentence number". For example, the filename "1901_bof_t05.wav" indicates that this file has been recorded by speaker ID 1901 of black (b), older (o) and female (f) demographic background who has used a trustworthy (t) intent when speaking sentence #5 (i.e., "You should visit more often"). The audio file 1901_bof_n05.wav is from the same speaker, speaking the same exact sentence but in this instance, they have used their natural speaking voice (i.e., neutral "n" intent). See Table 2.7 for more information.

"Python_SourceCode_SpeechDB": This sub-folder contains a .txt file listing all relevant Python package dependencies with their respective versions, and a "Scripts" sub-folder containing the "main.py" file for running the analyses seen in the Technical Validation section of this paper.

Table 2.7: Dataset's audio file name abbreviations

Speaker	Abbreviation	Audio filename examples
White	w	`1893_wof_t05.wav`
Black	b	`1901_bof_t05.wav`
South Asian	a	`2017_aof_t05.wav`

Continued on next page

Table 2.7: Dataset's audio file name abbreviations (Continued)

Speaker	Abbreviation	Audio filename examples
Younger	y	1906_b**y**f_t05.wav
Older	o	1901_b**o**f_t05.wav
Male	m	2233_bo**m**_t05.wav
Female	f	1901_bo**f**_t05.wav
Neutral	n	1901_bof_**n**05.wav
Trustworthy	t	1901_bof_**t**05.wav

2.4. Technical validation

The present recordings relied on speakers' intention to convey trustworthiness. To evaluate whether the captured voice samples exhibit measurable differences between neutral speech and speech with a trustworthy intent, a set of commonly used acoustic and spectral features were analysed (Brockmann-Bauser et al., 2021; Da Silva et al., 2011; Maltezou-Papastylianou et al., 2025) —- see also Table 1. These features were then used as input to classifiers to determine whether successful classification was possible, thereby validating the presence of measurable acoustic differences between the two speech intent conditions. Specifically, the speech audio dataset has been validated using established classification methods, i.e., Random Forest (RF) (Badillo et al., 2020; Couronné, Probst, & Boulesteix, 2018; Fife & D'Onofrio, 2023; Pargent, Schoedel, & Stachl, 2023; Rehman et al., 2024) and Logistic Regression (LR) (Couronné et al., 2018; Nick & Campbell, 2007; C.-Y. J. Peng, Lee, & Ingersoll, 2002). The present study investigated how trustworthy intentions during speech production relate to acoustic features across demographically diverse speakers. As the data were recorded in real-life settings outside a controlled lab environment, they may include technical variations such as differing microphone qualities and noise levels. While

these variations were anticipated, they reflect the practical challenges of data collection in non-controlled environments.

To handle the complexities of the current dataset (i.e., extracted acoustic features, diverse ethnic and age groups, speaker intent), a RF classification algorithm (126 trees; random state with a value of 1 for reproducibility purposes) was chosen for its ability to handle multi-dimensional data and robustness to noise. Moreover, RF enhances generalisability by aggregating predictions from multiple independent hierarchical models known as decision trees, and includes a built-in measure of feature importance (i.e., can assess the contribution rate of each acoustic feature towards the classification between trustworthy and neutral intents).

To further evaluate the robustness of the RF model's classification accuracy, its results were compared with another model, namely logistic regression (random state with a value of 1). For each classification method, a leave-one-speaker-out cross-validation (LOSO CV) strategy has been employed (Scherer, Stratou, Gratch, & Morency, 2013; Stumpf, Kadirvelu, Waibel, & Faisal, 2024). The added benefit of LOSO CV stems from the fact that it has allowed us to validate these models more thoroughly by assessing the model's sensitivity in discriminating trustworthy from neutral intent considering individual speaker idiosyncrasies.

2.4.1. Trustworthy intent classification

All extracted acoustic features have been used in both LR and RF models. As seen in Table 2.8, the overall (i.e., all data included) performance in detecting trustworthy speech, revealed similar metric scores between the two models. When splitting the data by ethnicity, some variation has been noted for black and south Asian ethnicities for both models. This variation may possibly be due to the unbalanced number of participants recruited per age-group for those two ethnicities in the dataset, considering that the white ethnic group and independent assessment of each age-group have gained better performance. See Table 2.9 for the confusion matrices results.

Table 2.8: LOSO CV classification results — Comparison of RF and LR trustworthy intent

Data	Random Forest				Logistic Regression			
	Accuracy	**Precision**	**Recall**	**F1 Score**	**Accuracy**	**Precision**	**Recall**	**F1 Score**
Overall	71%	73%	68%	70%	69%	71%	66%	68%
Per Ethnicity								
White	71%	70%	73%	71%	72%	73%	69%	71%
Black	68%	69%	66%	68%	68%	69%	66%	67%
South Asian	66%	68%	61%	64%	68%	69%	64%	66%
Per Age-group								
Younger adults	70%	71%	66%	68%	67%	69%	63%	66%
Older adults	69%	71%	66%	68%	72%	73%	69%	71%

Table 2.9: Confusion matrices results — Comparison of RF and LR trustworthy intent

Data	Random Forest				Logistic Regression			
	True Positives	**False Positives**	**True Negatives**	**False Negatives**	**True Positives**	**False Positives**	**True Negatives**	**False Negatives**
Overall	394	148	428	182	378	157	419	198
Per Ethnicity								
White	174	73	167	66	165	60	180	75
Black	111	49	119	57	111	51	117	57
South Asian	102	48	120	66	107	47	121	61
Per Age-group								
Younger adults	237	96	264	123	227	104	256	133

Continued on next page

Table 2.9: Confusion matrices results — Comparison of RF and LR trustworthy intent (Continued)

Data	Random Forest				Logistic Regression			
	True Positives	False Positives	True Negatives	False Negatives	True Positives	False Positives	True Negatives	False Negatives
Older adults	142	58	158	74	150	56	160	66

Moreover, these models have been evaluated through the Receiver Operating Characteristic (ROC) curves and compared the Area Under the Curve (AUC) values. The ROC curve illustrates classifier performance, while the AUC score from 0 – 1 (where 1 = perfect classifier) quantifies its ability to distinguish trustworthy from neutral intent (see Table 2.10). Both RF and LR models have reliably exhibited above average classification performance (RF AUC values between 71 – 77%; LR AUC values between 72 – 78%).

Table 2.10: AUC values — Comparison of RF and LR trustworthy intent

Data	Random Forest AUC values	Logistic Regression AUC values
Overall	77%	76%
Per Ethnicity		
White	77%	78%
Black	71%	74%
South Asian	73%	72%
Per Age-group		
Younger adults	75%	75%
Older adults	75%	76%

2.4.2. Acoustic feature importance

The Gini feature importance function was applied as part of the current RF analysis to delineate the contribution of each extracted acoustic feature towards the classification of trustworthy speaker intent – common across all speaker demographics (see Fig. 2.1 for the Gini output), as well as separately per age-group and ethnicity (see Fig. 2.2 and Fig. 2.3 for the Gini output). The Gini feature importance figures can be seen side by side for comparison with the LR acoustic significance findings (see Table 2.11, Table 2.12 and Table 2.13). Pitch, HNR, shimmer and CPP seem to be the common contributors across all speaker demographics, albeit HNR appears more prominently for LR. Moreover, significant acoustics seem to vary between models and individual demographics, with yet again the most common leaning towards, pitch and HNR. LTAS seems to be consistently low in terms of feature importance in the RF model. Overall, both models seem to offer similar observations in terms of acoustic significance towards the classification of trustworthy speaker intent. They seem to align with and offer additional insights to past research examining these acoustic features (Baus et al., 2019; Klofstad, 2016; Mahrholz et al., 2018; Schirmer et al., 2020; Torre et al., 2016; Tsantani et al., 2016).

Table 2.11: Common acoustic significance across all speaker demographics — LR acoustic feature contribution towards the classification of trustworthy intent. Classification accuracy was 69%

	Coef.(β)	**S.E.**	**z**	**p-value**	**95% C.I.**		**Odds Ratio (Exp(β))**
					Lower	**Upper**	
Voice duration	-0.27	0.18	-1.50	0.13	-0.62	0.08	0.77
f_0**, mean pitch**	0.02	0.00	7.44	**0.00**	0.02	0.03	1.02
f_0**, SD pitch**	0.03	0.01	5.83	**0.00**	0.02	0.04	1.03
HNR	-0.27	0.04	-7.76	**0.00**	-0.34	-0.20	0.76
Jitter, RAP	-38.33	20.52	-1.87	0.06	-78.54	1.89	0.00

Continued on next page

Table 2.11: Common acoustic significance across all speaker demographics — LR acoustic feature contribution towards the classification of trustworthy intent. Classification accuracy was 69% (Continued)

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
Shimmer, APQ3	-13.05	5.85	-2.23	**0.03**	-24.52	-1.59	0.00
CPP	-0.06	0.03	-1.73	0.08	-0.12	0.01	0.95
LTAS, mean	-0.03	0.01	-2.12	**0.03**	-0.05	0.00	0.97
LTAS, SD	-0.07	0.03	-2.14	**0.03**	-0.13	-0.01	0.93
LTAS, slope	0.00	0.02	-0.18	0.86	-0.04	0.03	1.00

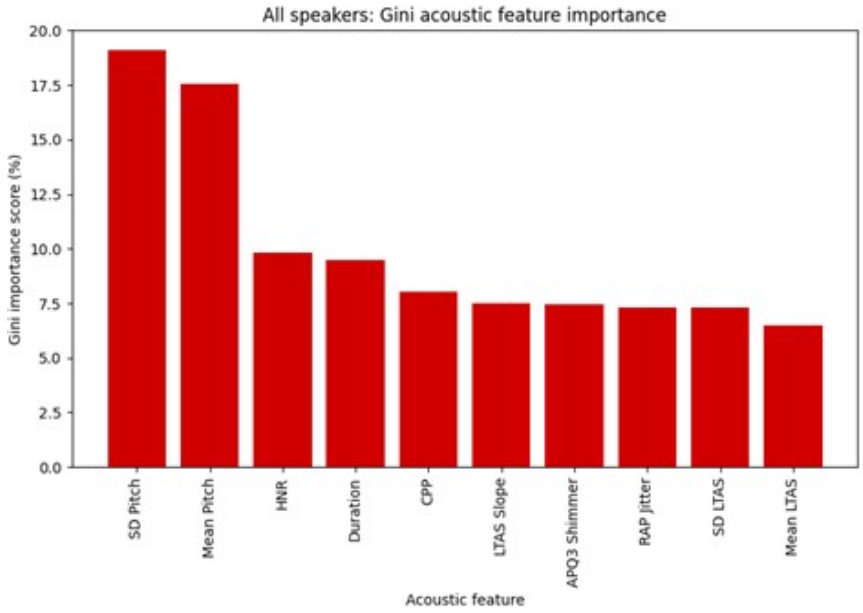

Figure 2.1: Common Gini feature importance across all speaker demographics: RF acoustic feature contribution in % towards the classification of trustworthy intent. Classification accuracy was 71%.

Table 2.12: LR acoustic feature contribution towards the classification of trustworthy intent, by speaker age-group

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
Younger adults							
Voice duration	-0.65	0.30	-2.20	**0.03**	-1.23	-0.07	0.52
f_0**, mean pitch**	0.02	0.00	6.53	**0.00**	0.02	0.03	1.02
f_0**, SD pitch**	0.03	0.01	4.16	**0.00**	0.01	0.04	1.03
HNR	-0.30	0.05	-6.36	**0.00**	-0.39	-0.21	0.74
Jitter, RAP	-32.80	25.27	-1.30	0.19	-82.33	16.73	0.00
Shimmer, APQ3	-19.76	8.40	-2.35	**0.02**	-36.22	-3.31	0.00
CPP	-0.03	0.04	-0.82	0.41	-0.11	0.05	0.97
LTAS, mean	-0.04	0.02	-2.66	**0.01**	-0.08	-0.01	0.96
LTAS, SD	-0.15	0.05	-3.17	**0.00**	-0.25	-0.06	0.86
LTAS, slope	0.02	0.02	0.72	0.47	-0.03	0.06	1.02
Older adults							
Voice duration	-0.03	0.26	-0.13	0.90	-0.54	0.48	0.97
f_0**, mean pitch**	0.02	0.01	3.35	**0.00**	0.01	0.03	1.02
f_0**, SD pitch**	0.04	0.01	4.28	**0.00**	0.02	0.06	1.04
HNR	-0.28	0.06	-5.02	**0.00**	-0.39	-0.17	0.75
Jitter, RAP	-60.32	37.04	-1.63	0.10	-132.91	12.28	0.00
Shimmer, APQ3	-10.04	9.10	-1.10	0.27	-27.87	7.79	0.00
CPP	-0.03	0.06	-0.60	0.55	-0.14	0.08	0.97
LTAS, mean	-0.01	0.02	-0.20	0.84	-0.05	0.04	1.00

Continued on next page

Table 2.12: LR acoustic feature contribution towards the classification of trustworthy intent, by speaker age-group (Continued)

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
LTAS, SD	0.01	0.05	0.22	0.83	-0.09	0.11	1.01
LTAS, slope	-0.06	0.04	-1.75	0.08	-0.13	0.01	0.94

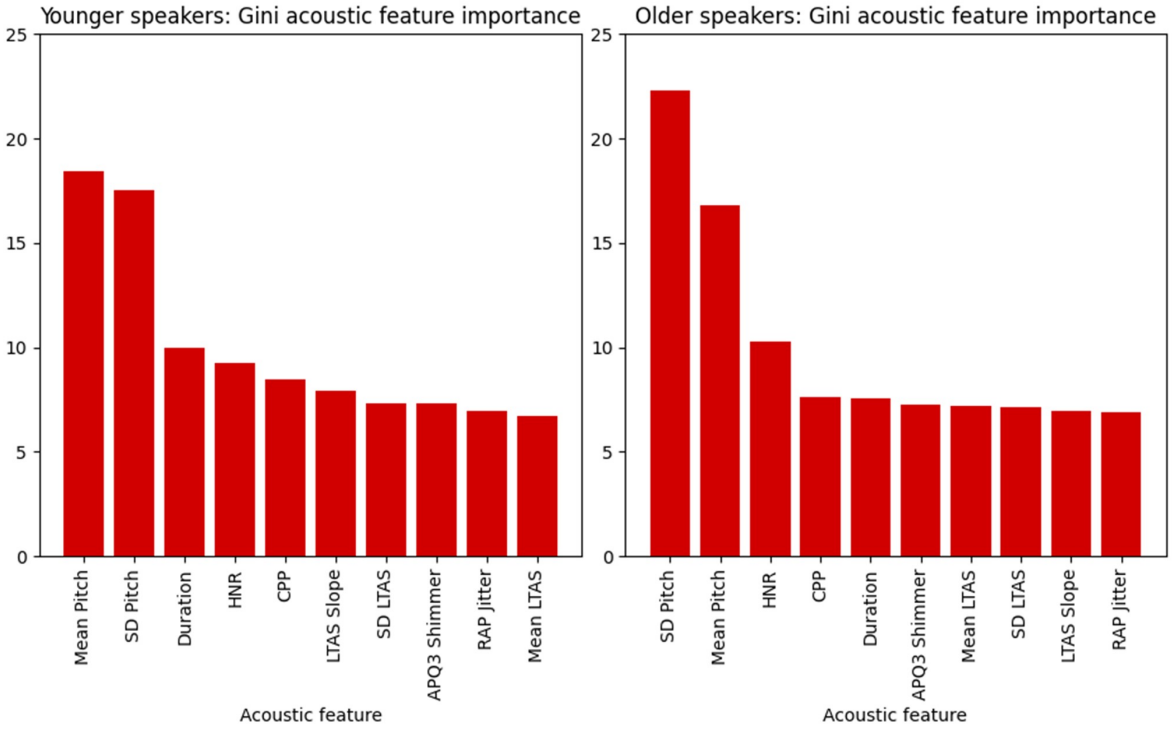

Figure 2.2: RF acoustic feature contribution in % towards the classification of trustworthy intent, by speaker age-group.

Table 2.13: LR acoustic feature contribution towards the classification of trustworthy intent, by speaker ethnicity

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
White ethnic							

Continued on next page

Table 2.13: LR acoustic feature contribution towards the classification of trustworthy intent, by speaker ethnicity (Continued)

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
Voice duration	0.26	0.30	0.85	0.40	-0.34	0.85	1.29
f_0**, mean pitch**	0.03	0.01	5.11	**0.00**	0.02	0.04	1.03
f_0**, SD pitch**	0.04	0.01	4.71	**0.00**	0.02	0.06	1.04
HNR	-0.36	0.06	-6.31	**0.00**	-0.48	-0.25	0.70
Jitter, RAP	-27.09	35.19	-0.77	0.44	-96.05	41.88	0.00
Shimmer, APQ3	-21.41	8.96	-2.39	**0.02**	-38.97	-3.86	0.00
CPP	-0.11	0.05	-1.99	0.05	-0.21	0.00	0.90
LTAS, mean	-0.02	0.02	-1.09	0.28	-0.07	0.02	0.98
LTAS, SD	-0.03	0.06	-0.48	0.63	-0.15	0.09	0.97
LTAS, slope	-0.06	0.03	-1.76	0.08	-0.12	0.01	0.94
Black ethnic							
Voice duration	-0.89	0.33	-2.68	**0.01**	-1.55	-0.24	0.41
f_0**, mean pitch**	0.03	0.01	5.15	**0.00**	0.02	0.04	1.03
f_0**, SD pitch**	0.02	0.01	2.41	**0.02**	0.00	0.04	1.02
HNR	-0.34	0.07	-4.65	**0.00**	-0.49	-0.20	0.71
Jitter, RAP	-98.56	46.27	-2.13	**0.03**	-189.25	-7.87	0.00
Shimmer, APQ3	-5.06	16.01	-0.32	0.75	-36.43	26.32	0.01
CPP	0.00	0.07	0.03	0.98	-0.13	0.13	1.00
LTAS, mean	-0.07	0.03	-2.66	**0.01**	-0.13	-0.02	0.93
LTAS, SD	-0.12	0.06	-2.08	**0.04**	-0.24	-0.01	0.89

Continued on next page

Table 2.13: LR acoustic feature contribution towards the classification of trustworthy intent, by speaker ethnicity (Continued)

	Coef.(β)	S.E.	z	p-value	95% C.I.		Odds Ratio (Exp(β))
					Lower	Upper	
LTAS, slope	-0.03	0.04	-0.89	0.37	-0.11	0.04	0.97
South Asian ethnic							
Voice duration	-0.30	0.34	-0.89	0.38	-0.97	0.37	0.74
f_0**, mean pitch**	0.02	0.01	3.44	**0.00**	0.01	0.03	1.02
f_0**, SD pitch**	0.03	0.01	2.70	**0.01**	0.01	0.05	1.03
HNR	-0.18	0.06	-2.76	**0.01**	-0.30	-0.05	0.84
Jitter, RAP	-24.57	33.82	-0.73	0.47	-90.85	41.71	0.00
Shimmer, APQ3	1.27	11.95	0.11	0.92	-22.15	24.68	3.55
CPP	-0.03	0.06	-0.53	0.60	-0.14	0.08	0.97
LTAS, mean	-0.02	0.02	-0.87	0.38	-0.07	0.03	0.98
LTAS, SD	-0.10	0.06	-1.73	0.08	-0.22	0.01	0.90
LTAS, slope	0.05	0.03	1.69	0.09	-0.01	0.11	1.05

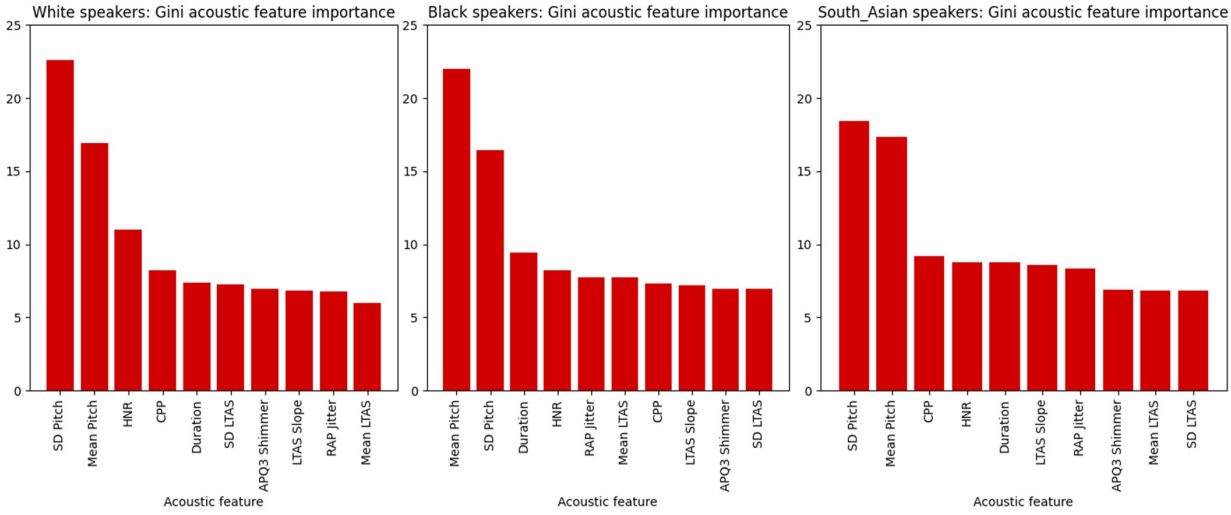

Figure 2.3: RF acoustic feature contribution in % towards the classification of trustworthy intent, by speaker ethnicity.

2.4.3. Conclusion

In this paper, a new speech dataset of 1,152 audio recordings from 96 speakers of different ethnicities (white, black, south Asian) and age groups (18 – 90 years old) was presented; this dataset allows the production of trustworthy intent as perceived by the speakers themselves, in spoken English, to be investigated. The classification of acoustic and spectral features extracted from the audio samples, yielded accuracies of about 70% and AUC values between 71 and 78% for both linear and non-linear classification models (RF and LR). Results suggest that mean f_0, SD f_0, HNR, CPP and shimmer are the most common and relevant features for discriminating natural speaking voice (i.e., neutral intent) and speech produced with the intent to sound trustworthy across all speaker demographics. LTAS seems to be the least influential factor, albeit not the case for black ethnicity in LR. Overall, the current findings seem to align with and offer additional insights to past research in the field (Baus et al., 2019; Klofstad, 2016; Mahrholz et al., 2018; Schirmer et al., 2020; Torre et al., 2016; Tsantani et al., 2016). Further analysis is needed to gain deeper insights into the production, recognition and perception of trustworthiness in spoken language, and this dataset can serve as a good resource to the research community and contribute to future research and insights in this multi-disciplinary area.

2.5. Usage notes

All data are readily accessible to the public under the terms of a CC-By Attribution 4.0 International license on an OSF repository (Maltezou-Papastylianou et al., 2024b). The present study encourages the research community to leverage and appropriately acknowledge this speech audio dataset in their analyses and publications by citing the work mentioned in the README.md file on the OSF repository.

2.6. Code availability

The Python source code employed to evaluate this dataset is openly accessible on the OSF repository (Maltezou-Papastylianou et al., 2024b). Please read the README.md file in the repository for more information on how to run the scripts yourself.

Transition to the next Chapter

Chapter 2 introduced a novel, demographically diverse speech dataset containing both neutral and trust-signalling utterances. Beyond its value as an open-access, standardised resource (Maltezou-Papastylianou, Scherer, & Paulmann, 2024a; Maltezou-Papastylianou et al., 2024b), the chapter demonstrated that vocal intent is systematically encoded across speakers via acoustic features such as pitch, HNR, and LTAS, laying a methodological and conceptual foundation for subsequent perceptual research.

Building directly on this foundation, Chapter 3 transitions from speaker-side vocal production to listener-side perception. It investigates how speakers' intentional voice modulation and specific acoustic and voice quality features shape listener perceptions of trustworthiness, warmth, and competence — three core components of social perception. This chapter will probe not only the influence of vocal intent on these social perceptions, but how these perceptions interact with acoustic profiles to guide first impressions in a voice-only modality.

While Chapter 2 established the building blocks for studying vocal trust signals, Chapter 3 uses this structure to uncover the perceptual architecture through which listeners evaluate those cues. Together, the chapters mark the shift from expressive behaviour to impression formation — bridging production and perception in the vocal trustworthiness process.

Chapter 3

Trustworthiness impressions: Vocal predictors and perceptual links to warmth and competence

3.1. Introduction

Humans rapidly form social impressions of others —- often within milliseconds —- and in many modern interactions, these impressions are based solely on voice (Asch, 1946; Lavan, 2023; Maltezou-Papastylianou et al., 2025; Mileva, 2025). From phone interviews and telehealth consultations to interactions with voice assistants like Amazon's Alexa or Apple's Siri (Kepuska & Bohouta, 2018), listeners routinely make judgments about a speaker's trustworthiness, warmth, and competence in the absence of visual cues (S. J. Ko, Judd, & Stapel, 2009; Oleszkiewicz et al., 2017). These inferences carry real-world consequences, influencing hiring decisions, cooperation, persuasion, and even perceptions of automated systems in human–computer interaction.

Among these impressions, trustworthiness holds particular psychological and functional importance; it reflects an individual's perceived benevolence, competence, and integrity —- the foundation for interpersonal and societal trust (Castelfranchi & Falcone, 2010; Hardin, 2002; Mayer et al., 1995). According to Mayer's integrative model of trust (Mayer et

al., 1995), trustworthiness is a multidimensional social construct comprising ability (i.e., perceived intelligence, skill, and efficacy), benevolence (i.e., perceived kindness or sincerity), and integrity (i.e., perceived adherence to shared principles). These components conceptually align with the broader person perception literature, where ability parallels the dimension of competence, while benevolence maps onto warmth (Cuddy et al., 2008; Guldner et al., 2024; McAleer et al., 2014). Similarly, the stereotype content model (SCM) positions warmth and competence as the two core dimensions that stimulate how we evaluate others socially (Fiske, 2018; Fiske et al., 2007; Judd, James-Hawkins, Yzerbyt, & Kashima, 2005). In this view, trust is inferred when a person is perceived as both warm (i.e., well-intentioned) and competent (i.e., capable of enacting those intentions), making these dimensions highly relevant for understanding how vocal cues shape trust-related impressions.

As such, although trustworthiness is often treated as a distinct construct, these theoretical frameworks suggest it may be more perceptually entangled with warmth and competence (Belin et al., 2019; Cuddy et al., 2008; McAleer et al., 2014). Indeed, empirical work has shown that listeners' evaluations of vocal trustworthiness may be positively associated with these impressions (McAleer et al., 2014; Oleszkiewicz et al., 2017), raising the question of whether such judgments are separable in auditory perception —- or whether they arise from a shared inferential process. This study builds on that idea by evaluating the extent to which perceived trustworthiness aligns with perceptions of warmth and competence in vocal impressions.

In parallel, past research has demonstrated that specific vocal acoustic features influence these evaluations. For instance, higher pitch (f_0) has been linked to friendliness and warmth, whereas lower pitch conveys dominance and authority, particularly in male speakers (Fantini, Fussi, Crosetti, & Succo, 2017; Klofstad, Anderson, & Nowicki, 2015; S. J. Ko et al., 2009; O'Connor & Barclay, 2018; Ohala, 1983, 1995). Moreover, faster speech rates have been associated with higher ratings of competence and trustworthiness, likely due to perceptions of vocal effort, confidence, and engagement (Gussenhoven, 2002; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012). Beyond these classic parameters, voice quality features such as harmonics-to-noise ratio (HNR), shimmer, jitter, cepstral peak prominence (CPP),

and long-term average spectrum (LTAS) offer additional information about vocal clarity, breathiness, and resonance —- qualities that can shape impressions of speaker energy, physical condition and ageing (Behlau et al., 2023; Da Silva et al., 2011; Ferrand, 2002; Jalali-najafabadi et al., 2021; Pabon, Stallinga, Södersten, & Ternström, 2014). However, the perceptual weight of these cues —- relative to each other and across demographic variation — remains under-explored (Maltezou-Papastylianou et al., 2025).

Crucially, these impressions are not only shaped by vocal mechanics but also by speaker intent (Maltezou-Papastylianou et al., 2024a). Recent work has shown that listeners are sensitive to deliberate vocal modulation —- such as when speakers intentionally try to sound trustworthy (Belin et al., 2019; Leongómez et al., 2021). Such intentional shifts in pitch contour and vocal energy can significantly alter perceived trustworthiness, suggesting that these perceptions are not solely inferred from passive, stable vocal features, but also from strategic prosodic cues (Maltezou-Papastylianou et al., 2025). Considering the limited scope of past work on trustworthy voice modulation, this presents another key opportunity for the present study, to examine whether a deliberate intent to sound trustworthy compared to no such explicit intent in a demographically diverse sample boosts perceptions across all three impressions of trustworthiness, warmth and competence.

3.1.1. Motivation

This study was designed to investigate how vocal cues, speaker intent, and demographic variation influence perceptions of trustworthiness, warmth, and competence. Drawing on the integrative model of trust (Mayer et al., 1995) and the SCM (Fiske, 2018), it was examined: whether intentional vocal modulation to appeal to listeners' trust enhance these perceptions (H1); assess the contribution of fundamental frequency (f_0), voice duration, HNR, jitter, shimmer, CPP and LTAS in these judgements (H2, H4); assess whether higher perceptions of vocal trustworthiness align with increased ratings of warmth and competence, and vice versa (H3). The present study and its hypotheses have been pre-registered on the Open Science Framework platform (`https://osf.io/485a7`).

3.2. Methods

3.2.1. Ethics declaration

All procedures performed in this study were approved by the Ethics Subcommittee 3 of the University of Essex (ETH2223-0254) 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.

3.2.2. Stimuli

Thirty-six speakers from three ethnicities (white, black and south Asian) were recruited to record three sentences each (e.g., "I can drive you if you want."; "Hello, I arrived early."; "I will give you a lift home."). Demographics of speakers can be found in Table 3.1. Sample sizes were guided by averages reported in the systematic review (see Chapter 1, Table 1.3) and selected to ensure balanced representation across speaker ethnicity and age. Sex was also considered in the sampling design, though a minor imbalance occurred in the older Black subgroup (one extra female). As sex was not analysed in the empirical models, this small deviation does not affect the interpretation of the present results.

They were asked to speak the materials once with no specific social intent (referred to here as 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. For more information on the material and recording procedure see (Maltezou-Papastylianou et al., 2024b).

Table 3.1: Descriptive statistics of speaker demographics

Ethnicity	**Age-group**	**Sex**	**N**	**Mean age (years)**	**Age range (years)**	**SD**
White	Younger	Female	3	28.3	22 - 36	7.1
		Male	3	32.3	25 - 43	9.5
	Older	Female	3	76	67 - 87	10.2
		Male	3	67.3	62 - 76	7.6
Black	Younger	Female	3	26.3	22 - 31	4.5
		Male	3	32.3	24 - 37	7.2
	Older	Female	4	61	60 - 62	0.8
		Male	2	62	61 - 63	1.4
South Asian	Younger	Female	3	29.7	22 - 37	7.5
		Male	3	29	22 - 34	6.2
	Older	Female	3	65	62 - 68	4.2
		Male	3	73	61 - 90	15.1

The VoiceLab software (D. Feinberg, 2022; D. R. Feinberg & Cook, 2020) was used to extract several acoustic and spectral features to be examined in the present study: mean f_0, standard deviation (SD) of f_0 for perceived pitch variability, voice duration (used to measure speech rate), HNR, the relative average perturbation (RAP) for jitter, and the amplitude perturbation quotient 3 (APQ3) for shimmer, as seen in past research (Baus et al., 2019; McAleer et al., 2014). The additional voice quality features of CPP, and mean, SD and slope of LTAS were extracted too. f_0 was extracted in Hertz (Hz) using VoiceLab's Praat auto-correlation values, and analysed in Hz following z-standardisation of all acoustic predictors, rather than converted to semitones. See Table 3.2 for a description of the acoustic features examined in this study.

Table 3.2: Summary characteristics of speech acoustics examined

Acoustic signal	**Measured in**	**Key characteristics**
Fundamental frequency (f_0); perceived as pitch	Hertz (Hz)	f_0 is the lowest rate of vocal fold vibrations, with vocal intonation reflected in its variability within an utterance.
HNR	dB	Lower HNR indicates increased noise in a voice signal (Fernandes et al., 2018; Ferrand, 2002). Noise refers to any element disrupting the clarity and quality of the intended speech, often unrelated to the voice's fundamental frequency; it may stem from vocal fold alterations, muscle tension, respiratory patterns, ambient sounds, or electronic interference (Ferrand, 2002).
Jitter	%	Reveals micro-fluctuations in pitch caused by irregular vocal fold vibrations (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014). A lower percentage indicates a smaller pitch variation in speech.
Shimmer	dB	Measures micro-fluctuations in amplitude, indicating variations in voice intensity (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).

Continued on next page

Table 3.2: Summary characteristics of speech acoustics examined (Continued)

Acoustic signal	**Measured in**	**Key characteristics**
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).
LTAS	dB	A lower LTAS often reflects longer vocal tracts (Da Silva et al., 2011; Hammarberg et al., 1980; S. E. Linville, 2002; Löfqvist, 1986), associated with deeper, more resonant voices linked to dominance, particularly in males (Gussenhoven, 2002; Puts et al., 2007). Conversely, higher LTAS values indicate relatively greater high-frequency energy, which has been associated with impressions of approachability and reduced threat (Lavan et al., 2019; Ohala, 1983, 1995).

3.2.3. Participants

288 English-speaking adults, balanced across ethnicity (N = 96 per group — white, black and south Asian), age (N = 48 per group —- younger and older than 60 years) and sex (N = 24 per group) — were recruited in total to rate the audio stimuli (see Table 3.3 for more details on listener demographics.).

An a-priori power analysis was conducted using the software program G∗Power (Faul, Erdfelder, Buchner, & Lang, 2009; Faul, Erdfelder, Lang, & Buchner, 2007). The goal was to obtain a test power of 95% to detect a medium effect size at the standard alpha error probability of .05. All younger adult listeners and older white listeners were recruited online through Prolific(Prolific, 2014); most of the older black and older south Asian listeners were recruited through Prime Panels (Chandler, Rosenzweig, Moss, Robinson, & Litman, 2019; CloudResearch, 2015), a participant recruitment platform that aggregates several market research panels. Younger adults were up to a maximum age of 45 years, in an attempt to have a wide enough age-gap between younger and older speakers. All listeners reported normal hearing. Throughout this manuscript, the terms participants / listeners may be used interchangeably.

Data were quality-checked prior to analysis. Participants who gave invariant responses across all trials were excluded. Recruitment continued until a final sample of 288 valid participants was achieved, consistent with the pre-registered target sample size.

Table 3.3: Descriptive statistics of participant demographics

Ethnicity	**Age-group**	**Sex**	**N**	**Mean age (years)**	**Age range (years)**	**SD**
White	Younger	Female	24	30.46	20 - 43	7.23
		Male	24	28.63	18 - 44	8.71
	Older	Female	24	66.38	60 - 80	5.98
		Male	24	64.04	60 - 72	3.71
Black	Younger	Female	24	25.08	20 - 35	5.51
		Male	24	27.96	19 - 40	5.73
	Older	Female	24	65.79	60 - 78	5.16
		Male	24	64.08	60 - 81	4.43
South Asian	Younger	Female	24	22.63	19 - 33	3.59

Continued on next page

Table 3.3: Descriptive statistics of participant demographics (Continued)

Ethnicity	Age-group	Sex	N	Mean age (years)	Age range (years)	SD
		Male	24	23.00	18 - 39	5.99
	Older	Female	24	66.17	60 - 78	4.86
		Male	24	65.67	60 - 78	5.16

3.2.4. Rating procedure

The Qualtrics software was used for screening and directing participants to the rating study that took place online on a custom-made PHP web app. During the rating task, each individual listened to 72 audio recordings, comprising 12 speakers each producing three different sentences, repeated once under two vocal conditions (neutral vs. trustworthy intent). The order of stimulus presentation was fully randomised using the Fisher-Yates Shuffle algorithm (Eberl, 2016). After hearing each utterance, participants rated the speaker on three social impressions —- trustworthiness, warmth, and competence —- using a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree).

3.3. Results

To evaluate the hypotheses, a range of statistical analyses were employed: ANOVAs (H1), linear mixed-effects models (H2 and H4) and correlations (H3). Omega-squared (ω^2) was used as an indicator of effect size for the ANOVA results. Based on past literature, although context-dependent, ω^2 = .01 (i.e. 1% of variance explained) is generally interpreted as a small effect, ω^2 = .06 as a medium effect, and ω^2 = .14 as a large effect (Field, 2018; Kirk, 1996).

3.3.1. Effects of speaker intent on impressions of trustworthiness, warmth, and competence (H1)

A repeated-measures ANOVA was conducted to test hypothesis H1, which posed that utterances spoken with a trustworthy intent would receive higher ratings on perceived trustworthiness, warmth, and competence compared to neutral utterances. A significant main effect of speaker intent was found for all three perceived traits ($p < .001$), with medium to large effect sizes, supporting H1. Specifically, trustworthiness, $F(1, 287) = 238.94$, $p < .001$, $\omega^2 = .09$; warmth, $F(1, 287) = 349.36$, $p < .001$, $\omega^2 = .19$; competence, $F(1, 287) = 179.80$, $p < .001$, $\omega^2 = .06$. Therefore, ratings across all three impressions were consistently higher for material delivered with a trustworthy intent compared to materials where speakers did not intent to sound trustworthy. See Table 3.4 for mean trustworthiness, warmth and competence impression ratings across both vocal intents.

Table 3.4: Descriptive statistics of trustworthiness, warmth and competence rating scores per speaker demographics and intent, out of a total of 7 points

Speaker ethnicity	**Speaker age-group**	**Speaker intent**	**Trustworthiness mean score**	**Warmth mean score**	**Competence mean score**
White	Older	Neutral	4.51	4.27	4.60
		Trustworthy	4.37	4.29	4.40
	Younger	Neutral	4.90	4.62	5.09
		Trustworthy	5.37	5.39	5.40
Black	Older	Neutral	4.31	4.02	4.35
		Trustworthy	4.69	4.47	4.72
	Younger	Neutral	4.64	4.41	4.74
		Trustworthy	5.02	4.89	5.08

Continued on next page

Table 3.4: Descriptive statistics of trustworthiness, warmth and competence rating scores per speaker demographics and intent, out of a total of 7 points (Continued)

Speaker ethnicity	**Speaker age-group**	**Speaker intent**	**Trustworthiness mean score**	**Warmth mean score**	**Competence mean score**
South Asian	Older	Neutral	4.20	4.05	4.13
		Trustworthy	4.31	4.28	4.28
	Younger	Neutral	4.53	4.37	4.49
		Trustworthy	4.94	5.04	4.91

Note: The present study has not controlled for age-groups. Age effects will be examined separately in the next chapter.

3.3.2. Acoustic predictors of the perceived trait impressions of trustworthiness, warmth, and competence (H2, H4)

Linear mixed-effects models (LMMs) were employed to identify common acoustic features across speaker demographics that influence listeners' perceptions of trustworthiness, warmth and competence (H2, H4). Acoustic features (mean and SD of f_0, voice duration, HNR, RAP jitter, APQ3 shimmer, CPP, and mean, SD and slope of LTAS) acted as the predictors, listener ratings as the target and listeners as the random effects. All three LMMs included random intercepts by listener to account for individual baseline differences in rating tendencies. Random slopes were not specified, as the hypotheses (H2, H4) concerned the average effects of acoustic features across listeners in the sample, rather than variability in predictor effects across listeners. For example, the model for trustworthiness ratings was specified as:

$$Trustworthiness \sim VoiceDuration + Mean f_0 + SD f_0 \times SpeakerEthnicity + HNR + Jitter + Shimmer + CPP + MeanLTAS + SDLTAS + LTASSlope + (1|Listener)$$

Speaker ethnicity was also included as a predictor to examine whether the effect of pitch variability (SD of f_0) on these trait impressions varied across ethnic groups, given prior evidence that intonational norms and expressive pitch use differ across cultural backgrounds,

yet most existing work has relied predominantly on white western speakers (Baus et al., 2019; Belin et al., 2019). By including ethnicity allowed this study to address whether findings generalise to more diverse speaker samples.

To allow for meaningful comparisons and prevent bias due to different units of measurement, all acoustic predictors were standardised (mean = 0, SD = 1). For example, features like mean f_0 and voice duration differ in units (Hertz vs seconds), making direct comparisons challenging. Standardisation ensures a fair assessment of how each acoustic and voice quality feature influences perceptions of trustworthiness, warmth, and competence. For more details on the LMM results and list of acoustic features examined, see Table 3.5 – Table 3.7.

This made it possible to directly assess the contribution of each vocal quality, such as how changes in pitch or voice duration specifically impacted perceptions of trustworthiness, warmth, and competence.

3.3.2.1. Perceived trustworthiness ratings

Voice duration, HNR and shimmer had a significantly negative association with trustworthiness ratings, whereas mean f_0, mean LTAS and LTAS slope were also significant but positively associated. A negative association means that as these acoustic factors increased, trustworthiness ratings decreased. On the other hand, a positive association means that as certain acoustic factors increased, trustworthiness ratings also increased. Interestingly, the effect of pitch variability (SD of f_0) on perceived trustworthiness was positive for white speakers, smaller in magnitude for black speakers, and negative for south Asian speakers, indicating a lack of generalisability across ethnicities. Additionally, the differences in how listeners grouped these factors through their ratings was represented by σ^2 = .40, indicating some variability in individual rating tendencies. For more details see Table 3.5.

Table 3.5: LMM summary table for trustworthiness ratings

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
Intercept	4.59	0.04	116.31	**0.00**	4.51	4.67
Speaker ethnicity [T.Black]	0.00	0.02	-0.18	0.86	-0.05	0.04
Speaker ethnicity [T.S.Asian]	0.01	0.02	0.52	0.60	-0.03	0.06
Voice duration	-0.17	0.01	-18.65	**0.00**	-0.18	-0.15
f_0**, mean**	0.37	0.02	19.63	**0.00**	0.34	0.41
f_0**, SD**	0.07	0.02	4.58	**0.00**	0.04	0.10
f_0**, SD x ethnicity [T.Black]**	-0.05	0.02	-2.65	**0.01**	-0.09	-0.01
f_0**, SD x ethnicity [T.S.Asian]**	-0.15	0.02	-6.91	**0.00**	-0.19	-0.11
HNR	-0.27	0.02	-16.22	**0.00**	-0.30	-0.24
Jitter, RAP	-0.02	0.01	-1.44	0.15	-0.04	0.01
Shimmer, APQ3	-0.11	0.01	-9.12	**0.00**	-0.13	-0.08
CPP	0.02	0.01	1.87	0.06	-0.00	0.04
LTAS, mean	0.14	0.02	8.69	**0.00**	0.11	0.18
LTAS, SD	-0.02	0.02	-1.05	0.30	-0.05	0.02
LTAS, slope	0.03	0.01	2.77	**0.01**	0.01	0.06
Grouping variable	0.40	0.03				

3.3.2.2. Perceived warmth ratings

Voice duration, HNR, jitter, shimmer and LTAS slope showed a significantly negative relationship with warmth ratings, whereas mean f_0 and mean LTAS were significantly positive. The effect of pitch variability (SD of f_0) on perceived warmth was significantly positive for white speakers, significantly negative for south Asian speakers and non-significant toward black speakers. For SD f_0 in particular, there was a positive relationship

with warmth ratings for white speakers but a negative relationship for south Asian speakers. The variance of the listeners' grouping variable was σ^2 = .25. For more details see Table 3.6.

Table 3.6: LMM summary table for warmth ratings

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
Intercept	4.38	0.03	131.76	**0.00**	4.31	4.44
Speaker ethnicity [T.Black]	0.03	0.02	1.15	0.25	-0.02	0.08
Speaker ethnicity [T.S.Asian]	0.11	0.02	4.53	**0.00**	0.06	0.16
Voice duration	-0.15	0.01	-15.55	**0.00**	-0.17	-0.13
f_0**, mean**	0.47	0.02	22.73	**0.00**	0.43	0.51
f_0**, SD**	0.10	0.02	5.93	**0.00**	0.07	0.14
f_0**, SD x ethnicity [T.Black]**	-0.00	0.02	-0.11	0.91	-0.04	0.04
f_0**, SD x ethnicity [T.S.Asian]**	-0.17	0.03	-7.15	**0.00**	-0.22	-0.12
HNR	-0.33	0.02	-18.24	**0.00**	-0.36	-0.29
Jitter, RAP	-0.02	0.01	-2.01	**0.04**	-0.05	0.00
Shimmer, APQ3	-0.13	0.01	-10.40	**0.00**	-0.16	-0.11
CPP	-0.01	0.01	-0.93	0.35	-0.03	0.01
LTAS, mean	0.21	0.02	11.75	**0.00**	0.18	0.25
LTAS, SD	0.01	0.02	0.27	0.79	-0.03	0.04
LTAS, slope	-0.03	0.01	-2.02	**0.04**	-0.05	-0.00
Grouping variable	0.25	0.01				

3.3.2.3. Perceived competence ratings

Voice duration, HNR and shimmer showed a significantly negative association with ratings of competence, while mean f_0, CPP, mean LTAS and LTAS slope were reported as significantly

positive. However, SD f_0 was only positively associated to competence ratings toward white speakers, while for black and south Asian speakers it was negatively associated. The variance of the listeners' grouping variable was σ^2 = .41. For more details see Table 3.7.

Table 3.7: LMM summary table for competence ratings

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
Intercept	4.76	0.04	120.99	**0.00**	4.69	4.84
Speaker ethnicity [T.Black]	-0.05	0.02	-2.21	**0.03**	-0.09	-0.01
Speaker ethnicity [T.S.Asian]	-0.12	0.02	-5.54	**0.00**	-0.16	-0.08
Voice duration	-0.19	0.01	-22.45	**0.00**	-0.21	-0.18
f_0**, mean**	0.27	0.02	14.40	**0.00**	0.23	0.30
f_0**, SD**	0.07	0.02	4.35	**0.00**	0.04	0.10
f_0**, SD x ethnicity [T.Black]**	-0.07	0.02	-3.80	**0.00**	-0.11	-0.03
f_0**, SD x ethnicity [T.S.Asian]**	-0.14	0.02	-6.56	**0.00**	-0.18	-0.10
HNR	-0.25	0.02	-15.99	**0.00**	-0.29	-0.22
Jitter, RAP	0.00	0.01	-0.38	0.71	-0.02	0.02
Shimmer, APQ3	-0.10	0.01	-8.41	**0.00**	-0.12	-0.07
CPP	0.03	0.01	3.03	**0.00**	0.01	0.05
LTAS, mean	0.15	0.02	9.38	**0.00**	0.12	0.18
LTAS, SD	-0.03	0.02	-1.51	0.13	-0.06	0.01
LTAS, slope	0.03	0.01	2.73	**0.01**	0.01	0.05
Grouping variable	0.41	0.03				

To summarise, across all three first impressions —- trustworthiness, warmth, and competence —- multiple acoustic features significantly predicted listener ratings, supporting H2. In particular, voice duration, mean pitch (f_0), pitch variability (SD of f_0), HNR, shimmer,

and LTAS measures consistently emerged as significant predictors. In line with H4, the analysis specifically tested the interaction between pitch variability (SD of f_0) and speaker ethnicity, to examine whether this effect generalised across ethnic groups. Other potential interactions, such as listener demographics, were not examined, as they were beyond the scope of the current study. Pitch variability results showed a positive association with all three impressions for white speakers, but this effect was weaker for black speakers and reversed (negative) for south Asian speakers. Thus, these interaction effects indicate that the relationship between pitch variability and trait impressions does not generalise equally across speaker ethnicities, offering only partial support for H4.

3.3.3. The relationship between perceived trustworthiness, warmth, and competence (H3)

To test the hypothesis that perceived trustworthiness, is positively related to perceptions of warmth and competence (H3), Spearman's rank correlation was used. The ratings for trustworthiness, warmth and competence were averaged across all listeners for each audio file. This approach helped to examine how these impressions were related to the audio stimuli themselves, rather than individual listener differences.

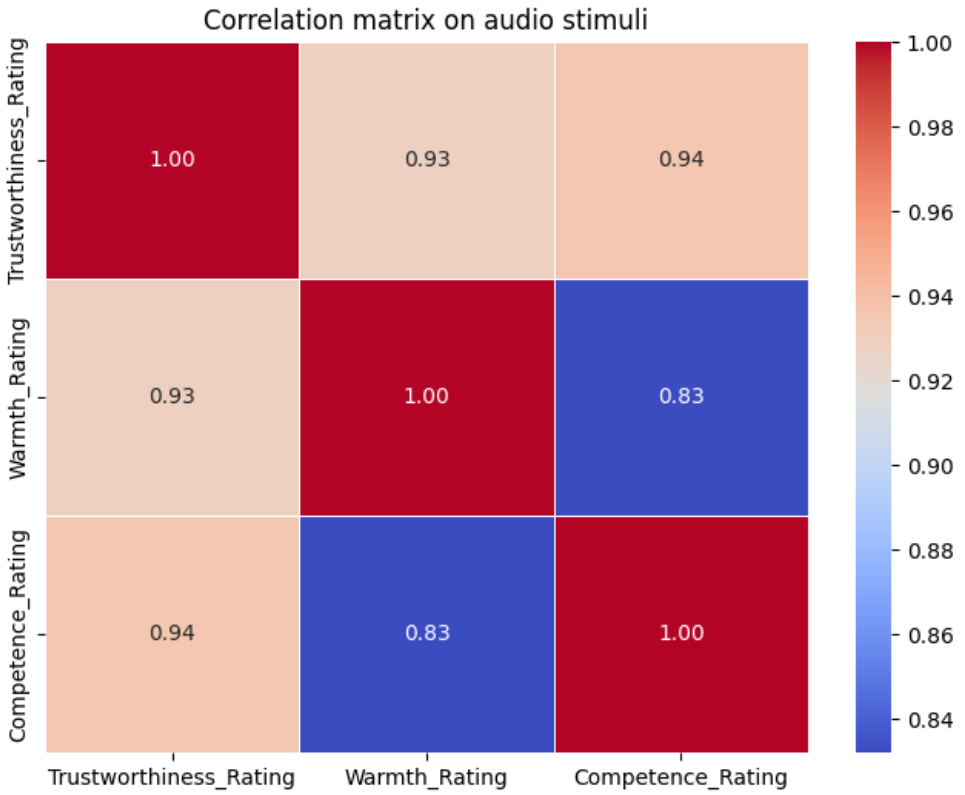

Figure 3.1: Spearman's rank correlation matrix (1 = perfect relationship, 0 = no relationship) between ratings of perceived trustworthiness, warmth and competence in relation to the audio stimuli.

As seen in Fig. 3.1, a significant, positive correlation was found between perceptions of trustworthiness and those of warmth, $r(214) = .93$, $p < .001$, and competence, $r(214) = .94$, $p < .001$. A significant correlation was also found between perceived warmth and competence, $r(214) = .83$, $p < .001$. These findings support H3, showing that audio files rated as more trustworthy were also rated higher in warmth and competence, and thus, suggesting that listeners tend to evaluate these impressions in a closely interrelated manner when judging this study's vocal stimuli.

3.4. Discussion

This study examined how vocal cues shape first impressions, with a focus on trustworthiness, warmth, and competence — three core dimensions of social perception and interpersonal evaluation (Fiske et al., 2007; Mayer et al., 1995). Grounded in the integrative model of trust by Mayer et al. (1995), and the Stereotype Content Model (SCM) by Fiske (2018); Fiske et al. (2007), the study has investigated whether these social impressions are perceptually interlinked in a voice-based setting, rather than evaluated independently. In addition, it was investigated whether deliberate vocal modulation intended to convey trustworthiness, alongside specific acoustic cues — voice duration, mean pitch (f_0), pitch variability (SD of f_0), HNR, jitter, shimmer, CPP and LTAS measures —- could predict listener ratings of trait impressions. In particular, the study examined whether the effect of pitch variability generalises across the speaker ethnicities (white, black, south Asian) represented in the vocal stimuli of the present study. The following paragraphs evaluate the findings in relation to these aims and theoretical frameworks.

3.4.1. Interrelations between impressions of trustworthiness, warmth, and competence

One of the central aims of this study was to determine whether voice-based trustworthiness is evaluated as a distinct perceptual construct, or whether it reflects a deeper overlap with warmth and competence — two core dimensions of social cognition according to the Stereotype Content Model (Fiske, 2018; Fiske et al., 2007). The present findings revealed strong, positive associations (see Fig. 3.1) between perceived trustworthiness, warmth, and competence. These strong associations suggest that listeners may not always differentiate clearly between these dimensions when judging brief vocal utterances. However, it's important to recognise that just because these impressions are strongly correlated does not mean that listeners perceive them as exactly the same thing, or that they are formed in the same way (cf., Cuddy et al., 2008; Judd et al., 2005). These constructs remain theoretically

distinct within models such as the SCM (Fiske et al., 2007) and the integrative model of trust (Mayer et al., 1995). Instead, what the current findings may reflect is a functional overlap in voice-based evaluations, whereby listeners infer trustworthiness from a combination of warmth and competence cues, particularly in the absence of richer contextual or visual information —- supporting earlier voice-based findings (Guldner et al., 2024; McAleer et al., 2014).

Our results further corroborate past behavioural work on short speech stimuli by Lavan (2023), showing that listeners can form first impressions in approximately 400 milliseconds, and likely rely on holistic processing of socially desirable vocal characteristics (Lavan, 2023; Maltezou-Papastylianou et al., 2025; McAleer et al., 2014). In particular, Lavan (2023) showed that dominance impressions, like competence, tend to be judged more rapidly than trustworthiness, which tends to strengthen over a more gradual time-course within those milliseconds. The reasoning behind the timing difference seems to be attributed to the nature of social attributions being more gradual in assessment over intellect attributions. Notably, it was currently found that trustworthiness correlated more strongly with warmth than with competence, reinforcing the idea that vocal signals of intent (e.g., friendliness, sincerity) may carry more weight than signals of capability when inferring trustworthiness — particularly in early or decontextualised interactions (Belin et al., 2019; Castelfranchi & Falcone, 2010; Fiske et al., 2007; Hardin, 2002). This asymmetry highlights the need to reconsider how trustworthiness is operationalised in auditory studies. While current theoretical models on social perception may treat warmth, competence and trustworthiness as analytically separable (Belin et al., 2019; McAleer et al., 2014), the current findings suggest that perceived benevolence or approachability may play a more central role in early vocal judgements of trust in social settings. This interpretation is further supported by prior studies where warmth-related descriptors frequently co-occurred with perceived trustworthiness, more strongly so in social and emotional over intellectual scenarios (Cuddy et al., 2008; Guldner et al., 2024; Maltezou-Papastylianou et al., 2025; Oleszkiewicz et al., 2017).

Taken together, these findings draw attention to the need for future research to

critically examine how dimensions of trait impressions are operationalised and interpreted in multidisciplinary voice-based perception studies. For researchers, this highlights the importance of using multi-trait measurement approaches and modelling perceptual dependencies, particularly when studying trustworthiness. For applied contexts —- such as virtual assistants, interview scenarios, or persuasive communication — these insights emphasise that fostering vocal warmth may be especially effective in conveying trustworthiness. Practitioners working on vocal training, digital voice design, or social skill development should recognise that warmth-related impressions may dominate listeners' assessments, particularly in early or lower-information encounters, such as in visually or physically absent interactions.

Given the strong perceptual interdependence between trustworthiness, warmth, and competence, it became particularly relevant to explore whether speakers can deliberately influence these impressions by intentionally trying to gain listeners' trust. This question has been addressed by assessing whether an explicit vocal intent to sound trustworthy would enhance ratings across all three impressions.

3.4.2. The effect of speaker intent on impressions of trustworthiness, warmth, and competence

This study also set out to examine whether an explicit vocal intent to sound trustworthy would influence perceptions of trustworthiness, warmth, and competence. The results offered clear support for this hypothesis across diverse ethnic and age groups. Across all three impressions, utterances delivered with a trustworthy intent were rated significantly higher than those delivered with no such intent. Crucially, the effect sizes ranged from large for trustworthiness and warmth to medium for competence. These findings build on previous evidence that intentional prosodic modulation has a measurable impact on social trait attributions (Belin et al., 2019; Guldner et al., 2024).

That speakers can modulate their voices to influence how they are perceived reflects the strategic, goal-directed nature of human vocal behaviour. Prior work has demonstrated that

listeners form social impressions of trustworthiness rapidly and from minimal acoustic input (Lavan, 2023; McAleer et al., 2014), and that these impressions are sensitive not only to stable vocal properties but also to intentional shifts (Leongómez et al., 2021). The current findings extend this by showing that such modulation significantly shifts listeners' evaluations across all three core social impressions of trustworthiness, warmth and competence.

Additionally, the larger effect on warmth, followed by trustworthiness, supports theories suggesting that warmth (or benevolence) plays a primary role in trust-related impression formation (Asch, 1946; Fiske et al., 2007; Mayer et al., 1995; Oleszkiewicz et al., 2017). As such, it may strengthen existing literature, by suggesting that warmth may serve as a perceptual shortcut for inferring trustworthiness in lower-information, social settings like brief vocal-only encounters.

From an applied standpoint, these findings have broad implications (Behlau et al., 2023; Fantini et al., 2017; Pabon et al., 2014). In professional communication, customer service, human-computer interactions, or vocal training programmes, speakers could be encouraged to adopt prosodic strategies that signal trustworthiness, particularly those that enhance vocal warmth; for instance, vary intonation more (increased pitch variability), maintain a brisk but steady speech rate, and reduce vocal roughness or breathiness, without hindering the speaker's individuality and authenticity. In contexts where trust must be established quickly and without visual cues (e.g., phone interviews, voice assistants, or telehealth) these vocal adjustments may enhance listener perceptions of sincerity and competence. However, it is important to approach such applications with caution. As the ability to influence impressions through voice alone becomes more sophisticated, so does the potential for misuse — emphasising the need for ethical guidelines in the design and deployment of voice-based systems, particularly in high-stakes or vulnerable settings. Future research should explore how different kinds of intent (e.g., to sound dominant, friendly, honest, competent) are encoded and decoded across diverse linguistic and cultural contexts.

3.4.3. Acoustic features predicting perceptions of trustworthiness, warmth, and competence

Our findings demonstrated that key acoustic features — voice duration, mean perceived pitch, pitch variability, shimmer, HNR, and LTAS —- consistently predicted listener perceptions of trustworthiness, warmth, and competence. Crucially, these associations held across a demographically diverse speaker sample encompassing different age groups (under and over 60 years) and ethnic backgrounds (white, black, and south Asian). This consistency across multiple age and ethnic groups within the current dataset highlights the internal robustness of these acoustic predictors. While further research is required to test generalisability across broader and more diverse populations, the present findings suggest that listeners may rely on consistent perceptual heuristics when evaluating vocal trustworthiness, warmth, and competence within the current data sample. The stability of these effects across diverse speaker demographics in the present study points to underlying regularities in how vocal impressions are socially interpreted. In the following sections, the contribution of each acoustic feature is examined in greater detail, discussing the current findings within the broader theoretical and empirical literature.

Across all three impressions, shorter voice duration — here reflecting a faster speech rate —- was associated with higher ratings of trustworthiness, warmth, and competence. This aligns with previous research suggesting that faster speech rates may be associated with increased vocal effort, which in turn can convey confidence, engagement, eagerness to help, and reliability (Gussenhoven, 2002; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012). In the context of the current brief utterances, where semantic content is minimal, such temporal cues may take on heightened importance, offering listeners a perceptual shortcut for inferring speaker credibility (Schirmer et al., 2020; Yokoyama & Daibo, 2012).

Mean pitch also showed a positive relationship with trustworthiness, warmth, and competence, supporting evidence that higher pitch may signal friendliness, openness and sincerity (Ohala, 1983, 1995; Oleszkiewicz et al., 2017). Similarly, this effect may be

especially salient in brief utterances where listeners rely more heavily on acoustic cues such as intonation to form rapid judgements (Belin et al., 2019; McAleer et al., 2014). However, it is important to acknowledge that pitch is not universally interpreted the same way: in other communicative contexts, lower pitch has been associated with authority and dominance (Klofstad et al., 2015; S. J. Ko et al., 2009; Large & Burnett, 2014; O'Connor & Barclay, 2018). Thus, while the present results support a link between higher pitch and positive first impressions, they also emphasise the importance of situational framing and communicative goals in interpreting pitch-related effects.

Features related to voice quality —- HNR, shimmer, CPP, and LTAS —- collectively emerged as significant predictors of social trait impressions. In the current data, lower shimmer and HNR were associated with higher ratings of trustworthiness, warmth, and competence, whereas higher LTAS measures were associated with higher ratings of all three impressions and higher CPP with higher ratings of competence. While these features differ in their acoustic definitions, their combined contribution paints a more complex picture than traditional interpretations might suggest.

While lower shimmer and HNR are typically linked to greater vocal noise or instability, often associated with vocal ageing or diminished physiological control (Farrús et al., 2007; Fernandes et al., 2018; Ferrand, 2002), they did not appear to be uniformly penalised in the current data. Instead, these characteristics, jointly with the rest of the acoustic findings, could have been interpreted as signs of vocal expressiveness and energy, often associated with such positive impressions — an interpretation consistent with findings from Schirmer et al. (2020) and echoed in the systematic review by (Maltezou-Papastylianou et al., 2025).

This interpretation is further supported by the positive associations found for CPP and LTAS measures (mean and slope), which reflect the strength of harmonic structure and the distribution of vocal resonance, respectively (Da Silva et al., 2011; Jalali-najafabadi et al., 2021; S. E. Linville, 2002). The present associations for LTAS can be better understood by considering that it reflects the balance of energy across low and high frequencies, which is influenced by physiological factors such as vocal tract length. Within the "frequency code" theory (Ohala, 1983, 1995), greater high-frequency energy (higher LTAS mean, flatter

slopes), seen primarily in the voices of women and children with shorter vocal folds, tend to be perceived as less threatening and more approachable (Lavan et al., 2019; Lim et al., 2022; Schweinberger et al., 2014). In contrast, relatively less high-frequency energy (steep slopes) can convey authority and dominance (O'Connor & Barclay, 2018; Puts et al., 2007). In the present study, LTAS mean was consistently associated with higher ratings of trustworthiness, warmth, and competence, while the slope of the LTAS was positively associated with competence and trustworthiness but negatively with warmth. This pattern suggests that both approachability and authority cues contribute to social evaluations of trustworthiness, which, in combination with other vocal characteristics, can portray speakers as persuasive and credible — positive traits that have been linked to trust formation (Gussenhoven, 2002; Lim et al., 2022; Mileva, 2025; Rodero et al., 2014). Together, these spectral features may potentially amplify impressions of clarity, brightness, and vocal engagement, qualities commonly linked to perceived youthfulness, energy, and sociability (Behlau et al., 2023; Hammarberg et al., 1980; Löfqvist, 1986; Pabon et al., 2014). Rather than operating in isolation, these voice quality cues appear to function as a perceptual cluster —- one that listeners may use as a heuristic to gauge a speaker's vitality, approachability, and emotional accessibility.

Thus, the current findings suggest that the synergistic nature of the identified acoustic and voice quality features —- speech rate, pitch, shimmer, HNR, CPP, and LTAS — may collectively signal expressiveness, energy, and youthfulness, contributing to more positive evaluations of trustworthiness, warmth, and competence. Notably, this association emerged even for the voice quality features, like HNR, that are traditionally linked to acoustic irregularity or vocal ageing, highlighting the need to reconsider their perceptual role in social evaluation (Ferrand, 2002). These results support a growing shift towards a social-perceptual perspective on voice quality — one that moves beyond technical acoustic metrics to consider how vocal aesthetics are interpreted through the lens of listener expectations, cultural norms, and demographic diversity (Maltezou-Papastylianou et al., 2024a, 2025).

Importantly, while most acoustic and voice quality features showed consistent effects across the present demographically varied speaker sample, currently, only pitch variability

was explicitly examined for ethnic differences in first impressions. This decision was driven by theoretical and empirical interest in pitch variability, which has been widely discussed as an influential marker of emotional expressiveness and social intent (Maltezou-Papastylianou et al., 2022, 2025), and thus, its interpretation may be more culturally contingent than more stable acoustic parameters. Specifically, it was currently found that pitch variability was positively associated with all three impressions for white speakers, but this effect was smaller for black speakers and reversed (negatively associated) for south Asian speakers. These results echo prior suggestions that prosodic expressiveness norms vary across cultures, and that listeners may interpret pitch modulation through culturally specific expectations (Baus et al., 2019; Jiang, Gossack-Keenan, & Pell, 2020; Sebastian & Ryan, 2018). Therefore, greater pitch variability may be perceived as warmer, more trustworthy, and emotionally expressive in some voices more than others — albeit, depending on ethnic and cultural background — particularly in the absence of contextual or visual cues. This draws attention to the importance of future research in examining how vocal variation is interpreted across different ethnolinguistic groups.

Finally, the demographically controlled breadth of this study marks a critical step toward more inclusive models of voice-based impression formation. Few prior studies have systematically examined whether acoustic predictors retain their relevance across speakers of different ethnicities and ages (Maltezou-Papastylianou et al., 2024a, 2025). The present findings show that while speech rate, mean pitch, shimmer, CPP, LTAS, and HNR consistently influenced ratings across groups, their perceptual impact appears to operate not in isolation, but jointly, offering more nuanced and complex impressions when interpreted together. Moreover, while those features showed stable effects across demographics, others, like pitch variability, were more selective in their influence. In doing so, this study offers new evidence to refine vocal perception theories by emphasising the layered teamwork between structural and expressive cues, and by highlighting which features remain consistent across age-related and ethnically diverse voices.

From a practical standpoint, these insights offer actionable implications for communication training, voice design, and social perception research. For speakers or professionals

aiming to optimise their vocal impression —- whether in customer service, healthcare, or public speaking —- emphasising vocal clarity, controlled pitch modulation, and efficient speech timing may enhance perceptions of trustworthiness, warmth and competence (Behlau et al., 2023; Pabon et al., 2014). For developers of virtual agents or synthetic voices, modelling both expressive and stable vocal qualities may improve listener engagement and trust.

3.4.4. Conclusion

This study emphasises the complex and layered nature of voice-based social judgements, revealing how acoustic features (i.e., speech rate, mean pitch and variability, HNR, shimmer, CPP and LTAS), speaker intent (trustworthy vs neutral), and demographic factors (age and ethnicity) collectively shape social perceptions of the interrelated but distinct constructs of trustworthiness, warmth, and competence. While intentional vocal modulation emerged as a significant influence, listeners also relied on a network of vocal cues —- such as pitch, speech rate, and vocal stability —- to form their impressions. Notably, demographic variables such as age and ethnicity modulated these effects in ways that warrant further investigation, suggesting that societal and cultural factors may interact with acoustic signals to produce nuanced listener responses.

These findings not only advance social perception theories by demonstrating how seemingly discrete cues converge to inform rapid interpersonal evaluations, but they also offer practical applications. Communication coaches and public speakers, for instance, can use targeted vocal exercises to optimise pitch variation or reduce breathiness, thereby enhancing perceived warmth or trustworthiness. Moreover, designers of voice-based artificially intelligent systems (e.g., voice assistants, automated call centres) may incorporate acoustic modifications that foster more positive user experiences across diverse cultural and linguistic contexts. By identifying how certain vocal parameters resonate with listeners' perceptions, this study opens avenues for further inquiry into how voice-based impressions can be harnessed or adapted to meet the communicative demands of an increasingly interconnected

and technology-driven society.

Transition to the next Chapter

Chapter 3 explored how vocal acoustic features shape trustworthiness perceptions, highlighting the role of vocal modulation across diverse speaker demographics. It also demonstrated that trustworthiness is closely intertwined with warmth and competence, suggesting that voice-based trustworthiness impressions emerge from a constellation of overlapping social impressions (Fiske et al., 2007; Oleszkiewicz et al., 2017). These impressions were systematically linked to acoustic features such as pitch, HNR, shimmer, CPP, and LTAS, reinforcing the idea that trustworthiness is not a stand-alone construct but part of a broader socio-acoustic perceptual space. Existing models of voice perception have primarily focused on the cognitive processing of speaker identity, emotion, and paralinguistic information (e.g., Belin, Bestelmeyer, Latinus, & Watson, 2011; Belin, Fecteau, & Bedard, 2004; Lavan et al., 2019; Schweinberger et al., 2014), with research on vocal trustworthiness often centred narrowly on pitch, speech rate, or HNR (e.g., McAleer et al., 2014; Schild, Stern, & Zettler, 2020; Yokoyama & Daibo, 2012). The present study extends this literature by demonstrating that listeners also draw on voice quality features —- such as CPP and LTAS —- which have rarely been examined in social impression formation. This highlights that voice-based trustworthiness evaluations rely on a richer constellation of acoustic cues than previously examined, particularly in voice-only settings where visual and contextual information is absent or limited.

Expanding this framework, Chapter 4 investigates how speaker-listener demographic group membership and listener predispositions influence vocal trustworthiness perceptions. In particular, it examines whether shared age and ethnicity between speakers and listeners affects trustworthiness ratings, and whether vocal intent can mitigate bias across different groups. Thereby, this chapter aims to deepen our understanding of how demographic alignment and social bias jointly shape the interpretation of vocal cues in impression formation of trustworthiness.

Chapter 4

Social group bias in vocal trust: Listener predispositions and the limits of speaker intent

4.1. General introduction

Social trust —- the expectation that others will act in a fair, honest, trustworthy and cooperative manner — plays a fundamental role in social cohesion, economic stability, and interpersonal relationships (Freitag & Bauer, 2013; Schilke et al., 2021). However, trust is not distributed equally across societies. For instance, research suggests that levels of social trust vary depending on cultural norms of collectivism (i.e., prioritising societal welfare over individual interests) and individualism (i.e., valuing autonomy and personal uniqueness over group loyalty) (Allik & Realo, 2004). In the UK —- a multicultural society with generally high social trust (Duffy, 2023) — recent trends show a decline in collectivist values and an increased emphasis on individualism, which may encourage more flexible, inclusive trust dispositions (Allik & Realo, 2004; Duffy, 2023; Guo, Zheng, Shen, Huang, & Ma, 2022; Haerpfer et al., 2022a, 2022b). These broader social capital shifts may shape how individuals evaluate others in everyday interactions, especially in low-information contexts. In such contexts, individuals often rely on rapid heuristics to assess others' trustworthiness, drawing on vocal cues, perceived group membership (e.g., age, ethnicity), and personal trust

predispositions (Maltezou-Papastylianou et al., 2025; Skoog Waller, Eriksson, & Sörqvist, 2015). For example, people often rely on social groupings to form first impressions, as identities tied to age, ethnicity, profession, or nationality —- such as "I am a grandparent", "I am British" or "I am an engineer" — shape how individuals perceive and interact with others.

Trustworthiness perceptions —- the extent to which someone is judged as reliable and well-intentioned —- are a core component of social trust and inform decisions about whom to trust, collaborate with, or avoid (Castelfranchi & Falcone, 2010; Hardin, 2002; Mayer et al., 1995). Perceived trustworthiness can be shaped by both vocal and facial modalities (Mileva, 2025), with voice-based interactions being evaluated by both the speaker's characteristics and the listener's trust predispositions (Maltezou-Papastylianou et al., 2025). In the absence of visual cues, listeners rely on vocal characteristics to infer social impressions such as warmth, competence, and intent, which in turn shape trustworthiness judgments (Belin et al., 2019; Rodero et al., 2014; Torre et al., 2016; Yokoyama & Daibo, 2012).

Understanding how voice-based trustworthiness judgments are formed, especially in increasingly digital and multicultural societies, is therefore of growing importance. The present study examines how listeners' trust predispositions, perceived group membership, and speakers' vocal intent shape these judgments, with a specific focus on whether trust tendencies amplify or attenuate in-group biases in vocal evaluations.

4.1.1. Motivation

Voice-based evaluations play a central role in everyday interactions, from job interviews and customer service to courtroom settings and digital technologies (Huang, Markovitch, & Stough, 2024; Kushins, 2014). Yet, little is known about how demographically-diverse speaker-listener relationship and individual predispositions interact to shape vocal trustworthiness judgments. Past research has separately explored social biases, vocal modulation, and trust tendencies, but rarely within an integrated framework (Correll, Hudson, Guillermo, & Earls, 2017; P. W. Linville, Salovey, & Fischer, 1986; Tuomela & Tuomela,

2005). This study addresses that gap by examining how speaker characteristics (age, ethnicity, vocal intent) and listener trust predispositions (generalised and particularised trust) jointly influence voice-based trustworthiness perceptions.

To disentangle these complex influences, the present study adopts a comprehensive design structured around four distinct but interconnected hypotheses. For clarity and ease of reading, the results are presented in three parts. Each part is self-contained, comprising its own background, results, and discussion section, followed by a general discussion and conclusion that synthesise the findings across all parts.

Part 1 (H1)[1] of this study, investigates whether speakers who match listeners in ethnicity and age (i.e., in-group members) are perceived as more trustworthy than out-group speakers. The aim was to explore how social group membership influences trustworthiness judgements during voice-based impression formation.

Part 2 (H2) builds on Part 1 by investigating whether a vocal intent to sound trustworthy enhances trustworthiness ratings for out-group speakers, compared to no such explicit intent (i.e., neutral or natural tone of voice). The aim was to assess whether expressing a trustworthy vocal intent allows out-group speakers to mitigate bias and improve their evaluations.

Lastly, Part 3 examines the role of listeners' trust predispositions in shaping voice-based trustworthiness judgements of speakers who either share or differ from their social group. Specifically, it tests whether generalised trust (H3) correlates positively with trustworthiness ratings overall, and whether particularised trust (H4) is negatively associated with ratings for out-group speakers but positively associated for in-group speakers.

By structuring the present analyses this way, this study clarifies how social categorisation, vocal modulation and individual predispositions operate separately and jointly in shaping voice-based trustworthiness perceptions. This multi-part approach ensures each hypothesis receives focused analysis while preserving the coherence of a single empirical investigation.

All present studies and their hypotheses have been pre-registered on the Open Science Framework platform (`https://osf.io/485a7`).

[1]Note: Hypotheses H1 and H2 in this manuscript correspond to H6.1 and H6.2 on OSF, while H3 and H4 correspond to H5.1 and H5.2 on OSF.

4.2. PART 1: Group membership and cognitive biases

Trustworthiness evaluations are also shaped by social categorisation processes, where individuals assess others based on age, ethnicity, and other identity markers, which can complement personal trust predispositions (Hornsey, 2008; Johnson & Johnson, 1991; Tuckman, 1965; Turner, 2010). Social identity and group membership theories suggest that individuals display in-group biases, favouring those who share their demographic characteristics while being more sceptical of out-group members (Hornsey, 2008; Johnson & Johnson, 1991; Z. Peng, Wang, Meng, Liu, & Hu, 2019; Perrachione, Chiao, & Wong, 2010). However, vocal-based trustworthiness evaluations do not always follow rigid in-group favouritism patterns and may be shaped by perceived speaker traits rather than explicit group categorisation (McGettigan & Lavan, 2023; Mulac & Giles, 1996). For instance, research suggests that vocal cues often serve as proxies for group membership when deeper familiarity is lacking (Dahlbäck et al., 2007; Geiger, Langlinais, & Geiger, 2023; Lavan, Mileva, & McGettigan, 2021; Sebastian & Ryan, 2018).

The similarity-attraction effect, other-race and out-group homogeneity (i.e., the tendency to view in-group members as more diverse than out-group members) exemplify how biases influence trust (Dahlbäck et al., 2007; Perrachione et al., 2010). For example, listeners tend to prefer and develop more positive attitudes toward speakers with accents, vocal styles, or other characteristics similar to their own, as these cues signal in-group membership (Dahlbäck et al., 2007; C. Nass & Lee, 2000; Z. Peng et al., 2019). This bias may lead listeners to favour in-group speakers even when such evaluations are not related to actual expertise or credibility (Fu et al., 2012; Montoya & Horton, 2013). Nonetheless, it remains under-explored whether these biases operate similarly across diverse ethnic and age-based groups in voice perception. Thus, in this part of the present study will examine whether in-group speakers —- based on ethnicity (i.e., white, black, south Asian) and age (i.e., younger and older adults) —- are perceived as more trustworthy than out-group speakers, extending past literature on how group membership and social identity influences voice-based trust-related evaluations.

4.2.1. Vocal perceptions across age-group membership and stereotypes

Research on voice perception and ageing suggests that age-related vocal cues play a significant role in trustworthiness evaluations. While older voices are often associated with competence and wisdom, there is also a tendency to perceive them as less powerful and expressive compared to younger voices (Löckenhoff et al., 2009; McGettigan & Lavan, 2023; Mergler & Goldstein, 1983; Montepare, Kempler, & McLaughlin-Volpe, 2014; Schirmer et al., 2020). By contrast, younger voices are frequently linked to warmth, sociability, and engagement, traits that positively influence trustworthiness evaluations (Montepare et al., 2014; Schirmer et al., 2020). Consequently, these factors were shown to negatively affect trustworthiness ratings toward older speakers in contrast to their younger counterparts (McGettigan & Lavan, 2023; Schirmer et al., 2020). Nonetheless, it was also shown that older listeners tended to rate voices as more trustworthy overall, than younger listeners (Schirmer et al., 2020).

While ageing may diminish perceptions of power and adaptability, older voices are universally associated with wisdom and life experience (Bangen, Meeks, & Jeste, 2013; Montepare & Zebrowitz, 1998; Scheibe, Kunzmann, & Baltes, 2009; Thomas, 2004; Zebrowitz & Montepare, 2013). Cross-cultural research has indicated that, despite expectations of functional decline with age, ageing may enhance trustworthiness perceptions in knowledge-based contexts, across diverse cultures (Löckenhoff et al., 2009). Nevertheless, listeners' cognitive ageing has been associated with misclassification of speakers' vocal expression, which may in turn, influence their evaluations (Maltezou-Papastylianou et al., 2022). Given all these conflicting aspects, it remains unclear whether younger and older speakers experience the same in-group biases observed in ethnic-based trustworthiness judgments.

4.2.2. Vocal perceptions across ethnic-group membership and stereotypes

Ethnic-based trust biases in voice perception are often mediated by accent clarity, linguistic expectations such as appropriate pronunciation and articulation, and stereotype-based judgments (Hanzlíková & Skarnitzl, 2017; Sharma, Levon, & Ye, 2022). Prior research suggests that native English speakers tend to be rated as more trustworthy than non-native speakers in English-speaking countries, regardless of listener ethnicity (Geiger et al., 2023). However, it remains unclear whether these biases are due to explicit ethnic categorisation, cultural expressiveness familiarity or implicit preferences for linguistic fluency and speech clarity (Coupland & Bishop, 2007).

According to stereotype content theory (Cuddy et al., 2008, 2009; Fiske et al., 2007), different ethnic groups are evaluated along the dimensions of warmth and competence, which may influence trustworthiness judgments. For example, south Asian speakers have been rated lower in perceived competence and speech prestige (Coupland & Bishop, 2007; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022). Meanwhile, black speakers may experience trust biases that vary depending on how closely their speech aligns with expected native English speaker norms of a country (Kushins, 2014).

The present study will explore whether white speakers are consistently rated as more trustworthy than black and south Asian speakers, irrespective of listener ethnicity. Furthermore, it will investigate whether intent influences trust evaluations beyond simple ethnic categorisation.

4.2.3. Methods

4.2.3.1. Ethics declaration

All procedures performed in this study were approved by the Ethics Subcommittee 3 of the University of Essex (ETH2223-0254) 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.

4.2.3.2. Stimuli

Thirty-six speakers from three ethnic backgrounds were recruited to record three seven-syllable sentences each (e.g., "I can drive you if you want."; "Hello, I arrived early."; "I will give you a lift home."). Speaker demographics are detailed in Table 4.1. All speakers reported normal hearing and were compensated for their participation. Speakers were instructed to deliver each sentence twice: once using their natural —- neutral tone of voice —- without any particular social intention, and once while deliberately trying to convey trustworthiness to gain the listener's trust. To minimise experimenter bias, they were not given any specific instructions on how to express trustworthiness. For more information on the material and recording procedure see (Maltezou-Papastylianou et al., 2024b).

Table 4.1: Descriptive statistics of speaker demographics

Ethnicity	**Age-group**	**Sex**	**N**	**Mean age (years)**	**Age range (years)**	**SD**
White	Younger	Female	3	28.3	22 - 36	7.1
		Male	3	32.3	25 - 43	9.5
	Older	Female	3	76	67 - 87	10.2
		Male	3	67.3	62 - 76	7.6
Black	Younger	Female	3	26.3	22 - 31	4.5
		Male	3	32.3	24 - 37	7.2
	Older	Female	4	61	60 - 62	0.8
		Male	2	62	61 - 63	1.4
South Asian	Younger	Female	3	29.7	22 - 37	7.5
		Male	3	29	22 - 34	6.2

Continued on next page

Table 4.1: Descriptive statistics of speaker demographics (Continued)

Ethnicity	Age-group	Sex	N	Mean age (years)	Age range (years)	SD
	Older	Female	3	65	62 - 68	4.2
		Male	3	73	61 - 90	15.1

4.2.3.3. Participants

A total of 288 English-speaking adults in the UK were recruited, balanced across ethnicity (N = 96 per group —- white, black and south Asian), age (N = 48 per group — younger and older than 60 years) and sex (N = 24 per group). To ensure a clear age distinction, younger adults were defined as those aged 45 and below, while older adults were classified as 60 years and above. This was done to create a wide-enough age gap between the groups for comparison. See Table 4.2 for more details on listener demographics. An a-priori power analysis was conducted using the software program G∗Power (Faul et al., 2009, 2007). The goal of the current study was to obtain a test power of 95% to detect a medium effect size at the standard alpha error probability of .05.

All younger adult listeners and older white listeners were recruited online through Prolific (Prolific, 2014); most of the older black and older south Asian listeners were recruited through Prime Panels (Chandler et al., 2019; CloudResearch, 2015), a participant recruitment platform that aggregates several market research panels. All listeners reported normal hearing. Throughout this paper, the terms participants / listeners may be used interchangeably.

Table 4.2: Descriptive statistics of participant demographics

Ethnicity	Age-group	Sex	N	Mean age (years)	Age range (years)	SD
White	Younger	Female	24	30.46	20 - 43	7.23

Continued on next page

Table 4.2: Descriptive statistics of participant demographics (Continued)

Ethnicity	Age-group	Sex	N	Mean age (years)	Age range (years)	SD
		Male	24	28.63	18 - 44	8.71
	Older	Female	24	66.38	60 - 80	5.98
		Male	24	64.04	60 - 72	3.71
Black	Younger	Female	24	25.08	20 - 35	5.51
		Male	24	27.96	19 - 40	5.73
	Older	Female	24	65.79	60 - 78	5.16
		Male	24	64.08	60 - 81	4.43
South Asian	Younger	Female	24	22.63	19 - 33	3.59
		Male	24	23.00	18 - 39	5.99
	Older	Female	24	66.17	60 - 78	4.86
		Male	24	65.67	60 - 78	5.16

4.2.3.4. Procedure

Qualtrics software was used to screen and direct participants to the online study hosted on a custom-made PHP web app. Participants first had to complete a short survey on social trust attitudes, then were asked to rate the auditory stimuli, and lastly, answered a set of multiple-choice questions identifying speakers' age and ethnicity (e.g., "What is the ethnicity of this speaker?"). The social trust questions were taken from the World Values Survey (Haerpfer et al., 2022a, 2022b), due to their extensive evaluation on validity and reliability on a global scale (Freitag & Bauer, 2013; H. H.-S. Kim, 2018; Newton & Zmerli, 2011).

During the rating task, each participant listened to stimuli spoken by twelve speakers, each speaker uttering three sentences twice with differing intent (i.e., 72 audio files in total). The presented audio stimuli were randomised using the Fisher-Yates Shuffle algorithm (Eberl, 2016). After each recording, participants were asked to respond to the statement

"This speaker sounds trustworthy" to indicate trustworthiness ratings on a Likert scale ranging from 1 (strongly disagree) – 7 (strongly agree). On average, listeners took 20 minutes to complete the study.

4.2.4. Results of Part 1

Since H1 examined whether in-group speakers would be perceived as more trustworthy than out-group speakers, the interactions involving group membership, such as speaker x listener ethnic and age groups were of particular interest. Thus, a mixed ANOVA was used instead of the pre-registered series of t-tests to account for the multiple factors and potential interactions. In terms of reporting the ANOVA results, omega-squared (ω^2) was used as an indicator of effect size. Even though effect sizes are context-dependent, an ω^2 = .01 (i.e., 1% of variance explained) is typically considered a small effect in the literature, an ω^2 = .06 a medium effect, and ω^2 = .14 a large effect (Field, 2018; Kirk, 1996). P-values were adjusted with the Greenhouse-Geisser correction when sphericity was violated, and all post-hoc comparisons were Holm-Bonferroni corrected (Abdi, 2010).

4.2.4.1. How speaker-listener group membership affect ratings of perceived trustworthiness

The mixed ANOVA analysis included speaker ethnicity (white, black, south Asian) and speaker age (older, younger) as within-subject variables and listener ethnicity (white, black, south Asian) and age (older, younger) as between-subject variables.

The significant main effect of speaker age-group, $F(1, 282) = 258.85$, $p < .001$, $\omega^2 = .10$, showed higher trustworthiness ratings in response to materials spoken by younger speakers ($M_{diff} = 0.44$). A main effect of speaker ethnicity, $F(1.90, 532.68) = 27.90$, $p < .001$, $\omega^2 = .02$, was followed up with post-hoc comparisons, which showed higher trustworthiness ratings for white speakers compared to black speakers ($M_{diff} = 0.20$, $SE = 0.03$, $p < .001$) and south Asian speakers ($M_{diff} = 0.17$, $SE = 0.03$, $p < .001$); no significant results between

black and south Asian speakers ($p = .26$). No significant main effects were found for listener ethnicity ($p = .23$) and age-group ($p = .53$).

Significant interactions were shown only for speaker x listener ethnicity, $F(1.90, 532.68) = 3.55$, $p = .008$, $\omega^2 = .003$, speaker ethnicity x listener age-group, $F(1.90, 532.68) = 7.46$, $p < .001$, $\omega^2 = .004$, and speaker ethnicity x age-group, $F(1.96, 552.06) = 3.86$, $p = .02$, $\omega^2 = .001$.

To answer H1, the significant two-way interaction of speaker x listener ethnicity was followed up with post-hoc comparisons. Post-hoc results showed no significant difference in white listeners' ratings toward white vs black speakers ($p = .12$), white vs south Asian speakers ($p = 1.00$) and black vs south Asian speakers ($p = 1.00$). On the other hand, black listeners, rated white speakers marginally higher on perceived trustworthiness than black speakers ($M_{diff} = 0.16$, $SE = 0.05$, $p = .05$), and significantly higher than south Asian speakers ($M_{diff} = 0.24$, $SE = 0.06$, $p = .001$). No significant difference was observed when black listeners rated black vs south Asian speakers ($p = 1.00$). South Asian listeners rated south Asian speakers significantly lower on trustworthiness ratings than white speakers ($M_{diff} = -0.23$, $SE = 0.06$, $p = .003$), but no significant difference compared to black speakers ($p = 1.00$). South Asian listeners rated white speakers as significantly more trustworthy than black speakers too ($M_{diff} = 0.31$, $SE = 0.05$, $p < .001$).

In summary, the current analysis revealed that younger speakers were rated as more trustworthy overall than older speakers, with a large effect size. There also appears to be a small but significant effect towards a general bias favouring white speakers in terms of perceived trustworthiness, compared to black and south Asian speakers, regardless of the listeners' ethnicity.

Considering these findings, a chi-square test was also conducted to assess whether listeners could accurately identify the speakers' ethnicity (white, black, south Asian, other) and age. The results showed a significant association between the listeners' guesses and the true ethnicities of the speakers, $\chi^2 (6, N = 3444) = 1468.79$, $p < .001$, indicating that listeners were able to identify the speakers' ethnicities significantly better than chance. Speakers from a white ethnic background were correctly identified 77.29% of the time,

black 37.66% (where 35.48% of the time they were recognised as white and 18.92% as south Asian) and south Asian 61.93%. The results also showed a significant association between the listeners' guesses and the true age-group of the speakers, χ^2 *(1, N = 3444) = 88.75, p < .001*. However, a bias toward younger age classifications was observed, as listeners predominantly categorised speakers as younger than 60 years old —- speakers from a younger age-group were correctly identified 97.97% of the time, while those from an older age-group were correctly identified only 9.52% of the time.

4.2.5. Discussion of Part 1

The expectation that in-group speakers would be perceived as more trustworthy than out-group speakers revealed mixed findings in Part 1 of the present study. Younger speakers were rated as more trustworthy than their older counterparts, regardless of listener age-group. Moreover, ethnicity did not show strong in-group effects — instead, white speakers were rated as more trustworthy than black and south Asian speakers, regardless of the listener's own ethnicity. These findings provide a more nuanced understanding of how age and ethnicity interact in vocal trust judgments and suggest that perceptions of trustworthiness are shaped by factors beyond simple in-group biases. Implications of findings are discussed in the following paragraphs.

4.2.5.1. What makes younger-sounding voices more trustworthy?

The consistent preference for younger-sounding voices across listener age-groups challenges the assumption that people generally favour their in-group. One possible explanation is that younger voices are often associated with warmth and energy, traits that contribute positively to trust-related evaluations (McGettigan & Lavan, 2023; Montepare & Zebrowitz, 1998; Mulac & Giles, 1996; Schirmer et al., 2020), whereas older voices have been linked to a more reserved or passive demeanour (McGettigan & Lavan, 2023; Mergler & Goldstein, 1983). Indeed, the current voice files were also rated for warmth in (Maltezou-Papastylianou,

Scherer, & Paulmann, 2023) and results suggest higher ratings for impressions of warmth for younger speakers across all ethnic groups compared to their older counterparts.

Since this study did not require speakers to demonstrate expertise —- an attribute often ascribed to older voices (Mergler & Goldstein, 1983; Montepare et al., 2014), judgments may have been more influenced by approachability and expressiveness rather than perceived authority (Löckenhoff et al., 2009). Had the study required speakers to demonstrate knowledge or competence in an intellectual context, older voices might have been perceived as more trustworthy due to the stereotype linking age with wisdom and experience (Bangen et al., 2013; Montepare & Zebrowitz, 1998; Scheibe et al., 2009; Thomas, 2004; Zebrowitz & Montepare, 2013). However, considering that the same voice material were also rated for competence in (Maltezou-Papastylianou et al., 2023) and younger adult speakers were still rated higher than their older counterparts, this suggests a contextual effect: the current stimuli did not require speakers to display intellectually-relevant traits, but instead leaned more toward social first impressions. This could explain why there was a general bias in favour of younger-sounding voices, as listeners may have associated them with greater engagement and sociability under the theme of general first impressions (Maltezou-Papastylianou et al., 2025). Hence, the current findings suggest that trustworthiness perceptions are context-dependent, shifting based on which perceived personality traits and vocal cues are more relevant in a given setting.

Alternatively, listeners may have struggled to correctly identify speaker age, minimising the "in-group/out-group recognition". Indeed, materials from older speakers were misclassified as younger 90.48% of the time, whereas younger voices were correctly identified in 97.97% of cases. This suggests that older listeners may not have consciously recognised older speakers as belonging to their in-group. If an older listener believes they are evaluating a younger-sounding speaker, their judgment might be influenced more by perceived vocal qualities associated with different age-groups (e.g., faster speech rate often associated with younger adults) rather than shared age-group membership (Belin et al., 2019; Rodero et al., 2014; S. M. Smith & Shaffer, 1995; Torre et al., 2016; Yokoyama & Daibo, 2012). This suggests that in-group bias may have been diminished or absent.

A broader question is why these misclassifications occurred. While the present study did not examine acoustic ageing cues, it is possible that certain older speakers did not exhibit factors typically associated with ageing (McGettigan & Lavan, 2023; Montepare et al., 2014; Schirmer et al., 2020), leading to ambiguity in age perception. For instance, past research has shown that voices with perceived faster speech rates, higher pitch and greater pitch variability were perceived as younger and more expressive than their actual chronological age and vice versa (Skoog Waller et al., 2015). This assumption led us to consult the mean acoustic cue results per age-group as reported by (Maltezou-Papastylianou et al., 2024a, 2024b), which were based on the same audio material. These results confirmed that older adults exhibited greater pitch variability — suggesting higher vocal expressiveness —- despite speaking more slowly and with a lower average pitch than their younger counterparts. As such, this pattern could also reflect shifting social perceptions of ageing, where listeners may subconsciously categorise voices as younger unless stronger ageing cues are present (Skoog Waller et al., 2015). Nonetheless, visual cues, which were not available in this audio-only study, can significantly influence age judgements (Lim et al., 2022; Mileva, 2025). The present design was intentionally restricted to audio to reflect voice-only contexts (e.g., phone calls or tele-services), where listeners must rely solely on vocal information.

Future research could investigate whether such misperceptions are solely driven by strong acoustic markers of ageing or broader cultural attitudes and expectations toward vocal ageing. Such an approach would help clarify whether listeners' misclassifications were acoustically driven or socially motivated. Another important avenue to explore is whether listeners' perception of how old a speaker sounds is more important in age-categorisation than a speaker's actual age.

Conversely, the current pattern of responses suggests that younger listeners may have shown a preference for younger-sounding speakers. This trend, if replicated, would be consistent with similarity-attraction bias, as they are more likely to interact with peers in their everyday lives. This aligns with previous research suggesting that social exposure reinforces trust-related biases and age-related stereotypes (Dahlbäck et al., 2007; Löckenhoff et al., 2009; C. Nass & Lee, 2000; Z. Peng et al., 2019). However, the asymmetry in findings

—- where older listeners did not show a strong in-group preference — suggests that familiarity with one's in-group alone does not fully explain the observed effects, which raises further questions for future research; perhaps these older listeners had more exposure in their daily lives with the younger generation (e.g., their younger adult children or grandchildren) and exhibit a noticeable shift in their familiarity spectrum, or may not categorise themselves as being in the older age category. Moreover, future studies could investigate whether explicitly providing speaker age information alters trustworthiness ratings, helping to disentangle whether perceived rather than actual age-group membership drives trust-related judgments.

These findings highlight the complexity of vocal trustworthiness perceptions, suggesting that listener judgments may be shaped more by perceived vocal traits associated with age —- such as pitch or expressiveness — than by actual or consciously recognised speaker age. The frequent misclassification of older voices further supports the idea that trustworthiness evaluations may rely on perceived rather than actual demographics. Future research should investigate which specific vocal cues drive these age-related impressions and how they shape social judgments in voice-based interactions.

4.2.5.2. Why were white speakers rated as more trustworthy?

Although ethnicity did not lead to strong in-group biases, white speakers were consistently rated as more trustworthy than black and south Asian speakers, regardless of the listeners' own ethnic background. This finding appears to align with research indicating that native English speakers in English-speaking countries are often perceived as more credible and desirable by both native and non-native listeners, particularly in professional contexts (Fu et al., 2012; Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017). However, it is important to clarify that this does not mean native English speakers are exclusively white. Rather, all white speakers in this study were native, English-born, and their vocal attributes such as pronunciation and articulation may have aligned more closely with listeners' expectations for fluent, white native British English, potentially contributing to their higher trustworthiness ratings. While ethnicity remains a key factor in trust-related judgments, which was examined in the present study and will be discussed in more detail shortly, future research should

systematically examine how linguistic fluency and potential accent familiarity interact with ethnic categorisation in shaping listener evaluations. Furthermore, while theories of identity in social groups (Hornsey, 2008; Johnson & Johnson, 1991; Tuckman, 1965; Turner, 2010), other-race and out-group homogeneity predict stronger biases favouring one's ethnic in-group (Correll et al., 2017; P. W. Linville et al., 1986; Tuomela & Tuomela, 2005), the present findings suggest that other vocal cues, such as age and intent, may override ethnic biases in trustworthiness judgments.

An interesting discrepancy emerged in how accurately listeners identified speaker ethnicity. Black speakers were significantly less likely to be correctly identified (37.66%) compared to white (77.29%) and south Asian (61.93%) speakers. Notably, black speakers were misclassified as white (35.48%) or south Asian (18.92%) almost as frequently as they were correctly identified, suggesting that listeners had greater difficulty categorising black voices by ethnicity. Given this high misclassification rate, it is likely that trustworthiness evaluations were shaped more by perceived vocal cues than by accurate ethnic categorisation. The fact that black speakers were frequently perceived as white yet still rated lower in trustworthiness undermines the assumption that perceived group membership alone drives listener evaluations. Instead, trustworthiness judgments may have been influenced by vocal features that conflicted with listeners' implicit expectations for how white, native English speakers are "supposed" to sound (Correll et al., 2017; Geiger et al., 2023; P. W. Linville et al., 1986). From an expectancy violation perspective (Burgoon, 2015; Burgoon & Hubbard, 2005), this mismatch between how a speaker was categorised and how they actually sounded may have created dissonance or uncertainty for listeners, thereby reducing perceived trustworthiness. If listeners unconsciously expected certain vocal traits to align with specific social identities, then encountering a voice that defied those expectations — for example, a black speaker who was misclassified as white but whose prosody or articulation did not conform to stereotypical white British norms — may have been interpreted as incongruent or ambiguous, triggering less favourable evaluations. Thus, while ethnicity remains a salient cue, it likely interacts with deeper vocal and social-cognitive processes, making trustworthiness judgments more complex than categorical group biases alone.

In contrast, south Asian speakers were more accurately identified than black speakers but received similar lower trustworthiness ratings than white speakers. One possible explanation is that, south Asian voices may have more consistently activated existing stereotypes. Prior research has shown that south Asian accents, even when subtle, are often negatively stereotyped in professional and social contexts and rated lower in prestige, pleasantness, and attractiveness in native English-speaking countries like the UK (Coupland & Bishop, 2007; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022). As such, higher recognition accuracy may have reinforced rather than disrupted these stereotypes, contributing to consistently lower trustworthiness evaluations. This pattern suggests that trustworthiness judgments are not simply a product of ethnic categorisation, but emerge from an interaction between perceived ethnicity and linguistic features such as accent strength, speech fluency, and prosody. These findings raise important real-world implications for contexts like hiring, customer service, and other professional settings, where biases based on pronunciation or articulation differences may disadvantage south Asian speakers regardless of their actual communicative clarity. Future studies should explore whether interventions, such as listener training or exposure to a wider range of speech fluency types, can reduce these biases, and further investigate which specific acoustic features may drive negative evaluations in this group. The current findings also raise the question of whether speaker intent, when deliberately expressed through vocal cues, can mitigate these biases and improve out-group evaluations. The analyses in Part 2 explore this possibility by examining the role of vocal intent in shaping trustworthiness perceptions across demographic groups.

4.3. PART 2: Can vocal intent mitigate out-group bias?

Building on the observations from the first analysis in Part 1, can out-group speakers (based on ethnicity and age) actively influence trustworthiness perceptions through vocal intent? Prior studies have shown that deliberate vocal adjustments, such as faster speech rates, higher pitch and expressiveness, were often interpreted as signals of credibility, engagement

and younger age with positive association to trustworthiness, warmth, honesty and reduced dominance (Belin et al., 2019; Maltezou-Papastylianou et al., 2024a; McAleer et al., 2014; Rodero et al., 2014). Expectancy violation theory further suggests that out-group speakers who defy negative stereotypes, by deliberately signalling trustworthiness, may receive more favourable evaluations than anticipated (Burgoon, 2015; Burgoon & Hubbard, 2005).

However, the effectiveness of vocal intent may depend on the listener's pre-existing biases (S. K. Kang & Bodenhausen, 2015; Leongómez et al., 2021; Maltezou-Papastylianou et al., 2025). If expectations about an out-group speaker's trustworthiness are strongly negative, an assumption could be that vocal intent may have limited effect (Keeley, English, Irons, & Henslee, 2013). Conversely, when expectations are more flexible or ambiguous, vocal intent may serve as a stronger cue for trustworthiness. This Part 2 analyses investigates whether speakers who explicitly signal a trustworthy intent receive higher trustworthiness ratings than those with no such explicit intent, particularly among out-group speakers.

4.3.1. Results of Part 2

To evaluate the pre-registered H2 hypothesis this study employed a mixed ANOVA. In terms of reporting the ANOVA results, omega-squared (ω^2) was used as an indicator of effect size. Even though effect sizes are context-dependent, an ω^2 = .01 (i.e., 1% of variance explained) is typically considered a small effect in the literature, an ω^2 = .06 a medium effect, and ω^2 = .14 a large effect (Field, 2018; Kirk, 1996).

4.3.1.1. How out-group speaker intent affects ratings of perceived trustworthiness

To test H2 —- posing that out-group speakers expressing a trustworthy intent would be rated as more trustworthy than those using a neutral tone — a mixed ANOVA was conducted. Interactions between speaker intent and speaker-listener group membership (based on age and ethnicity) were of particular interest. Although a series of t-tests was originally pre-registered, a mixed ANOVA was chosen instead to account for the multiple factors involved and to

examine both main effects and interactions. The analysis included speaker intent (neutral, trustworthy), speaker ethnicity (white, black, south Asian), and speaker age (younger, older) as within-subject variables, and listener ethnicity and age-group as between-subject variables. Greenhouse-Geisser corrections were applied when sphericity assumptions were violated, and post-hoc comparisons were adjusted using the Holm-Bonferroni method (Abdi, 2010).

Significant main effects were identified for speaker intent, $F(1, 282) = 46.15$, $p < .001$, $\omega^2 = .03$, with high perceived trustworthiness ratings for voices that were expressed with a trustworthy intent ($M_{diff} = 0.27$). The main effect of speaker age-group, $F(1, 282) = 119.21$, $p < .001$, $\omega^2 = .09$, showed higher trustworthiness ratings in response to materials spoken by younger speakers ($M_{diff} = 0.50$). A main effect of speaker ethnicity, $F(1.92, 540.52) = 16.51$, $p < .001$, $\omega^2 = .02$, was followed up with post-hoc comparisons, which showed higher trustworthiness ratings for voices of white speakers compared to black speakers ($M_{diff} = 0.12$, $SE = 0.05$, $p = .02$) and south Asian speakers ($M_{diff} = 0.29$, $SE = 0.06$, $p < .001$); voices of black speakers were rated higher on perceived trustworthiness over south Asian speakers ($M_{diff} = 0.17$, $SE = 0.05$, $p < .001$). Non-significant main effects were shown for listener ethnicity ($p = .30$) and age-group ($p = .65$).

Significant interactions were found for speaker x listener age-groups, $F(1, 282) = 5.20$, $p = .02$, $\omega^2 = .003$, speaker ethnicity x age-group, $F(1.99, 560.86) = 6.78$, $p = .001$, $\omega^2 = .006$, speaker ethnicity x intent, $F(1.98, 557.77) = 3.71$, $p = .03$, $\omega^2 = .002$, speaker age-group x intent, $F(1, 282) = 23.74$, $p < .001$, $\omega^2 = .01$, speaker ethnicity x age-group x intent, $F(1.96, 551.57) = 7.07$, $p < .001$, $\omega^2 = .005$. Finally, there was also a four-way interaction between speaker ethnicity x age-group x intent x listener age-group, $F(1.96, 551.57) = 3.50$, $p = .03$, $\omega^2 = .002$.

To answer H2, which tested for vocal intent effects on out-group speakers, the significant four-way interaction between speaker intent x ethnicity x age-group x listener age-group was followed up with post-hoc analyses. Results showed nuanced effects between listener and out-group speaker characteristics on trustworthiness ratings. No significant variations between the two intent conditions (neutral vs trustworthy intent) were shown when younger listeners rated older white and older south Asian speakers' perceived trustworthiness (p

= *1.00*). However, ratings did vary significantly for older black speakers, with higher trustworthiness rating toward trustworthy intent over neutral intent, M_{diff} = *0.56, SE = 0.12, p < .001*. In terms of older listeners' ratings toward out-group speakers' perceived trustworthiness, younger black speakers with a trustworthy intent were rated significantly higher than with neutral intent, M_{diff} = *0.45, SE = 0.12, p = .04*. Although not significant, a marginal effect was shown for younger white (M_{diff} = *0.39, SE = 0.11, p = .054*) and younger south Asian (M_{diff} = *0.38, SE = 0.11, p = .087*) speakers, with higher ratings toward trustworthy intent over neutral. In short, findings showed that vocal intent primarily increased trustworthiness ratings for black speakers, particularly younger ones rated by older listeners.

4.3.2. Discussion of Part 2

While the analysis from Part 1 showed that trustworthiness ratings reflected group membership biases, especially in relation to ethnicity than age, here a shift can be seen: once vocal intent was explored more directly, trustworthiness ratings showed the sensitivity of listeners toward how speakers sound rather than who they are. Specifically, a speaker's deliberate attempt to convey trustworthiness significantly improved trustworthiness ratings overall, aligning with past findings on vocal modulation (Belin et al., 2019; Leongómez et al., 2021; Maltezou-Papastylianou et al., 2024a; Rodero et al., 2014). Yet, the most salient effects emerged between speaker intent and age-groups, particularly for younger black speakers evaluated by older listeners irrespective of ethnicity. This marks a shift from the ethnicity-driven biases in Part 1 analyses to a more nuanced, cue-driven mechanism of trust in Part 2 analyses. Thus, suggesting that when vocal intent is introduced, listeners may weigh social categories differently, prioritising expressiveness over group affiliation, as discussed in the following paragraphs.

4.3.2.1. Why did vocal intent have a stronger effect for younger out-group speakers?

The most consistent effects of vocal intent in out-groups were observed for younger speakers rated by older listeners. This aligns with the broader finding that older listeners generally rated younger voices as more trustworthy, suggesting they may have been more receptive to vocal intent cues when evaluating younger speakers. Therefore, since faster speech rates and higher pitch have been linked to more positive perceptions of credibility, engagement, and trustworthiness (Rodero et al., 2014; S. M. Smith & Shaffer, 1995; Yokoyama & Daibo, 2012), it raised the need to consult the acoustic cue findings reported for the present materials in Maltezou-Papastylianou et al. (2024a, 2024b). Those results showed that the younger speakers of the current study already possess such vocal qualities naturally associated with trustworthiness; for instance, a combination of faster speech rate, higher perceived pitch and greater pitch variability, with lower and perceptually more expressive voice quality signals of harmonics-to-noise ratio, shimmer and long-term average spectrum. By explicitly attempting to sound more trustworthy, they had amplified these cues, making their intent more salient and effective to listeners (Maltezou-Papastylianou et al., 2024a, 2024b).

Conversely, vocal intent had limited impact on younger listeners' ratings of older speakers. This suggests that older voices may have been perceived as less malleable in their ability to sound deliberately more trustworthy (McGettigan & Lavan, 2023; Mergler & Goldstein, 1983; Montepare et al., 2014). Prior work suggests that age-related vocal characteristics, such as slower speech rate and lower pitch, are often associated with authority rather than warmth and expressiveness (Belin et al., 2019; Maltezou-Papastylianou et al., 2025; Mulac & Giles, 1996; Rodero et al., 2014), which may have made older speakers' vocal adjustments less effective in shaping listeners' trustworthiness judgements.

These findings suggest that older listeners were not simply responding to group similarity but to how voices expressed trustworthiness. When younger speakers amplified qualities like warmth or engagement through vocal intent, older listeners were especially responsive. This marks a shift from static identity-based biases (as seen in Part 1) toward more dynamic, cue-based evaluations.

4.3.2.2. Why did vocal intent boost trustworthiness perceptions only for some ethnic groups?

Although vocal intent boosted trustworthiness ratings across all younger speakers, its effect was strongest for younger black speakers rated by older listeners, and for older black speakers rated by younger listeners. This pattern may be interpreted through the lens of expectancy violation theory in combination with social categorisation ambiguity. For instance, a potential explanation is that listener expectations for black speakers were more ambiguous since they were frequently misclassified, which according to past research on out-group homogeneity and related biases, ambiguity in social categorisation leads to greater reliance on individuating cues such as vocal intent (Correll et al., 2017; S. K. Kang & Bodenhausen, 2015; P. W. Linville et al., 1986; Z. Peng et al., 2019). The earlier findings of the present study showed that black speakers were frequently misclassified as white, yet they were still rated lower in trustworthiness. This suggests that listeners may not have based their judgments purely on perceived ethnic group membership. Instead, evaluations could have been influenced by vocal characteristics that did not fully match listeners' expectations for white native English speakers (Fu et al., 2012; Geiger et al., 2023; Tuomela & Tuomela, 2005). Because black speakers were misclassified more often, listener expectations for how they "should" sound may have been weaker or more variable, creating a greater opportunity for expectancy violations to shift trustworthiness ratings. As such, when black speakers deliberately signalled a trustworthy intent, this may have been perceived as an unexpected but positive violation of expectations, amplifying its impact on ratings (Burgoon, 2015; Burgoon & Hubbard, 2005). In other words, since black speakers were not consistently identified as belonging to a single ethnic category in the present study, vocal intent may have served as a particularly strong, individuating trustworthiness cue in resolving listener uncertainty.

In contrast, older south Asian speakers were more accurately identified, and their trustworthiness ratings remained relatively stable, even with differing vocal intent. If trust-related factors were already well-established for south Asian speakers, there may have been less room for expectancy violation effects to occur. However, the marginal increase

in trustworthiness ratings for younger south Asian speakers suggests that vocal intent was still beneficial, albeit to a lesser extent than for black speakers. One possible explanation is that listeners may rely more on linguistic characteristics such as speech pronunciation and articulation when evaluating trustworthiness in south Asian voices, meaning vocal intent alone may not have been enough to override pre-existing evaluations (Coupland & Bishop, 2007; Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022). If trustworthiness ratings for south Asian speakers were influenced more by perceived accent clarity or fluency as suggested by past studies (Coupland & Bishop, 2007; Sharma et al., 2022), then simply attempting to sound more trustworthy may not have significantly altered listener judgments, albeit heading to a more positive direction. Future research should examine whether vocal adjustments in articulation and prosody, in conjunction with vocal intent, may boost trustworthiness perceptions in south Asian speakers further.

In addition, the stereotype content theory may further explain these asymmetries in the current results between black and south Asian speakers (Cuddy et al., 2008, 2009; Fiske et al., 2007). This theory suggests that different ethnic groups are perceived along the dimensions of warmth and competence, which may interact with vocal expressiveness in shaping trustworthiness perceptions (Cuddy et al., 2008, 2009; Hanzlíková & Skarnitzl, 2017). For instance, as it was previously mentioned, past work suggests that south Asian accents, even when subtle, are often associated with lower prestige and perceived competence (Coupland & Bishop, 2007; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022), potentially making vocal intent less effective in enhancing trustworthiness perceptions. In contrast, black speakers — whose speech patterns may be perceived as more aligned with the white, native English speakers of the present study — may have had more room for vocal intent to positively influence their perceived trustworthiness.

For younger white speakers, vocal intent also had a marginal boost on trustworthiness ratings, while no such effect was observed for older white speakers, possibly due to a ceiling effect (Keeley et al., 2013); since white speakers were already rated as the most trustworthy group overall, there may have been limited room for additional increases based on vocal intent. This further aligns with research on expectancy confirmation and halo effect,

where previous positive evaluations (e.g., on natural speaking tone) are reinforced rather than dramatically altered by additional individuating cues such as an explicit vocal intent (Burgoon, 2015; Burgoon & Hubbard, 2005; Keeley et al., 2013; Sebastian & Ryan, 2018).

Considering the above findings, while vocal intent does improve trustworthiness ratings overall, it does not appear to operate equally as a universal trust-enhancing cue for improving out-groups' perceived trustworthiness. Rather, out-group speakers' vocal intent seems to interact with listeners' initially rated expectations, speaker perception, acoustic and linguistic traits and social categorisation ambiguity. When listener expectations are more flexible or uncertain toward an out-group —- such as for younger and older black speakers — vocal intent may serve as a salient and effective trustworthiness cue. Conversely, when expectations are more rigid or pre-established (e.g., south Asian speakers being more accurately identified and rated the lowest), intent alone may not override ingrained perceptions as effectively as one could expect. Rather than uniformly increasing trustworthiness, vocal intent may serve as an amplifying factor, enhancing positive biases when expectations are flexible but failing to override rigidly negative biases. Future studies should experimentally manipulate vocal intent (e.g., intonation, articulation, speech rate) to determine which aspects of vocal expressiveness drive trustworthiness judgments more effectively across different ethnic groups. Furthermore, cross-cultural studies could explore whether these vocal intent effects vary based on broader linguistic norms and implicit cultural biases.

The present findings on vocal intent have important implications for speech training programs and diversity-focused hiring assessments. If vocal intent is more effective when listener expectations are flexible, then training speakers to modulate vocal intent cues may be useful in professional settings where bias reduction is critical. Future work should explore whether interventions focusing on vocal delivery in different scenarios can mitigate trustworthiness biases across demographic groups, particularly in contexts such as job interviews, political speeches, and legal testimonies.

In sum, while Part 1 analyses of this study found clear group-based preferences, follow-up analyses show that these can be moderated —- or even reversed —- when vocal intent provides stronger, more individuating cues. Listeners appear more willing to update their

judgments when category boundaries are ambiguous or expectations are less fixed, especially for speakers who deliberately signal warmth or trustworthiness.

4.4. PART 3: Trust predispositions — Generalised vs particularised trust

Beyond group membership (Part 1) and vocal intent (Part 2), what role could individual trust predispositions have potentially played in the current findings? Trust predispositions serve as socio-cognitive heuristics that guide trust-related judgments in daily interactions (Castelfranchi & Falcone, 2010; Glanville & Shi, 2020; Hardin, 2002), meaning that in the absence of prior knowledge about a speaker, trust predispositions can serve as cognitive shortcuts, shaping initial judgments of vocal trustworthiness. This reliance on pre-existing trust tendencies has direct implications on voice-based judgements to guide trustworthiness evaluations of unfamiliar speakers (Bauer & Freitag, 2018; Hardin, 2002; Jiang et al., 2020). These predispositions are often classified into generalised trust —- trust extended broadly to strangers and those outside one's immediate group (i.e., out-group trust) — and particularised trust, which is reserved for close, familiar individuals within one's group (i.e., in-group trust) (Bauer & Freitag, 2018; Hardin, 2002; Schilke et al., 2021; Uslaner, 2002). These two trust dispositions have been studied extensively, but their relationship remains debated.

Some studies suggest that generalised and particularised trust can coexist, with particularised trust reinforcing generalised trust through cognitive biases such as the halo effect, where positive impressions extend across social judgments (Cao, Zhao, Ren, & Zhao, 2015; Freitag & Traunmüller, 2009; Huang et al., 2024). Others argue that particularised trust fosters in-group favouritism at the expense of out-group trust, reinforcing social divisions and exclusionary tendencies (Fisher, Van Heerde, & Tucker, 2010; Fu et al., 2012; Uslaner, 1999). Therefore, trust-related processes are not objective or universally applicable; they are shaped by individual trust predispositions, situational factors, and societal influences (Freitag & Bauer, 2016; Schilke et al., 2021; Tschannen-Moran & Hoy, 2000).

These opposing perspectives have direct implications for voice-based trustworthiness

judgments. If particularised trust fosters out-group distrust, listeners with high particularised trust may systematically rate out-group speakers as less trustworthy. Conversely, if generalised trust serves as a compensatory mechanism, it may counteract this bias and lead to more balanced evaluations (Delhey & Welzel, 2012; Freitag & Traunmüller, 2009; Huang et al., 2024). To shed more light on this debate, the present study will investigate how generalised and particularised trust interact in shaping vocal trustworthiness ratings. Specifically, the current analyses of Part 3 will examine whether individuals with higher generalised trust rate all speakers more favourably in terms of trustworthiness ratings, and if those with higher particularised trust rate in-group speakers as more trustworthy compared to out-group speakers.

4.4.1. Results of Part 3

To evaluate the pre-registered hypotheses (H3 and H4) the present study employed correlation analyses

Table 4.3: Mean scores of participants' trust propensity out of a total of 12 points

Ethnicity	**Age-group**	**Generalised trust mean score**	**Particularised trust mean score**
White	Younger	7.77	9.48
	Older	8.35	10.23
Black	Younger	7.06	8.96
	Older	7.17	9.29
South Asian	Younger	7.25	9.67
	Older	6.94	9.88

4.4.1.1. Assessing the relationship between generalised trust and trustworthiness perceptions

H3 hypothesised that generalised trust would be positively correlated with trustworthiness ratings. A Spearman's rank correlation was used between the mean trustworthiness ratings and generalised trust score of each participant. The results revealed a significant positive correlation, indicating that participants with higher generalised trust ratings tended to rate speakers as more trustworthy irrespective of intent, $r(286) = 0.28, p < 0.001$. See Table 4.3 for the mean scores of generalised and particularised trust per demographic group.

4.4.1.2. Correlation between particularised trust and trustworthiness ratings for in-group and out-group speakers

To examine hypothesis H4, which hypothesised that particularised trust would be negatively related to trustworthiness ratings for out-group speakers and positively related for in-group speakers, Spearman's rank correlation was used. For each listener, two separate analyses were conducted: one for in-group speakers and one for out-group speakers. The distinction between these groups was based on demographic similarities between the listener and the speaker. A listener was classified as being part of the in-group if their age-group and ethnicity matched the speaker's, whereas they were considered part of the out-group if either their age or ethnicity differed from the speaker's. Both in-group ($r(286) = 0.14, p = .02$) and out-group ($r(285) = 0.23, p < .001$) results were significant and both of them showed a positive correlation between particularised trust and trustworthiness ratings.

Considering the surprisingly positive relationship toward out-groups between particularised trust and trustworthiness ratings, a follow-up analysis was conducted to assess whether vocal intent could explain these results by interacting with particularised trust. Thus, a linear mixed-effects model (LMM) was employed on out-group data, with speaker intent and particularised trust scores acting as the predictors, listener trustworthiness ratings as the target and listeners as the random effects. Results revealed significant main effects for speaker intent ($\beta = .27, SE = .13, p = .045$) and particularised trust ($\beta = .37, SE = .09, p <$

.001), but no significant interaction between the two ($p = .19$). Thus, these findings show that out-group speakers who expressed speech with a trustworthy intent, received significantly higher trustworthiness ratings than those with neutral intent, supporting the idea that vocal adjustments in out-groups influence trustworthiness perceptions. However, listeners high in particularised trust gave higher ratings to out-group speakers regardless of whether the speaker intended to sound trustworthy or not. Therefore, vocal intent alone cannot explain the positive association between particularised trust and out-group evaluations.

4.4.2. Discussion of Part 3

The Part 3 analyses of this study, which focused on listeners' trust predispositions, wanted to examine how individual differences in generalised and particularised trust could have influenced voice-based trustworthiness evaluations. Findings revealed that higher generalised trust scores predicted higher trustworthiness ratings across all speakers, while higher particularised trust scores were unexpectedly associated with higher trustworthiness ratings for both in-group and out-group speakers. These results suggest that trust predispositions may not always operate in a rigid, categorical manner but instead interact more fluidly with speaker evaluations in voice-based conditions.

4.4.2.1. Generalised trust as a broad social trust heuristic

A key finding was that individuals high in generalised trust rated all speakers as more trustworthy, regardless of whether they were in-group or out-group members. This supports the idea that generalised trust functions as a broad, socio-cognitive heuristic that influences immediate social evaluations, even when minimal information is available (Bauer & Freitag, 2018; Freitag & Bauer, 2013; Hardin, 2002). The present findings also echo the view that generalised trust may stem from an individual's sense of long-term interpersonal security and stability, making them more likely to extend trust indiscriminately (Uslaner, 2002). It may be further argued that it also reflects an underlying belief in the goodwill of others, independent

of direct social experiences of others, promoting a more open rather than cautious approach in social judgements (Freitag & Traunmüller, 2009).

Unlike face-to-face interactions, where trust assessments integrate multiple cues (i.e., facial expressions, body gestures, and situational context), voice-based evaluations rely primarily on rapid auditory impressions (Maltezou-Papastylianou et al., 2025; McAleer et al., 2014). The persistence of generalised trust effects in an auditory-only setting suggests that generalised trust biases extend beyond visual social cues, reinforcing its role as an evaluative lens in conditions with less explicit information.

Interestingly, the current results also mirror the broader age effects in the present study. Older listeners —- who overall rated out-group (i.e., younger) speakers as more trustworthy —- may have been influenced by a trend of higher levels of generalised trust (see Table 4.3), predisposing them to evaluate all speakers as more trustworthy than their younger-listener counterparts, independent of demographic characteristics. Conversely, younger listeners, who were more selective in their trustworthiness ratings, their lower average scores on trust, may have prompted them to engage in more cautious trust evaluations.

Overall, the current results support the idea that generalised trust operates across sensory modalities, extending beyond face-to-face contexts into auditory settings (Maltezou-Papastylianou et al., 2025). Since voice-based evaluations rely primarily on rapid auditory impressions rather than more complex or explicit visual cues — like face expressions and body language —- this suggests that generalised trust biases persist even with reduced information.

4.4.2.2. Particularised trust: A shift toward flexible trust orientations?

In contrast to generalised trust, particularised trust has traditionally been conceptualised as reinforcing in-group favouritism —- leading individuals to withhold trust from out-group members (Fisher et al., 2010; Uslaner, 1999, 2002). However, the present study found that higher particularised trust was associated with more positive evaluations of both in-group and out-group speakers, challenging the assumption that particularised trust is strictly exclusionary.

A key distinction between generalised and particularised trust is that the former remained a broad, stable tendency, while the latter showed unexpected fluidity in out-group evaluations. One possible explanation is that particularised trust, while traditionally reinforcing in-group bias, may extend to out-groups in settings where individuating cues (such as vocal expressiveness) reduce reliance on rigid social categorisation (Burgoon & Hubbard, 2005; S. K. Kang & Bodenhausen, 2015; Keeley et al., 2013; P. W. Linville et al., 1986). Specifically, in voice-based evaluations, listeners lack visual identity markers, forcing them to rely more on vocal characteristics (Belin et al., 2019; Torre et al., 2020; Yokoyama & Daibo, 2012). Thus, voice-based interactions may have created a perceptual space where listeners with high particularised trust were more receptive to individuating vocal cues such as intent rather than broader and rigid social categories. In other words, particularised trust may not inherently promote bias but instead amplify evaluations in response to trust-relevant cues —- whether those cues come from in-group or out-group members. This interpretation complements past evidence which suggested that the two forms of trust predispositions can co-exist, with one reinforcing the other, likely due to cognitive biases such as the halo effect, where positive impressions can shape broader trust judgments (Cao et al., 2015; Freitag & Traunmüller, 2009; Huang et al., 2024). Such positive coexistence may be more prominent in multi-cultural societies such as the UK (Glanville & Shi, 2020).

It has been noted in the past that low collectivism societies tend to interact with strangers more, and social norms encourage engagement with people outside one's immediate circle (Allik & Realo, 2004; Delhey & Welzel, 2012; Glanville & Shi, 2020; Guo et al., 2022). Thus, fostering greater fluidity between particularised trust and generalised trust, making it more likely that someone who trusts their close acquaintances will also extend trust to the broader society (Allik & Realo, 2004; Delhey & Welzel, 2012; Glanville & Shi, 2020). Considering that present participants were recruited in multi-cultural England, this aspect of social capital could have influenced the present findings as an increase in social trust and decrease in collectivism has been observed in the UK (Allik & Realo, 2004; Duffy, 2023; Haerpfer et al., 2022a, 2022b). Importantly, the current results suggest that particularised trust may not always manifest as direct in-group favouritism but instead functions as a

flexible evaluative bias, shaped by available information and broader social capital attitudes, particularly in relation to social trust and collectivism (Allik & Realo, 2004; Guo et al., 2022).

Finally, as it was already implied, the present findings connect to broader study results on vocal intent. Since a trustworthy vocal intent overall increased trustworthiness ratings, with more nuanced findings when examining its effects more closely for specific speaker-listener out-group demographics, it is possible that individuals with high particularised trust were particularly responsive to these trust-enhancing cues. However, after a follow-up examination, it was revealed that speaker intent did not interact strongly with particularised trust, and consequently, could not explain the positive relationship toward out-groups between particularised trust and trustworthiness ratings.

Therefore, another possible explanation for the present findings is that particularised trust operates differently in higher vs lower-information and explicit vs perceptual conditions; prior studies have examined particularised trust in visual or face-to-face settings, where group identity is immediately visible and thus more likely to shape trust decisions with stronger social-demographic cues (Fisher et al., 2010; Freitag & Traunmüller, 2009). In contrast, voice-only settings obscure visual group markers, making trust assessments more reliant on non-explicit, perceived cues (Maltezou-Papastylianou et al., 2025; McAleer et al., 2014).

4.4.2.3. Reconciling trust orientations and their role in social judgments

These findings contribute to ongoing debates on how trust predispositions influence social judgments in lower-information and non-explicit settings. While generalised trust remained a broad evaluative bias, particularised trust showed flexibility, particularly in voice-based interactions. Hence, an important observation from the present study is that particularised trust may not always function as a rigid in-group bias, but instead operate flexibly on a spectrum depending on societal norms and communication modalities.

If particularised trust is increasingly extended beyond in-group members in multicultural contexts, does this signal a new, emerging model of selective but adaptive trust? Future

research should examine whether particularised trust is becoming more dynamic in diverse societies, particularly in domains where voice-based interactions are the primary mode of evaluation. Additionally, future studies should investigate whether particularised trust shifts occur only in non-explicit, perceptual, lower-information settings (such as voice-only interactions) or whether they extend to richer, multimodal communication formats where visual and situational cues are available. Comparing voice-based and face-to-face evaluations could provide deeper insights into whether these patterns reflect a structural change in how social trust functions or are limited to specific communicative conditions.

In an increasingly digital world where voice is often the primary medium for communication — such as in virtual assistants, hiring interviews, and legal testimonies — understanding how trust is formed and modified in auditory contexts is critical. Understanding how trust predispositions shape voice-based social judgments can therefore provide insights into improving communication strategies in multicultural and digital settings.

4.5. General discussion

This comprehensive investigation set out to unpack how trustworthiness judgments from voice are shaped by interconnected layers of speaker characteristics, listener predispositions, and social expectations. Across all current analyses, the results revealed that vocal trust is far from a neutral or automatic process. Instead, it emerges through a layered, often subtle interaction between who is speaking, how they sound, and what the listener brings to the table. Each study built upon the last to reveal different dimensions of this process —- from social categorisation biases, to speaker agency, to listener-driven heuristics. Although listener ethnicity was included in the analyses to examine in-group and out-group dynamics, it did not show significant main effects here; this factor will be revisited in Chapter 5.

This work showed that vocal trustworthiness was not simply a matter of in-group preference. Listeners consistently rated younger voices as more trustworthy than older ones,

regardless of their own age — suggesting that vocal qualities associated with youth (e.g., warmth, energy) were more persuasive than shared identity (McGettigan & Lavan, 2023; Mulac & Giles, 1996). Ethnic biases were more pronounced: white speakers received the highest ratings across the board, while black and south Asian speakers were rated lower, even by listeners from the same ethnic group (Sharma et al., 2022). Crucially, however, this wasn't a simple case of ethnic in-group favouritism. Black speakers were frequently misclassified —- often as white —- yet still received lower ratings. This mismatch showed more reliance on perceived ethnicity than actual, although it pointed more strongly to a deeper potential mechanism: listeners weren't judging based on categorisation alone, but on how a speaker's vocal cues aligned with their listeners' expectations on potential vocal group stereotyping (Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022). In other words, trustworthiness wasn't affected because someone was perceived from a certain ethnicity, but because they didn't "sound" enough like their perceived ethnicity (Coupland & Bishop, 2007; Geiger et al., 2023; Jiang et al., 2020). As such, misperceptions don't necessarily dissolve bias; they can simply redirect it toward new, subtler forms.

Consequently, in a follow-up analysis it was asked if out-group speakers can counter these biases through intentional vocal modulation. The answer was cautiously optimistic. Speakers who attempted to sound trustworthy were generally rated as more trustworthy than their neutral counterparts, albeit this effect wasn't distributed uniformly. Vocal intent had the strongest impact for younger black speakers rated by older listeners, and older black speakers rated by younger ones — precisely where ethnic categorisation was more ambiguous. These results support the idea that when listeners' expectations are less fixed, individuating cues like vocal expressiveness or accent carry more weight. Conversely, for speakers from more clearly recognised groups —- particularly older south Asian voices —— trustworthiness ratings barely shifted, suggesting that vocal intent alone may not be strong enough to override any embedded group stereotypes, which was supported by past research (Coupland & Bishop, 2007; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022). Notably, younger white speakers only saw a marginal boost, likely due to a ceiling effect, as their baseline ratings were already high (Keeley et al., 2013). Together, these patterns suggest that

vocal intent can work as a bias mitigation strategy but may depend on individual differences shaped by uncertainty, openness and socio-cultural factors. In this way, speaker agency has limits: impression formation can be nudged, but not fully erased, depending on how listeners are already primed to perceive the speaker.

Therefore, the spotlight was subsequently turned onto listeners themselves, to investigate how individual trust predispositions influence voice-based trustworthiness evaluations. As expected, individuals high in generalised trust rated all speakers more positively, reinforcing the idea that generalised trust functions as a broad social heuristic —- a kind of broad optimism about others' intentions (Bauer & Freitag, 2018; Freitag & Traunmüller, 2009; Hardin, 2002). Particularised trust yielded a more nuanced picture. Rather than rigidly reinforcing in-group favouritism (Fu et al., 2012), listeners with higher particularised trust rated both in-group and out-group speakers as more trustworthy regardless of vocal intent. Thus, suggesting that this form of trust may function more flexibly than traditionally theorised (Freitag & Bauer, 2013; Glanville & Shi, 2020). While vocal intent enhanced trustworthiness perceptions in general, it did not interact with particularised trust, indicating that these mechanisms may influence listeners' trustworthiness evaluations through separate pathways. These findings challenge binary views of particularised trust as solely exclusionary and points to the possibility that, in diverse and increasingly more individualistic societies with high social trust like the UK (Allik & Realo, 2004; Duffy, 2023; Haerpfer et al., 2022a), listeners with high particularised trust may extend trustworthiness beyond their immediate social circles, not because of speaker effort, but due to changing norms around social group predispositions and social exposure.

To conclude, this comprehensive study offers three key takeaways. First, trustworthiness evaluations from voice are not merely driven by static group affiliations but are shaped by interconnected relationship between speaker characteristics, listener predispositions, and situational cues. Second, while group biases may persist. especially along ethnic boundaries, these are not insurmountable: vocal expressiveness, particularly in younger out-group speakers, can serve as a meaningful cue that challenges listener expectations and improves evaluations. Third, trust predispositions —- especially particularised trust —- do

not always reinforce in-group favouritism. In diverse, high-trust socio-cultural environments like the UK, such predispositions may operate on a more flexible spectrum, enabling trust extension to out-group speakers when rigid identity markers are absent. These insights have direct relevance for designing inclusive communication strategies in professional, digital, and multicultural settings, where voice alone may influence real-world outcomes such as hiring, customer service, or public trust.

Transition to the next Chapter

Chapter 4 deepened the thesis' investigation on voice trustworthiness by showing through a multi-part study, how speaker-listener group membership and trust predispositions shape listeners' trustworthiness impressions. It highlighted that trustworthiness impressions are not formed in a vacuum: they are co-determined by speaker–listener demographic (age, ethnicity) dynamics, vocal intent, and listeners' broader, social world-views. These findings draw attention to the complexity of trustworthiness perception in human interactions — where vocal cues intersect with cognitive heuristics and social categorisation.

With the foundations laid across the human-focused speaker-listener chapters — tracing how trustworthiness is vocally expressed, socially interpreted, and shaped by individual differences — Chapter 5 delivers the thesis' final pivot: from human to machine. It offers a direct comparison of speaker nature (human vs synthesised voices), by assessing whether the trust-relevant factors identified in human speech — acoustic cues, vocal intent, ethnic variation, and listener predispositions — likewise influence impressions of real-world, commercially available synthesised voices.

In doing so, Chapter 5 addresses a pressing need in HAI research. As voice-based IAs become increasingly and seamlessly embedded in daily life —- from customer service to autonomous vehicles and healthcare — their design has largely centred on standardised, white, Western voice profiles, often overlooking the representation of ethnically diverse voices (e.g., Bilal & Barfield, 2021; Lima, Furtado, Furtado, & Almeida, 2019; Moussalli & Cardoso, 2017). This lack of diversity risks reinforcing social inequalities and limiting technology acceptance across user groups (e.g., Gluszek & Dovidio, 2010; Visser & El Fakiri, 2016). Given that trust is fundamental to both human interaction and technology adoption (Nam & Lyons, 2020; Simpson, 2007), understanding whether trust-relevant mechanisms identified in human speech

Transition to the next Chapter (continued)

also shape evaluations of voice-based IAs is not only theoretically valuable but critical for the development of inclusive and trustworthy voice technologies.

Therefore, Chapter 5 brings the thesis' multi-stage investigation full circle by offering a holistic view of whether similar social and perceptual mechanisms extend to voice-based IAs, and setting the stage for an integrative discussion of its broader theoretical and applied implications.

Chapter 5

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

5.1. 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 (Castelfranchi & Falcone, 2010; Hardin, 2002; Oleszkiewicz et al., 2017; 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), humanoid robots (Kouravanas & Pavlopoulos, 2022; Radford et al., 2015; Shigemi, Goswami, & Vadakkepat, 2018), and virtual agents (Yuan, Dennis, & Riemer, 2019). This tendency to anthropomorphise technology emphasises the potential of human-agent interactions (HAI) to simulate aspects of human-human communication (C. I. 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 (C. Nass et al., 1994) suggests that humans attribute social characteristics to machines based on cues like voice, forming impressions similar to those made in human interactions (Aylett, Vinciarelli, & Wester, 2017; Large et al., 2019; C. I. 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 across both human and synthesised (i.e., artificial) voices. The following sections review the relevant literature that informs this approach.

5.1.1. 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; Tanis & Postmes, 2005). For instance, Geiger et al. (2023) found that in a US 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

(Baquiran & Nicoladis, 2020), particularly in the U.S., where English dominates professional and academic settings (Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017). Alternatively, they may reflect a similarity-attraction bias, whereby participants favour speakers who seem linguistically or culturally similar to themselves (Dahlbäck et al., 2007; Montoya & Horton, 2013; C. Nass & Lee, 2000). 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, He, & Li, 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.

5.1.2. 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 (Baylor & Kim, 2003; Bilal & Barfield, 2021; Liao & He, 2020; Moreno & Flowerday, 2006; Tamagawa, Watson, Kuo, MacDonald, & Broadbent, 2011). The similarity-attraction effect remains relevant, especially when evaluating out-group or unfamiliar speakers (Aylett et al., 2017; Dahlbäck et al., 2007; C. Nass & Lee, 2000). 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; 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). Questionnaires like the Negative Attitudes to Robots Scale (NARS) reveal how individual differences shape perceptions of IAs (Kühne, Fischer, & Zhou, 2020; Lim et al., 2022; Nomura et al., 2006a; Nomura, Suzuki, Kanda, & Kato, 2006b). 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, Balkenius, & Johansson, 2022; Lim et al., 2022). 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, Lin, Panganiban, & Long, 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.

5.1.3. 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; S. Ko et al., 2020; Lim et al., 2022). 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; J. Kim et al., 2023; Rodero et al., 2014; S. M. 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). 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).

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). Analogously, a halo effect (i.e., a person's overall positive impression influencing judgments about specific traits) extends to perceived trustworthiness of machines (Gabrieli, Ng, & Esposito, 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) 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.

5.1.4. 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, Comber, & Gooberman-Hill, 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). Firstly, it was hypothesised that listeners with more negative attitudes toward robots (higher NARS scores) would rate synthesised voices lower than human voices, regardless of speaker intent or demographics (H1). It was 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 the present 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, it was 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, an exploratory analysis will also be conducted 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 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 the present study's focus. By narrowing the scope, this study aimed to ensure clearer and better motivated hypotheses. The role of speaker-listener sex can be explored separately in future work.

5.2. Methods

5.2.1. 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.

Table 5.1: Summary characteristics of speech acoustics examined

Acoustic features	**Typically measured in...**	**Key characteristics**
Fundamental frequency (f_0); perceived as pitch	Hertz (Hz)	f_0 is the lowest rate of vocal fold vibrations, with vocal intonation reflected in its variability within an utterance.

Continued on next page

Table 5.1: Summary characteristics of speech acoustics examined (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
Amplitude; perceived as loudness	Decibels (dB)	Reflects variations in air pressure.
HNR	dB	Lower HNR indicates increased noise in a voice signal (Fernandes et al., 2018; Ferrand, 2002). Noise refers to any element disrupting the clarity and quality of the intended speech, often unrelated to the voice's fundamental frequency; it may stem from vocal fold alterations, muscle tension, respiratory patterns, ambient sounds, or electronic interference (Ferrand, 2002).
Jitter	%	Reveals micro-fluctuations in pitch caused by irregular vocal fold vibrations, where a lower percentage indicates a smaller pitch variation in speech (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).
Shimmer	dB	Measures micro-fluctuations in amplitude, indicating variations in voice intensity (Baus et al., 2019; Felippe et al., 2006; Schweinberger et al., 2014).

Continued on next page

Table 5.1: Summary characteristics of speech acoustics examined (Continued)

Acoustic features	**Typically measured in...**	**Key characteristics**
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).
LTAS	dB	A lower LTAS often reflects longer vocal tracts (Da Silva et al., 2011; Hammarberg et al., 1980; S. E. Linville, 2002; Löfqvist, 1986), associated with deeper, more resonant voices linked to dominance, particularly in males (Gussenhoven, 2002; Puts et al., 2007).

5.2.2. 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."). 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. 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. For more information on the human stimuli and recording procedure see Maltezou-Papastylianou et al. (2024a, 2024b).

The IA voices were generated using Narakeet text-to-speech (TTS) web tool (`https://www.narakeet.com/`), no particular intent was specified. Narakeet was selected because at the time, it was the only publicly available tool I could identify that offered both South Asian and Black English-speaking TTS voices. This enabled the inclusion of a demographically diverse synthesised voice sample, in line with the aims of this research.

Although direct comparisons across TTS providers in the literature are limited, recent studies suggest that Narakeet voices are broadly comparable to major systems such as Google Cloud Text-to-Speech in terms of perceived quality and intelligibility (e.g., Kumar, Kumar, Sathe, & Pati, 2025; Norval, Wang, & Sun, 2023). For instance, mean opinion scores (MOS) provided by human evaluators and reported in benchmarking studies, place Narakeet TTS (MOS score: 3.65) within a similar quality and performance range as other commercial tools such as Google TTS (MOS score: 3.71) (Kumar et al., 2025).

The VoiceLab software (D. Feinberg, 2022; D. R. Feinberg & Cook, 2020) was used to extract several acoustic and spectral features to be examined in the present study: mean f_0, standard deviation of f_0 for perceived pitch variability, voice duration (measuring speech rate), and the voice quality features of HNR, jitter, shimmer, cepstral peak prominence (CPP), long-term average spectrum (LTAS), standard deviation of the LTAS, and LTAS slope —- indicative of the clarity and noise in a signal often reflected in perceived vocal breathiness, hoarseness or roughness. In particular, as seen in past research (Baus et al., 2019; McAleer et al., 2014), jitter was measured using the value of relative average perturbation (RAP), shimmer was measured as the value of amplitude perturbation quotient 3 (APQ3), and for f_0, VoiceLab's auto-correlation values were used. For further description of each acoustic feature see Table 5.1. Summary descriptives of each feature per demographic group can be found in Table 5.2 and Table 5.3 for human voices, and Table 5.4 for synthesised voices.

Table 5.2: 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	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 (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 (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]	8.90 [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]	27.00 [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]	16.30 [0.60]	21.70 [0.78]	16.55 [1.05]	17.00 [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 5.3: 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	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 (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 (Hz)**	35.63 [19.63]	58.48 [11.14]	26.62 [3.11]	18.00 [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]	16.60 [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 5.4: 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	1.41 [0.17]	1.46 [0.20]	1.49 [0.26]	1.58 [0.33]	1.50 [0.14]	1.62 [0.23]
f_0, mean (Hz)	119.61 [13.69]	180.75 [2.74]	115.25 [7.85]	189.00 [5.92]	160.28 [6.38]	238.95 [4.25]
f_0, SD (Hz)	37.41 [8.22]	44.10 [5.45]	20.04 [4.16]	24.51 [7.40]	39.30 [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)	23.90 [1.21]	24.45 [0.94]	25.72 [1.46]	25.40 [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]	3.60 [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]	-11.40 [1.09]	-14.13 [1.67]	-12.73 [1.65]	-14.82 [0.91]	-18.05 [1.66]

5.2.3. Participants

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 5.5 for more details

on listener demographics. Guidance from past research was followed, which 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.

Table 5.5: Descriptive statistics of participant demographics

Ethnicity	**Sex**	**N**	**Mean age (years)**	**Age range (years)**	**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

5.2.4. Rating procedure

During the study, 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 its subscales, suggests a less favourable evaluation of the interaction. Subsequently, 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).

5.3. Results

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 ω^2 = .01 (i.e., 1% of variance explained) is typically considered a small effect in the literature, an ω^2 = .06 a medium effect, and ω^2 = .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. All post-hoc comparisons were Holm-Bonferroni corrected (Abdi, 2010). For more information on the mean trustworthiness ratings per listener-speaker ethnicity, and speaker intent and nature, see Fig. 5.1.

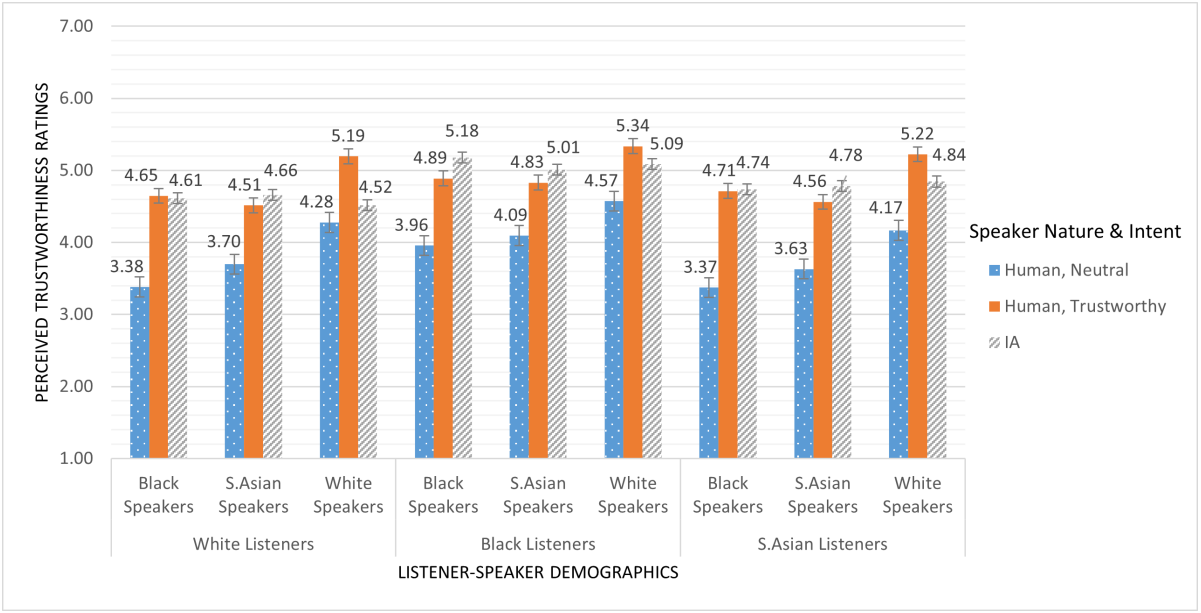

Figure 5.1: Listeners' mean trustworthiness ratings (1-strongly disagree to 7-strongly agree) per speaker nature, intent and demographic group.

5.3.1. Exploratory analysis

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, 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 5.6 for further details.

Table 5.6: Exploratory mixed-effects model results summary table

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
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 pitch**	0.02	0.001	20.88	**0.00**	0.02	0.02
f_0**, SD pitch**	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				

5.3.2. 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 (human vs 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, σ^2 = 0.24, reflects the amount of variability in trustworthiness ratings attributable to differences between participants' baseline trustworthiness ratings, independent of the fixed effects. Full results are presented in Table 5.7 and Fig. 5.1.

Table 5.7: Mixed-effects model results summary table

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
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

Continued on next page

Table 5.7: Mixed-effects model results summary table (Continued)

	Coef.(β)	S.E.	z	p-value	95% C.I.	
					Lower	Upper
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**	0.00	0.06
Grouping variable	0.24	0.02				

5.3.3. 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 designed to sound trustworthy (i.e. neutral), and that these ratings would be influenced by the speaker's ethnicity. 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, $F(1, 177) = 158.07$, $p < .001$, $\omega^2 = .27$, showing higher trustworthiness ratings for synthesised voices compared to human voices. The main effect of speaker ethnicity was significant, $F(1.94, 342.50) = 25.89$, $p < .001$, $\omega^2 = .05$, and so was listener ethnicity, $F(2, 177) = 11.003$, $p < .001$, $\omega^2 = .04$. 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 significance noted between white and south Asian listeners (*p = 0.54*).

Speaker nature x speaker ethnicity was the only significant interaction, *F(1.68, 296.59) = 31.85, p < .001,* ω^2 *= .05*. 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 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 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.

5.3.4. 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.

The main effect of speaker ethnicity was significant, $F(1.92, 338.95) = 19.56$, $p < .001$, $\omega^2 = .03$, and similarly listener ethnicity, $F(2, 177) = 6.29$, $p = .002$, $\omega^2 = .02$. 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 ($p = 0.13$). 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 and south Asian speakers ($M_{diff} = 0.24$, $SE = 0.11$, $p = .04$). No significant difference found between white and south Asian listeners ($p = 0.25$).

The only significant interaction was between speaker nature x speaker ethnicity, $F(1.73, 306.65) = 21.93$, $p < .001$, $\omega^2 = .04$. 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 black human speakers with south Asian human speakers ($p = 0.57$), nor with black and south Asian IA speakers ($p = 1.00$). No significance between south Asian human speakers and south Asian IA speakers either ($p = .088$).

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.

5.4. Discussion

The present research investigated how listener biases, 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.

5.4.1. 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, voice duration (here reflects 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.

Shorter sentence duration —- here indicative of faster speech rates — was associated with higher trustworthiness ratings. This aligns with research showing that faster speech can convey engagement, credibility and persuasiveness (Rodero et al., 2014; S. M. 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; J. Kim et al., 2023). 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; 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's 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 (S. E. 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; Oleszkiewicz et al., 2017). In short, not all acoustic cues are equally salient in all situational contexts — what enhances trust in one setting may be neutral or even detrimental in another.

These results offer valuable guidance for synthesised voice design. While not all vocal parameters need to be optimised simultaneously, the current 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 first impressions, adapting these to different usage scenarios (e.g., healthcare vs customer service).

5.4.2. 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; Nomura et al., 2006a). 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; C. Nass et al., 1994; C. I. Nass & Brave, 2005).

Unlike prior studies reporting broader effects of NARS scores (Krantz et al., 2022; Lim et al., 2022; Nomura et al., 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, the present 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). 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 the current study's initial prediction: listeners with higher scepticism toward emotional interactions rated synthesised voices as more trustworthy than human voices. A potential explanation may lie in the stimuli design. As seen in Chapter 2, all sentences were deliberately designed to avoid or minimise emotional bias and loaded language (e.g., "Hi, the shops are still open."), and human speakers were asked to speak the materials twice: once neutrally, and once with an intent to gain the listener's trust — no examples were provided on how they should sound to mitigate experimenter bias. In contrast the synthesised voices were created with Narakeet's default settings (standard volume, normal speed) and with no particular intent specified. Consequently, for listeners high on Subscale 3, this restrained delivery by synthesised voices may have reduced discomfort with emotional ambiguity, offering a sense of stability and impartiality in their interaction, in comparison to the emotional unpredictability and richness of human voices. This interpretation is consistent with the earlier acoustic findings of this study, which indicated that the current synthesised voices used occupy a perceptual middle ground between human neutral and human trustworthy speech. In comparison, human voices attempting to sound trustworthy may have inevitably introduced subtle emotional cues in their tone of voice, which high-Subscale 3 listeners could have interpreted as unwanted, ambiguous, or disingenuous. From this perspective, the current findings and interpretation seem to align with CASA and uncanny valley theories (Mori et al., 2012; C. Nass et al.,

1994; C. I. Nass & Brave, 2005), 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 and human-robot interactions.

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), 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.

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

Interestingly, the real-world synthesised voice stimuli used in the present study were rated as more trustworthy than the 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 Table 5.2 – Table 5.4). 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. 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, Goslin, White, & Zanatto, 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 (Kühne et al., 2020; Lee & Nass, 2010; C. Nass et al., 1994). 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.

5.4.4. 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 (C. I. 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, 2018; Yokoyama & Daibo, 2012). 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.

5.4.5. 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 this

study's 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 the current dataset (see Table 5.2 and Table 5.3). 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, US and Canada (Baquiran & Nicoladis, 2020; Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017). Similarly, the presence of lower pitch variability or more disfluent-sounding profiles 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; Moreno & Flowerday, 2006). 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; Montoya & Horton, 2013; Tamagawa et al., 2011). 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).

These dynamics are reinforced by past work showing that non-native or accented speakers are often rated less favourably on social impressions, even when content is controlled (Baquiran & Nicoladis, 2020; Cambre & Kulkarni, 2019; Dahlbäck et al., 2007; Geiger et al., 2023). 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.

5.4.6. 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 the current synthesised voice stimuli reflected real-world 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. Thus, it should be acknowledged that the representativeness of Narakeet TTS voices across the full spectrum of synthetic voice designs (e.g., expressive, neural, or domain-specific models) has not been formally established. The present findings should therefore be interpreted as an initial step, with future research needed to test whether trust-related impressions generalise across other TTS platforms and prosodic styles.

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. The controlled, perception-based design of this study 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.

5.4.7. 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 neutral and trustworthy human voices —- and were rated as more trustworthy than human voices with neutral intent (see Table 5.2 – Table 5.4 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. 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.

Chapter 6

General discussion and conclusion

This thesis set out to investigate how trustworthiness is expressed and perceived through voice, across both human and real-world IA speakers, and among demographically diverse speaker-listener groups. While trust is a central component of social interaction, the vocal cues that shape trustworthiness judgements among older (60+ years) or ethnically (black, south Asian) minoritised groups remain poorly understood, especially in voice-only settings (Maltezou-Papastylianou et al., 2025). This gap has limited the inclusivity/demographic representation, generalisability, and applied relevance of existing work. Addressing this, the present research undertook a systematic and multi-method exploration—beginning with a comprehensive review of the field and followed by four interlinked empirical studies, which have already been or will be published in peer-review journals. Across these, the thesis provides new theoretical, methodological, and practical insights into how speaker intent, acoustic features, social identity, listener predispositions, and human-AI dynamics jointly shape vocal trust.

6.1. A recap of this research journey

To establish a clear foundation, the thesis began with a systematic review (Chapter 1), which provided the first structured synthesis of voice-based trustworthiness research. It mapped out the fragmented state of the literature on voice trustworthiness by thematically organising

prior work according to speaker nature (human vs voice-based IAs) and communicative context (e.g., generic first impressions, public-facing roles, telehealth advice). The review identified not only commonly examined vocal cues under which trustworthiness impressions are evaluated, but also revealed recurring limitations: conceptual vagueness on definitions of trust and trustworthiness, methodological variations, and limited demographic diversity. Crucially, three areas were especially under-explored: how speakers themselves try to express trustworthiness through their voice; how listener evaluations are shaped by social group membership (e.g., ethnicity, age) and individual predispositions; and how human voices compare to real-world synthesised voices under the same evaluative lens. This review laid the conceptual and methodological foundation for the thesis, informed the design of subsequent studies, and produced a summary framework (Table 1.7) to clarify limitations and guide future research.

To address this gap, the thesis pursued a multi-stage investigation, comprising four interlinked studies — progressing from speaker-centred production to listener-centred perception, and finally to cross-domain comparison between human and real-world synthesised voices. To summarise, the dataset-descriptive study (Chapter 2) set out with how speakers encode their trustworthy intentions, and moving to how listeners interpret it across acoustic cues and impressions of trustworthiness, warmth and competence (Chapter 3). Then, probing how group membership and trust predispositions shape these impressions (Chapter 4), and finally comparing these patterns across human and commercially available, real-world synthesised voices (Chapter 5). Each study was designed to build on the previous — methodologically, conceptually, and in scope — enabling a multi-layered understanding of trust-related cues across both human and voice-based IA interactions.

To support this thesis and the research community, a standardised and open-access speech audio dataset varying in terms of speaker age (younger and older than 60 years), ethnicity (white, black, south Asian), sex and intent (trustworthy intent vs natural tone of voice — termed as "neutral" intent), was developed and validated (Chapter 2; Maltezou-Papastylianou et al., 2024a, 2024b). Acoustic features such as perceived pitch, HNR, shimmer and LTAS varied significantly and systematically with vocal intent across all demographic groups,

framing speaker vocal intent as a meaningful behavioural signal and not just a theoretical construct. The resulting dataset represents a unique resource for the field and allows for more generalisable, inclusive conclusions about how trustworthiness is encoded and decoded in speech. Hence, it served as the basis for the perception experiments that followed.

A key methodological strength of this thesis is the recruitment of a speaker-listener sample that reflects both ethnic and age diversity, who remain under-represented in voice-based perception research (Maltezou-Papastylianou et al., 2025; Taylor & Rommelfanger, 2022). While recruitment presented logistical challenges — such as limited access to older ethnic minority participants, lower digital literacy in some groups, and mistrust toward research processes — it was essential to ensure that the findings reflect a broader population beyond younger, white, or digitally confident individuals (Taylor & Rommelfanger, 2022). These challenges and barriers, well-documented in prior research (Ellard-Gray, Jeffrey, Choubak, & Crann, 2015; McDougall Jr, Simpson, & Friend, 2015; Sun et al., 2024), were evident in the present study and required extended timelines and, at times, support from specialist recruitment services (Chandler et al., 2019; CloudResearch, 2015). Though largely behind the scenes, these efforts were central to building a more inclusive and standardised speech audio dataset (Maltezou-Papastylianou et al., 2024b). To the best of my knowledge, this is the largest open-access dataset on vocal trustworthiness to date (1,152 audio files) that includes this level of demographic representation — positioning the thesis as a contribution both in its empirical findings and in modelling a more inclusive approach to voice-based research.

6.2. Key empirical contributions

Rather than treating each study as a stand-alone contribution, this section draws out the overarching findings across five thematic domains: (1) conceptual framing of trustworthiness, (2) listener predispositions, (3) speaker-level vocal attributes across domains, (4) the role of vocal intent, and (5) socio-cognitive mechanisms of group bias.

6.2.1. Reframing vocal trustworthiness: Intersections with warmth and competence

This thesis extends prior work on vocal trustworthiness (Guldner et al., 2024; McAleer et al., 2014) by clarifying how it relates to impressions of warmth and competence —– two core dimensions in social perception —– within a demographically diverse, voice-based context. Drawing on the Stereotype Content Model (Cuddy et al., 2009; Fiske, 2018) and the integrative model of trust (Mayer et al., 1995), trustworthiness is widely theorised as a function of both perceived intent (warmth) and ability (competence). However, existing voice perception models have typically focused on acoustic signal decoding or speaker identity (e.g., Belin et al., 2004; McAleer et al., 2014), rarely integrating these broader social cognition frameworks. This thesis bridges that gap, demonstrating how vocal trustworthiness impressions are shaped by the perceptual relationship of warmth, competence, and modulated acoustic cues.

Findings from Chapter 3, where participants rated the same speech samples across all three impressions, revealed strong, positive associations between trustworthiness, warmth, and competence. However, trustworthiness aligned more closely with warmth than with competence. This directional asymmetry suggests that, in brief and low-stakes socially framed vocal interactions, listeners prioritise perceived benevolence over perceived ability when forming trustworthiness judgments. Instead, competence may be more influential in intellectual or goal-oriented situations, where credibility is sought. This reinforces the idea that trustworthiness emerges as a composite percept, shaped by overlapping, but not equal, contributions from warmth and competence. Importantly, this has implications for applied settings, suggesting that cues of friendliness, engagement and sincerity may be more impactful in social, early-stage or auditory encounters, while cues of competence and credibility to shine in higher-stakes and professional contexts.

Acoustic analyses supported these perceptual patterns, revealing consistent predictors across all three traits. Trustworthy voices in this thesis, were typically characterised

by faster speech rates, higher mean pitch, and greater pitch variability — cues often associated with energy, engagement, and social presence (Maltezou-Papastylianou et al., 2025). Interestingly, lower values of voice quality features such as HNR and shimmer — which denote greater vocal irregularity or variability —- and higher mean LTAS — which denotes greater resonance —- were also associated with higher ratings of trustworthiness, warmth, and competence (Maltezou-Papastylianou et al., 2023). This pattern reflects a perceptual preference for voices that sound richer, more expressive, and authentic in brief, social interactions, rather than overly smooth, mechanised or monotone. While the role of specific acoustic cues is discussed in more detail in subsequent sections, this convergence points to a shared perceptual heuristic through which listeners assess vocal traits, particularly in voice-only settings. In sum, the thesis reconceptualises trustworthiness not as a discrete perceptual dimension but as a socially grounded, acoustically mediated construct that reflects both theoretical models and real-world communicative demands.

These results highlight the need for future models of vocal impression formation to treat trustworthiness not as a discrete trait, but as one modulated by warmth, competence, and contextual framing, and inferred through a constellation of acoustic cues. Taken together, the findings offer a more integrated view of how vocal trust is perceived: one that respects its conceptual distinctiveness but recognises its perceptual dependency on adjacent, core social dimensions. This has implications for both theoretical modelling and applied contexts; particularly, in environments where trust needs to be established rapidly and without visual cues, such as interviews, automated systems, or remote service interactions.

6.2.2. Trust predispositions as perceptual filters

This thesis provides one of the first integrated examinations of how social trust predispositions —- both human-focused and technology-directed —- influence voice-only trustworthiness impressions. While prior research has extensively examined how perceived pitch and other acoustic features shape vocal trustworthiness judgements (e.g., McAleer et al., 2014; Torre et al., 2020), models of voice perception have predominantly focused on the

processing of speaker identity, affect, or prosody (e.g., Belin et al., 2011, 2004; Lavan et al., 2019; Schweinberger et al., 2014), often overlooking the role of more stable listener-level characteristics such as trust predispositions, in modulating their judgements.

The present findings challenge this speaker-centric focus, be it acoustic-based or perceptual. They demonstrate that trustworthiness impressions do not emerge from vocal cues alone, but from the interaction between speaker behaviour and the listener's socio-cognitive dispositions and worldview —- both in human-human interaction and in HAI. This was particularly evident in human-human voice interaction (Chapter 4), where listeners with higher generalised trust rated all speakers more favourably, irrespective of shared social group membership. Extending previous findings on interpersonal openness (Hardin, 2002; Uslaner, 2002), these results show that even in voice-only, thin-slice impressions, listeners apply this heuristic of openness. Crucially, particularised trust, which is traditionally associated with in-group favouritism (Hornsey, 2008; Z. Peng et al., 2019), did not reinforce rigid group-based bias in the present study. Instead, listeners higher in particularised trust rated both in-group and out-group speakers as more trustworthy. These findings show a shift from traditional binary views of particularised trust as solely exclusionary (Fisher et al., 2010; Fu et al., 2012; Uslaner, 1999). They point to the possibility that, high social trust in diverse and increasingly more individualistic societies like the UK (Allik & Realo, 2004; Duffy, 2023; Haerpfer et al., 2022a, 2022b), listeners with high particularised trust may extend positive trustworthiness evaluations beyond their immediate social circles, not necessarily because of speaker effort, but due to changing social norms and exposure to diversity.

Chapter 5 extended these findings to voice-based IAs, showing that listener predispositions continued to exert a measurable influence on trust-related evaluations of synthesised voices. Notably, the NARS scale revealed two contrasting trends: Listeners with higher discomfort toward robots rated expressive synthesised voices as less trustworthy, whereas those sceptical of emotionally expressive robots rated them more positively. Thus, emotional expressiveness, often promoted as a universal trust-enhancing feature in the design of voice-based IAs (e.g., Large et al., 2019; C. I. Nass & Brave, 2005), was not equally effective

across listeners. The present findings highlight the influence of listener predispositions on either enhancing or suppressing trust, depending on the alignment between perceived vocal cues and pre-existing listener attitudes. These findings have important implications for the growing body of research advocating for a univeral increase in human-likeness and emotional expressiveness in synthesised voices to enhance user trust (e.g., Mara, Appel, & Gnambs, 2022; Torre et al., 2020; Yuan et al., 2019). The present findings suggest that emotional expressiveness is not universally effective, and in fact, may backfire among listeners who may associate different voice styles with disingenuousness, uncertainty, or uncanny valley in IAs.

In sum, this thesis positions listener predispositions —– both in human-human and HAI contexts —– as essential components of vocal trust evaluation. It calls for future models of voice perception and trustworthiness to move beyond more static, speaker-centric frameworks and adopt more dynamic, interactional perspectives that integrate listener variability. From an applied standpoint, the findings caution against over-reliance on a single "optimal" vocal profile for trust enhancement. Instead, voice-based systems, particularly in high-stakes or sensitive domains such as healthcare or financial support, may benefit from adaptive design strategies, including offering users a choice between more personalised vocal expressions, and transparently communicating system capabilities to align with listener expectations.

6.2.3. Acoustic signatures of trust: Convergence across production, perception, and synthesis

Across three empirical chapters (Chapters 2, 3, and 5), this thesis identified a stable cluster of acoustic features — namely, speech rate, mean pitch, HNR, shimmer and LTAS — that shaped perceptions of trustworthiness. These findings emerged from different perspectives: speaker self-modulation (Chapter 2), listener impression formation (Chapter 3), and voice identity comparisons (Chapter 5).

This convergence suggests that trustworthiness perceptions are consistently enhanced when voices exhibit a combination of faster pacing, higher mean pitch, greater pitch

variability, and voice quality characteristics that introduce subtle variability and resonance —– such as lower shimmer and HNR values, and higher mean LTAS. Together, these acoustic cues shape a voice that sounds livelier, more expressive, engaged, and socially present, in contrast to voices that may come across as flat, monotone, overly polished or emotionally disengaged. Faster speech rates may signal energy, fluency, and competence; higher pitch and greater variation may convey friendliness, enthusiasm, or attentiveness; while lower shimmer and HNR, alongside higher LTAS energy, contribute to a richer, more textured vocal quality, often experienced as authentic, natural, human-like expressiveness and variation, in contrast to overly smooth or "synthetic-sounding" speech (McAleer et al., 2014; Torre et al., 2020, 2016; Yokoyama & Daibo, 2012). Importantly, these patterns held across demographically diverse human voices and real-world, commercially available synthesised voices. This wide-reaching consistency indicates that listeners rely on a core set of acoustic heuristics as perceptual shortcuts to infer social impressions about trustworthiness, rather than solely speaker nature or identity.

These acoustic markers align with listeners' expectations for trustworthiness, particularly in brief or low-stakes social encounters, such as first impressions or casual conversation, where visual cues are absent and vocal signals become the primary basis for inferring speaker intent or social openness (Maltezou-Papastylianou et al., 2024a; Uslaner, 2002). In such contexts, voices that sound engaged, spontaneous, and emotionally connected are more likely to foster impressions of warmth, sincerity, and trust (Maltezou-Papastylianou et al., 2025; O'Connor & Barclay, 2018). However, these findings further emphasise that trust-enhancing acoustic profiles are unlikely to be universally optimal. The social meaning of vocal cues is context-dependent, shaped by listener expectations and situational demands (Maltezou-Papastylianou et al., 2025). While the vocal profile identified here may promote approachability and social connection, other settings may favour attributes linked to authority, expertise, or control, such as a combination of lower pitch, reduced variability, and faster pacing (J. Kim et al., 2023; Rodero et al., 2014; Schirmer et al., 2020). Ultimately, the optimal acoustic profile for trustworthiness likely depends on whether the interaction prioritises emotional engagement, informational credibility, or social dominance, and may

interact with listener preferences shaped by speaker characteristics such as gender, age, or ethnicity.

Moreover, the real-world synthesised voices in this thesis consistently fell between human neutral and trust-modulated voices in their acoustic profiles; a pattern that I characterise as occupying a *"perceptual middle ground"*, a term introduced here to describe this balance point in vocal trustworthiness design. While they were not rated as highly as human voices conveying a trustworthy intent, their acoustic design appeared to strike a balance —— possibly aiding their trustworthiness evaluations. Listeners appear to anchor their judgments in consistent auditory signals that convey vocal effort, clarity, and prosodic balance — qualities likely evolved for interpreting speaker intent in natural, human-human interactions (Gussenhoven, 2002; C. I. Nass & Brave, 2005; Ohala, 1995). This has implications for voice technology: thoughtfully engineered acoustic profiles may enhance listener trust without requiring full human-likeness or expressiveness, avoiding issues like uncanny valley effects or overinflated user expectations (Mori et al., 2012; C. Nass et al., 1994). Rather than aiming for a one-size-fits-all "trustworthy voice", theoretical models (e.g., Castelfranchi & Falcone, 2010; C. Nass et al., 1994; Soroka et al., 2003) and applied voice design should recognise trustworthiness as a context-sensitive, interactional construct. For voice-based IAs, flexible acoustic strategies —— such as offering user-selectable voice styles or dynamically adjusting vocal delivery to suit interactional goals —— may provide practical avenues to fostering trust while accommodating diverse user preferences and communicative settings.

In sum, this thesis provides rare cross-validation of trust-relevant acoustic features, spanning production, perception, and synthetic replication. This convergence builds a powerful case for the existence of generalisable perceptual mechanisms influencing trust-related judgements in voice, and positions vocal acoustics as a central channel for encoding and decoding social intent across increasingly human–robot hybrid environments.

6.2.4. Intentional voice modulation: Potential and limits for social signalling

A recurring question across the thesis was whether trustworthiness could be intentionally enhanced through vocal performance. Findings from Chapters 2, 4, and 5 collectively suggest: yes, but not universally.

Speakers asked to sound trustworthy (Chapter 2) intuitively adjusted their speech in a similar acoustic manner, suggesting the existence of a socially shared or culturally internalised vocal template for signalling trustworthiness (Maltezou-Papastylianou et al., 2024b). Listener ratings in Chapters 4 and 5 confirmed that such vocal intent often translated into higher trustworthiness impressions, especially for speakers whose baseline trustworthiness impression was lower, or whose social identity was more ambiguous. This was most noticeable for younger black speakers evaluated by older listeners, whose group categorisation was less consistent and vocal cues may have carried greater perceptual weight. In contrast, for clearly identified speakers — especially older south Asian voices — the impact of vocal modulation was smaller, possibly reflecting stronger stereotype anchoring or impressions on linguistic characteristics such as accent, speech pronunciation and articulation (Geiger et al., 2023; Hanzlíková & Skarnitzl, 2017; Sharma et al., 2022).

Chapter 5 extended this investigation to a cross-domain comparison: human voices modulated to convey a trustworthy intent were rated highest; however, real-world synthesised voices were rated as more trustworthy than human voices with no such intent (i.e., neutral). This pattern highlights the dual pathway to vocal trustworthiness: speaker intent and acoustic engineering. The practical implication is clear: while vocal modulation remains a powerful tool, its effectiveness depends on speaker identity, listener profile, and sociocultural expectations. One-size-fits-all approaches are unlikely to succeed, and calls into question simplistic notions that "sounding more trustworthy" is a universally available or effective remedy for social bias.

From a theoretical perspective, this thesis contributes to debates on impression malleability and bias mitigation in communication. The current findings clarify a key boundary condition in voice perception research: while an explicit vocal intent can enhance trustworthiness impressions, it does so within clear perceptual limits shaped by

speakers' vocal cues, listeners' social categorisation, and the situational context (Maltezou-Papastylianou et al., 2025).

More broadly, these findings reinforce that there is no universal vocal cue profile for eliciting trust or overcoming bias, whether in human communication or the design of synthesised voices. They also echo a central argument of this thesis: trustworthiness impressions are shaped through a combination of processes rather than fixed responses to vocal cues alone. While vocal modulation can offer perceptual benefits, its success is shaped as much by social dynamics and situational context as by signal properties. Accordingly, attempts to foster trustworthiness —– whether through speaker training or technological design —– must be grounded in an understanding of the social, cultural, and perceptual constraints within which voice operates.

6.2.5. Social categorisation and perceptual asymmetry

Building on the previous sections, Chapter 4 further problematised assumptions around in-group bias. Contrary to standard social identity theory (e.g., Bailey et al., 2016; Hornsey, 2008; Montoya & Horton, 2013; Z. Peng et al., 2019), speaker-listener shared ethnicity or age did not consistently increase trustworthiness ratings. Instead, younger and white-sounding voices were favoured across listener groups. Misclassification rates — particularly for older and black speakers — suggest that perceptual ambiguity, rather than objective similarity, governed evaluation.

This finding reframes in-group bias in voice-based contexts as perceptual rather than based on actuality. Biases did not emerge from who the speaker actually was, but from who the listener believed them to be, and future work should explore this further. In voice-only interactions, such as phone interviews, telehealth, or virtual assistants, this has pressing implications: speakers' demographic representation may not shield against bias if listeners' perceptual heuristics dominate the interpretive process. These insights call for more nuanced voice system design, one that accounts for how identity is perceived, not just encoded. More broadly, they suggest that models of voice perception (e.g., Belin

et al., 2011; Schweinberger et al., 2014) must move beyond static identity recognition to incorporate perceptual categorisation uncertainty and misclassification processes as central to understanding vocal bias.

6.2.6. Summary

Collectively, these thematic contributions form a cohesive narrative of how trustworthiness is communicated through voice. The thesis began by clarifying the conceptual architecture of vocal trustworthiness, before moving through the influence of listener predispositions, the power and limits of acoustic features, the role of intentional modulation, and the socio-cognitive mechanisms steering bias. Across these stages, the findings emphasise that voice-based trustworthiness is not static, nor reducible to fixed speaker traits or acoustic cues — challenging current voice perception models (e.g., Belin et al., 2011, 2004; Lavan et al., 2019; Schweinberger et al., 2014). Rather, it emerges through the combined process between context-sensitive settings, acoustic signals, social categories, and listeners' pre-existing attitudes, across both human and synthesised voices. This richer account not only advances theoretical understanding on vocal trustworthiness but also offers practical guidance for designing voice-based systems that foster trust across diverse demographics and real-world settings. The next section builds on these insights to outline concrete design implications and grounds future research in a more inclusive and all-encompassing understanding of how trustworthiness is communicated through voice.

6.3. Design implications for voice-based interaction: Insights from human and synthesised speech

Across the empirical studies in this thesis, a recurring theme has been the complexity of designing trustworthy voices – whether for human speakers (e.g., public figures, healthcare professionals) or voice-based IAs (e.g., voice assistants, humanoid robots). While no singular

acoustic profile guarantees positive impressions of trustworthiness, the evidence assembled here offers actionable insights into how and when certain vocal features, speaker traits, and user predispositions interact to shape trustworthiness impressions. Table 6.1 below distils these findings into design guidelines, drawing directly from observed results and contextualising them for applied use – from conversational IA development to social training and public speaking. These are not prescriptive rules, but empirically informed principles that highlight what works, for whom, and under what conditions.

Table 6.1: Evidence-based design recommendations derived from the thesis for enhancing vocal trustworthiness in human and synthesised voice applications

Guidelines	**Design insights**	**Evidence from thesis**
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.	Chapter 3: Trait asymmetry findings; theoretical integration with SCM and Mayer et al.'s trust model.
G2. Expressiveness can enhance trust — but must fit context and user	Intentional modulation (e.g., speaking with warmth or sincerity) increased trust ratings, especially for human voices. However, expressiveness may backfire for sceptical users (e.g., high-NARS individuals) or where emotional neutrality is expected (e.g., security alerts).	Chapters 4 and 5: Effects of vocal intent; NARS-related trust penalties.

Continued on next page

Table 6.1: Evidence-based design recommendations derived from the thesis for enhancing vocal trustworthiness in human and synthesised voice applications (Continued)

Guidelines	Design insights	Evidence from thesis
G3. Calibrate human-likeness in synthetic voice design	The synthesised voices in this thesis 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.	Chapter 5: Comparative acoustic and perceptual analysis of human vs synthesised voices.
G4. Anticipate and accommodate user predispositions	Listener biases influenced trust outcomes. 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.	Chapters 4 and 5: Trust predispositions (NARS, generalised/-particularised trust).

Continued on next page

Table 6.1: Evidence-based design recommendations derived from the thesis for enhancing vocal trustworthiness in human and synthesised voice applications (Continued)

Guidelines	Design insights	Evidence from thesis
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.	Chapters 3-5: Cross-group trustworthiness bias effects; acoustic controls.
G6. Time and pitch are powerful cues —- but require restraint	Faster speech 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.	Chapters 3-5: Acoustic predictors of trust-related traits.
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.	Chapters 3-5: Interaction between vocal intent and demographics.

Continued on next page

Table 6.1: Evidence-based design recommendations derived from the thesis for enhancing vocal trustworthiness in human and synthesised voice applications (Continued)

Guidelines	Design insights	Evidence from thesis
G8. Cultural norms shape how vocal cues are interpreted	Features such as pitch variability were expressed and received differently across ethnic groups —- 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.	Chapters 2-5: Ethnicity-based variation in acoustic cue interpretation.
G9. When in doubt, design for a *"perceptual middle ground"* in acoustic expressiveness	In this thesis, trustworthy perceptions of synthesised voices tended to occupy acoustic values between neutral and trustworthy-intended human speech. This "perceptual middle ground" as I call it, 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.	Chapter 5: Acoustic patterns in synthesised voice ratings.

Continued on next page

Table 6.1: Evidence-based design recommendations derived from the thesis for enhancing vocal trustworthiness in human and synthesised voice applications (Continued)

Guidelines	Design insights	Evidence from thesis
G10. First impressions are rapid —- design accordingly	Strong trait impressions were formed from brief utterances, often in under two seconds. For both humans and voice-based IAs, early speech cues (e.g., pitch, pacing) significantly shaped perceived trustworthiness, warmth and competence. This is especially relevant in onboarding scenarios, help requests, or cold calls.	Chapters 3 and 5: Trait evaluations from brief, audio clips.

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. These 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.

6.4. Limitations and future directions

While this thesis offers a comprehensive examination of how vocal trustworthiness is communicated, interpreted, and shaped across both human and synthesised voices, it is important to acknowledge several methodological and conceptual limitations that shape the scope and applicability of the findings.

First, although the speaker and listener sample was one of the most demographically

inclusive in existing voice-based trust research – with deliberate efforts made to include older and ethnically diverse participants across both roles – recruitment constraints and the online nature of the studies meant that participation was limited to English-speaking adults based in England. As such, cultural and linguistic generalisability remains limited. Future work would benefit from cross-cultural replications in non-Western or multilingual contexts, where vocal norms and trust-related expectations may differ significantly. Likewise, while listener sex and speaker sex were recorded, they were not actively analysed in the empirical studies for reasons of analytical scope. Yet, as summarised in this thesis' systematic review (Chapter 1; Maltezou-Papastylianou et al., 2025), gendered vocal stereotypes are known to influence trait evaluations. Female voices are often associated with warmth and trust, and lower-pitched male voices with dominance or competence, albeit context-dependent (O'Connor & Barclay, 2018; Ohala, 1995; Schirmer et al., 2020). Yet, these effects remain under-explored in voice-based trustworthiness research involving demographically diverse speakers, and particularly in interaction with other social cues such as ethnicity, age, or synthesised voice design. Future research should examine whether gendered vocal expectations modulate trust-related impressions differently across speaker-listener demographics, and human vs synthesised voices.

Nevertheless, it is worth clarifying that the imbalance in older ethnic minority speakers applied to the full dataset described in Chapter 2, which was published as a descriptor paper of the dataset. For the experimental studies reported in Chapter 3 – Chapter 5, I drew on a randomised, balanced subset of speakers across ethnicities and age-groups to minimise such imbalances. As such, the under-representation of older minority speakers did not directly bias the experimental findings, though broader limitations on cultural and linguistic generalisability remain.

A second limitation lies in the controlled, audio-only design of the empirical studies. While this allowed for the isolation of vocal cues without confounding visual or contextual stimuli, it does not reflect all real-world listening scenarios, especially those where listeners have access to multimodal cues (e.g., facial expressions, gestures, or situational context). Prior multimodal studies have shown that facial cues often interact with vocal signals in

shaping trait impressions, including trustworthiness (e.g., Elkins et al., 2012; Maxim et al., 2023; Mileva et al., 2018). For example, speakers perceived as trustworthy in voice-only scenarios, have been judged less favourably when paired with incongruent facial expressions, highlighting the powerful influence of visual information (Elkins & Derrick, 2013; Mileva et al., 2018). This is consistent with findings from this thesis' systematic review (Chapter 1; Maltezou-Papastylianou et al., 2025), which reported that pitch effects on perceived trustworthiness were weaker or less consistent in multimodal compared to unimodal (i.e., voice-only) settings. Accordingly, trust-enhancing vocal cues identified in this thesis may be amplified, softened, or reinterpreted when paired with visual characteristics such as facial expressions, demographics, body language, or virtual avatars. For this reason, future work should explore whether the perceptual patterns reported here hold in more ecologically valid or immersive settings, including live conversations, video-based assessments, or embodied voice-based agents.

Third, the synthesis of empirical findings highlighted that certain trust-enhancing cues – such as higher pitch, faster speech rate, or increased expressiveness – tend to benefit perceptions of warmth and trustworthiness, particularly in socially framed tasks. However, these effects were not uniformly effective across speakers, listeners, or domains. For instance, speaker intent was more effective for younger speakers than for older ones, and listener predispositions (e.g., high negative attitudes toward robots) sometimes counteracted otherwise positive cues. These nuanced effects highlight the need for future studies to adopt more context-sensitive designs —– varying communicative goals, speaker roles, and listener expectations —– to better understand when and why certain cues succeed or fail. Investigating multimodal voice agents that can also express contextually aligned emotional cues, or allow user-driven adaptation (e.g., adjusting speaking style or vocal tone dynamically), could offer valuable insight into how perceived trust can be flexibly shaped.

Finally, although the thesis measured trustworthiness, warmth, and competence using well-validated first-impression metrics, these ratings were based on single-exposure judgements, often formed within seconds. While this aligns with real-world "thin-slice" impression formation (e.g., Gheorghiu et al., 2020; Lavan, 2023), it limits our understanding

of how vocal trustworthiness evolves over time or through repeated interactions. Longitudinal designs, or paradigms that examine how trust develops (or deteriorates) over sustained voice-based encounters, would deepen the field's understanding of voice-mediated trust as a temporally sensitive process. This is especially relevant in applied domains where trust unfolds gradually, such as telehealth, education, or long-term HAI.

In sum, while the thesis presents a well-controlled and rich investigation into vocal trustworthiness, future research should aim to expand the ecological, demographic, and contextual reach of the findings. By doing so, it can further test the robustness of the acoustic, psychological, and social mechanisms identified here, and move toward more inclusive, multimodal, and user-sensitive models of vocal trustworthiness in both human and synthesised voice communication.

6.5. Concluding reflections

This thesis set out to investigate how trustworthiness is communicated and perceived through voice – across both human and synthesised speakers – and how this process is shaped by acoustic features, speaker intent, listener predispositions, and sociocultural context. By combining a systematic synthesis of the existing literature with a series of original empirical studies, it contributes a clearer and more integrated understanding of vocal trustworthiness as a multi-dimensional construct. Across the work, vocal cues such as pitch, speech rate, and signal clarity consistently emerged as reliable perceptual heuristics; however, their influence was not static. Trustworthiness was shown to be context-sensitive, modulated by speaker–listener identity, expressive intent, and listeners' own biases. Importantly, this thesis advances the field by placing these effects within diverse, real-world voice samples – including those often under-represented in voice research – and by directly comparing the multi-faceted interaction between human and synthesised voices.

Taken together, these findings offer both theoretical insight and practical guidance for future research and voice interface design. They highlight that trust, when mediated by voice,

is not simply a function of signal quality or expressive cues in isolation, but the product of an interaction between acoustic structure, social context, and listeners' own socio-cognitive filters. By clarifying when and how vocal trustworthiness cues succeed, and when they may vary, this thesis provides a more grounded foundation for understanding voice-based communication in a world increasingly shaped by seamlessly embedded intelligent voice technologies and cross-cultural interaction.

Beyond its academic contribution, this thesis represents a personal journey of perseverance, learning, and deep curiosity about the intersection of human perception and voice-based technologies. Conducting this work —- particularly with hard-to-reach, older and ethnically diverse participant groups —- required not only methodological rigour but also a commitment to inclusive research. It is my hope that the findings will not only inform future work in voice perception and AI design, but also contribute meaningfully to more human-centred, equitable approaches in both science and technology.

References

Abdi, H. (2010). Holm's sequential bonferroni procedure. *Encyclopedia of research design*, *1*(8), 1–8.

Allik, J., & Realo, A. (2004). Individualism-collectivism and social capital. *Journal of cross-cultural psychology*, *35*(1), 29–49.

Asch, S. E. (1946). Forming impressions of personality. *The journal of abnormal and social psychology*, *41*(3), 258.

Aylett, M. P., Vinciarelli, A., & Wester, M. (2017). Speech synthesis for the generation of artificial personality. *IEEE transactions on affective computing*, *11*(2), 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*(4), 219–224.

Badillo, S., Banfai, B., Birzele, F., Davydov, I. I., Hutchinson, L., Kam-Thong, T., … Zhang, J. D. (2020). An introduction to machine learning. *Clinical pharmacology & therapeutics*, *107*(4), 871–885.

Baier, A. (2014). Trust and antitrust. In (pp. 604–629). Routledge.

Bailey, P. E., Szczap, P., McLennan, S. N., Slessor, G., Ruffman, T., & Rendell, P. G. (2016). Age-related similarities and differences in first impressions of trustworthiness. *Cognition and Emotion*, *30*(5), 1017–1026.

Bangen, K. J., Meeks, T. W., & Jeste, D. V. (2013). Defining and assessing wisdom: A review of the literature. *The American Journal of Geriatric Psychiatry*, *21*(12),

1254–1266.

Baquiran, C. L. C., & Nicoladis, E. (2020). A doctor's foreign accent affects perceptions of competence. *Health Communication*, *35*(6), 726–730.

Batsaikhan, M., He, T.-S., & Li, Y. (2021). Accents, group identity, and trust behaviors: evidence from singapore. *China Economic Review*, *70*, 101702.

Bauer, P. C., & Freitag, M. (2018). Measuring trust. *The Oxford handbook of social and political trust*, *15*.

Baus, C., McAleer, P., Marcoux, K., Belin, P., & Costa, A. (2019). Forming social impressions from voices in native and foreign languages. *Scientific Reports*, *9*(1). doi: 10.1038/s41598-018-36518-6

Baylor, A., & Kim, Y. (2003). The role of gender and ethnicity in pedagogical agent perception. In (pp. 1503–1506). Association for the Advancement of Computing in Education (AACE).

Behlau, M., Madazio, G., Pacheco, C., Vaiano, T., Badaró, F., & Barbara, M. (2023). Coaching strategies for behavioral voice therapy and training. *Journal of Voice*, *37*(2), 295–e1.

Belin, P., Bestelmeyer, P. E. G., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725.

Belin, P., Boehme, B., & McAleer, P. (2019). Correction: The sound of trustworthiness: Acoustic-based modulation of perceived voice personality. *Plos one*, *14*(1), e0211282.

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129–135.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, *10*(1), 122–142.

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. Int.*, *5*(9), 341–345.

Brion, S., Lount Jr, R. B., & Doyle, S. P. (2015). Knowing if you are trusted: Does meta-

accuracy promote trust development? *Social Psychological and Personality Science*, *6*(7), 823–830.

Brockmann-Bauser, M., Van Stan, J. H., Sampaio, M. C., Bohlender, J. E., Hillman, R. E., & Mehta, D. D. (2021). Effects of vocal intensity and fundamental frequency on cepstral peak prominence in patients with voice disorders and vocally healthy controls. *Journal of Voice*, *35*(3), 411–417.

Bryant, D., Borenstein, J., & Howard, A. (2020). Why should we gender? the effect of robot gendering and occupational stereotypes on human trust and perceived competency. In (pp. 13–21).

Burgoon, J. K. (2015). Expectancy violations theory. *The international encyclopedia of interpersonal communication*, 1–9.

Burgoon, J. K., & Hubbard, A. S. E. (2005). Cross-cultural and intercultural applications of expectancy violations theory and interaction adaptation theory. *Theorizing about intercultural communication*, 149–171.

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*(CSCW), 1–19.

Cao, L., Zhao, J., Ren, L., & Zhao, R. (2015). Do in-group and out-group forms of trust matter in predicting confidence in the order institutions? a study of three culturally distinct countries. *International Sociology*, *30*(6), 674–693.

Cascio Rizzo, G. L., & Berger, J. A. (2023). The power of speaking slower. *Available at SSRN*.

Castelfranchi, C., Cesta, A., Conte, R., & Miceli, M. (1993). Foundations for interaction: The dependence theory. In (pp. 59–64). Springer.

Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model*. John Wiley & Sons.

Chan, M. P. Y., & Liberman, M. (2021). An acoustic analysis of vocal effort and speaking style. In (Vol. 45). AIP Publishing.

Chandler, J., Rosenzweig, C., Moss, A. J., Robinson, J., & Litman, L. (2019). Online panels

in social science research: Expanding sampling methods beyond mechanical turk. *Behavior research methods*, *51*, 2022–2038.

Clark, L., Doyle, P., Garaialde, D., Gilmartin, E., Schlögl, S., Edlund, J., ... Edwards, J. (2019). The state of speech in hci: Trends, themes and challenges. *Interacting with computers*, *31*(4), 349–371.

CloudResearch. (2015). *Prime panels by cloudresearch: Online research panel recruitment.* Retrieved from `https://www.cloudresearch.com/products/prime-panels/?ga_ref=home`

Correll, J., Hudson, S. M., Guillermo, S., & Earls, H. A. (2017). Of kith and kin: Perceptual enrichment, expectancy, and reciprocity in face perception. *Personality and Social Psychology Review*, *21*(4), 336–360.

Coupland, N., & Bishop, H. (2007). Ideologised values for british accents 1. *Journal of sociolinguistics*, *11*(1), 74–93.

Couronné, R., Probst, P., & Boulesteix, A.-L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*, 1–14.

Cuddy, A. J. C., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, *40*, 61–149.

Cuddy, A. J. C., Fiske, S. T., Kwan, V. S. Y., Glick, P., Demoulin, S., Leyens, J., ... Sleebos, E. (2009). Stereotype content model across cultures: Towards universal similarities and some differences. *British journal of social psychology*, *48*(1), 1–33.

Dahlbäck, N., Jönsson, A., & Ahrenberg, L. (1993). Wizard of oz studies: why and how. In (pp. 193–200).

Dahlbäck, N., Wang, Q., Nass, C., & Alwin, J. (2007). Similarity is more important than expertise: Accent effects in speech interfaces. In (pp. 1553–1556).

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*(4), 411–419.

Delhey, J., & Welzel, C. (2012). Generalizing trust: How outgroup-trust grows beyond

ingroup-trust. *World Values Research, WVR*, *5*(3).

Del Popolo Cristaldi, F., Granziol, U., Bariletti, I., & Mento, G. (2022). Doing experimental psychological research from remote: how alerting differently impacts online vs. lab setting. *Brain Sciences*, *12*(8), 1061.

Deng, M., Chen, J., Wu, Y., Ma, S., Li, H., Yang, Z., & Shen, Y. (2024). Using voice recognition to measure trust during interactions with automated vehicles. *Applied Ergonomics*, *116*, 104184.

Deutsch, M. (1960). The effect of motivational orientation upon trust and suspicion. *Human relations*, *13*(2), 123–139.

Duffy, B. (2023). *The state of social trust: how the uk comparesinternationally* (Tech. Rep.). The Behavioural Insights Team, The Policy Institute, King's College London.

Eberl, M. (2016). Fisher-yates shuffle. *Arch. Formal Proofs*, *2016*, 19.

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group decision and negotiation*, *22*(5), 897–913.

Elkins, A. C., Derrick, D. C., Burgoon, J. K., & Nunamaker Jr., J. F. (2012). Predicting users' perceived trust in embodied conversational agents using vocal dynamics - proceedings of the 2012 45th hawaii international conference on system sciences. , 579–588. Retrieved from `https://doi.org/10.1109/HICSS.2012.483` doi: 10.1109/HICSS.2012.483

Ellard-Gray, A., Jeffrey, N. K., Choubak, M., & Crann, S. E. (2015). Finding the hidden participant: Solutions for recruiting hidden, hard-to-reach, and vulnerable populations. *International journal of qualitative methods*, *14*(5), 1609406915621420.

Fairbanks, G. (1960). Voice and articulation drillbook.

Fantini, M., Fussi, F., Crosetti, E., & Succo, G. (2017). Estill voice training and voice quality control in contemporary commercial singing: an exploratory study. *Logopedics Phoniatrics Vocology*, *42*(4), 146–152.

Farrús, M., Hernando, J., & Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. International Speech Communication Association (ISCA).

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using g* power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149–1160.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, *39*(2), 175–191.

Feinberg, D. (2022). Voicelab: Software for fully reproducible automated voice analysis. In (pp. 351–355).

Feinberg, D. R., & Cook, O. (2020). Voicelab: Automated reproducible acoustic analysis.

Felippe, A. C. N. d., Grillo, M. H. M. M., & 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*(4), 480–487.

Field, A. (2018). *Discovering statistics using ibm spss statistics*. Sage publications limited.

Fife, D. A., & D'Onofrio, J. (2023). Common, uncommon, and novel applications of random forest in psychological research. *Behavior Research Methods*, *55*(5), 2447–2466.

Fisher, J., Van Heerde, J., & Tucker, A. (2010). Does one trust judgement fit all? linking theory and empirics. *The British Journal of Politics and International Relations*, *12*(2), 161–188.

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current directions in psychological science*, *27*(2), 67–73.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in cognitive sciences*, *11*(2), 77–83.

Freitag, M., & Bauer, P. C. (2013). Testing for measurement equivalence in surveys: Dimensions of social trust across cultural contexts. *Public opinion quarterly*, *77*(S1),

24–44.

Freitag, M., & Bauer, P. C. (2016). Personality traits and the propensity to trust friends and strangers. *The Social Science Journal*, *53*(4), 467–476.

Freitag, M., & Traunmüller, R. (2009). Spheres of trust: An empirical analysis of the foundations of particularised and generalised trust. *European journal of political research*, *48*(6), 782–803.

Frühholz, S., & Schweinberger, S. R. (2021). Nonverbal auditory communication-evidence for integrated neural systems for voice signal production and perception. *Progress in Neurobiology*, *199*, 101948.

Fu, F., Tarnita, C. E., Christakis, N. A., Wang, L., Rand, D. G., & Nowak, M. A. (2012). Evolution of in-group favoritism. *Scientific reports*, *2*(1), 460.

Gabrieli, G., Ng, S., & Esposito, G. (2021). Hacking trust: The presence of faces on automated teller machines (atms) affects trustworthiness. *Behavioral Sciences*, *11*(6), 91.

Gambetta, D. (2000). Can we trust trust. *Trust: Making and breaking cooperative relations*, *13*(2000), 213–237.

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*(3), 953–965.

Gelfand, S. A. (2017). *Hearing: An introduction to psychological and physiological acoustics*. CRC Press.

Germine, L., Nakayama, K., Duchaine, B. C., Chabris, C. F., Chatterjee, G., & Wilmer, J. B. (2012). Is the web as good as the lab? comparable performance from web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review*, *19*, 847–857.

Gheorghiu, A. I., Callan, M. J., & Skylark, W. J. (2020). A thin slice of science communication: Are people's evaluations of ted talks predicted by superficial impressions of the speakers? *Social Psychological and Personality Science*, *11*(1), 117–125.

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*, 102571.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (2000). Measuring trust. *The quarterly journal of economics*, *115*(3), 811–846.

Glanville, J. L., & Shi, Q. (2020). The extension of particularized trust to generalized and out-group trust: The constraining role of collectivism. *Social Forces*, *98*(4), 1801–1828.

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*(2), 214–237.

Goodman, K., & Mayhorn, C. (2023). It's not what you say but how you say it: Examining the influence of perceived voice assistant gender and pitch on trust and reliance. *Applied Ergonomics*, *106*. doi: 10.1016/j.apergo.2022.103864

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition: attitudes, self-esteem, and stereotypes. *Psychological review*, *102*(1), 4.

Groyecka-Bernard, A., Pisanski, K., Frąckowiak, T., Kobylarek, A., Kupczyk, P., Oleszkiewicz, A., ... Sorokowski, P. (2022). Do voice-based judgments of socially relevant speaker traits differ across speech types? *Journal of Speech, Language, and Hearing Research*, *65*(10), 3674-3694. doi: 10.1044/2022_JSLHR-21-00690

Guldner, S., Lavan, N., Lally, C., Wittmann, L., Nees, F., Flor, H., & McGettigan, C. (2024). Human talkers change their voices to elicit specific trait percepts. *Psychonomic Bulletin & Review*, *31*(1), 209–222.

Guo, Q., Zheng, W., Shen, J., Huang, T., & Ma, K. (2022). Social trust more strongly associated with well-being in individualistic societies. *Personality and Individual Differences*, *188*, 111451.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology..

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... Puranen, B. (2022a). World values survey: Round seven-country-pooled datafile version 5.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA*

Secretariat, *12*(10), 8.

Haerpfer, C., Inglehart, R., Moreno, A., Welzel, C., Kizilova, K., Diez-Medrano, J., ... Puranen, B. (2022b). *World values survey wave 7.* Retrieved from `https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp`

Hammarberg, B., Fritzell, B., Gaufin, J., Sundberg, J., & Wedin, L. (1980). Perceptual and acoustic correlates of abnormal voice qualities. *Acta oto-laryngologica*, *90*(1-6), 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*(3), 285–298.

Hardin, R. (2002). *Trust and trustworthiness*. Russell Sage Foundation.

Harrison McKnight, D., & Chervany, N. L. (2001). Trust and distrust definitions: One bite at a time. In (pp. 27–54). Springer.

Heffernan, K. (2004). Evidence from hnr that/s/is a social marker of gender. *Toronto Working Papers in Linguistics*, *23*.

Honing, H., & Reips, U.-D. (2008). Web-based versus lab-based studies: A response to kendall (2008).

Hornsey, M. J. (2008). Social identity theory and self-categorization theory: A historical review. *Social and personality psychology compass*, *2*(1), 204–222.

Horton, J. J., Rand, D. G., & Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, *14*, 399–425.

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*, 103600.

Jalali-najafabadi, F., Gadepalli, C., Jarchi, D., & Cheetham, B. M. G. (2021). Acoustic analysis and digital signal processing for the assessment of voice quality. *Biomedical Signal Processing and Control*, *70*, 103018.

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). *The measurement of the propensity to trust technology.* Springer International Publishing.

Jiang, X., Gossack-Keenan, K., & Pell, M. (2020). To believe or not to believe? how voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology*, *73*(1), 55-79. doi: 10.1177/1747021819865833

Johnson, D. W., & Johnson, F. P. (1991). *Joining together: Group theory and group skills*. Prentice-Hall, Inc.

Judd, C. M., James-Hawkins, L., Yzerbyt, V., & Kashima, Y. (2005). Fundamental dimensions of social judgment: understanding the relations between judgments of competence and warmth. *Journal of personality and social psychology*, *89*(6), 899.

Kamiloğlu, R. G., & Sauter, D. A. (2021). Voice production and perception..

Kang, G. S., & Heide, D. A. (1992). Canned speech for tactical voice message systems. In (pp. 47–56). IEEE.

Kang, S. K., & Bodenhausen, G. V. (2015). Multiple identities in social perception and interaction: Challenges and opportunities. *Annual review of psychology*, *66*(1), 547–574.

Kaur, N., & Singh, P. (2023). Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, *56*(7), 5837–5880.

Keeley, J. W., English, T., Irons, J., & Henslee, A. M. (2013). Investigating halo and ceiling effects in student evaluations of instruction. *Educational and Psychological Measurement*, *73*(3), 440–457.

Kepuska, V., & Bohouta, G. (2018). Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home). In (pp. 99–103). IEEE.

Kim, H. H.-S. (2018). Particularized trust, generalized trust, and immigrant self-rated health: cross-national analysis of world values survey. *public health*, *158*, 93–101.

Kim, J., Gonzalez-Pumariega, G., Park, S., & Fussell, S. R. (2023). Urgency builds trust: A voice agent's emotional expression in an emergency. In (pp. 343–347).

Kirk, R. E. (1996). Practical significance: A concept whose time has come. *Educational and psychological measurement*, *56*(5), 746–759.

Klofstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political psychology*, *37*(5), 725–738.

Klofstad, C. A., Anderson, R. C., & Nowicki, S. (2015). Perceptions of competence, strength, and age influence voters to select leaders with lower-pitched voices. *PloS one*, *10*(8), e0133779.

Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), 2698–2704.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? a cross-country investigation. *The Quarterly journal of economics*, *112*(4), 1251–1288.

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. , 174–193. Retrieved from https://doi.org/10.1007/978-3-030-60117-1_13 doi: 10.1007/978-3-030-60117-1_13

Ko, S. J., Judd, C. M., & Stapel, D. A. (2009). Stereotyping based on voice in the presence of individuating information: Vocal femininity affects perceived competence but not warmth. *Personality and Social Psychology Bulletin*, *35*(2), 198–211.

Kouravanas, N., & Pavlopoulos, A. (2022). Social robots: the case of robot sophia. *Homo Virtualis*, *5*(1), 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.

Krueger, F. (2021). *The neurobiology of trust*. Cambridge University Press.

Kumar, S., Kumar, S., Sathe, K., & Pati, J. (2025). Advancing bangla text-to-speech synthesis using a vits-based model with a custom dataset and comprehensive evaluation. *Discover Computing*, *28*(1), 183.

Kushins, E. R. (2014). Sounding like your race in the employment process: An experiment

on speaker voice, race identification, and stereotyping. *Race and Social Problems*, *6*, 237–248.

Kühne, K., Fischer, M., & 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*. doi: 10.3389/fnbot.2020.593732

Large, D. R., & Burnett, G. E. (2014). The effect of different navigation voices on trust and attention while using in-vehicle navigation systems. *Journal of safety research*, *49*, 69–e1.

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. In (pp. 49–59).

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), R143–R145.

Lau, D. C., Lam, L. W., & Wen, S. S. (2014). Examining the effects of feeling trusted by supervisors in the workplace: A self-evaluative perspective. *Journal of Organizational Behavior*, *35*(1), 112–127.

Lavan, N. (2023). The time course of person perception from voices: A behavioral study. *Psychological Science*, 09567976231161565.

Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic bulletin & review*, *26*, 90–102.

Lavan, N., Mileva, M., & McGettigan, C. (2021). How does familiarity with a voice affect trait judgements? *British Journal of Psychology*, *112*(1), 282–300.

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. In (pp. 1–15). IGI Global.

Leongómez, J. D., Pisanski, K., Reby, D., Sauter, D., Lavan, N., Perlman, M., & Varella Valentova, J. (2021). *Voice modulation: from origin and mechanism to social impact* (Vol. 376) (No. 1840). The Royal Society.

Leung, Y., Oates, J., & Chan, S. P. (2018). Voice, articulation, and prosody contribute

to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, *61*(2), 266–297.

Liao, Y., & He, J. (2020). Racial mirroring effects on human-agent interaction in psychotherapeutic conversations. In (pp. 430–442).

Lieberman, P., Laitman, J. T., Reidenberg, J. S., & Gannon, P. J. (1992). The anatomy, physiology, acoustics and perception of speech: essential elements in analysis of the evolution of human speech. *Journal of Human Evolution*, *23*(6), 447–467.

Lim, M. Y., Lopes, J. D. A., 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). , 538–545. Retrieved from https://doi.org/10.1109/RO-MAN53752.2022.9900772 doi: 10.1109/RO-MAN53752.2022.9900772

Lima, L., Furtado, V., Furtado, E., & Almeida, V. (2019). Empirical analysis of bias in voice-based personal assistants. In (pp. 533–538).

Linville, P. W., Salovey, P., & Fischer, G. W. (1986). Stereotyping and perceived distributions of social characteristics: An application to ingroup-outgroup perception. Academic Press.

Linville, S. E. (2002). Source characteristics of aged voice assessed from long-term average spectra. *Journal of Voice*, *16*(4), 472–479.

Löckenhoff, C. E., De Fruyt, F., Terracciano, A., McCrae, R. R., De Bolle, M., Costa, P. T., … Alcalay, L. (2009). Perceptions of aging across 26 cultures and their culture-level associates. *Psychology and aging*, *24*(4), 941.

Löfqvist, A. (1986). The long-time-average spectrum as a tool in voice research. *Journal of phonetics*, *14*(3-4), 471–475.

Mahendru, H. C. (2014). Quick review of human speech production mechanism. *International Journal of Engineering Research and Development*, *9*(10), 48–54.

Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker's personality are correlated across differing content and stimulus type. *PLoS ONE*, *13*(10). doi:

10.1371/journal.pone.0204991

Maloy, J. S. (2009). Two concepts of trust. *The Journal of Politics*, *71*(2), 492–505.

Maltezou-Papastylianou, C., Russo, R., Wallace, D., Harmsworth, C., & Paulmann, S. (2022). Different stages of emotional prosody processing in healthy ageing-evidence from behavioural responses, erps, tdcs, and trns. *Plos one*, *17*(7), e0270934.

Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2023). *Can i trust you? discovering how trustworthiness is communicated through vocal cues.* OSF. doi: 10.17605/OSF.IO/SFB3G

Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2024a). Acoustic classification of speech with trustworthy intent. In (pp. 961–964).

Maltezou-Papastylianou, C., Scherer, R., & Paulmann, S. (2024b). Trustworthy intent in speech (tis) corpora dataset. Retrieved from `https://doi.org/10.17605/OSF.IO/45D8J` doi: 10.17605/OSF.IO/45D8J

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*. Retrieved from `https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2025.1495456` doi: 10.3389/fpsyg.2025.1495456

Mara, M., Appel, M., & Gnambs, T. (2022). Human-like robots and the uncanny valley. *Zeitschrift für Psychologie*.

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*(3), 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. In (pp. 1–8).

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709–734.

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say 'hello'? personality

impressions from brief novel voices. *PLoS ONE*, *9*(3). doi: 10.1371/journal.pone .0090779

McDougall Jr, G. J., Simpson, G., & Friend, M. L. (2015). Strategies for research recruitment and retention of older adults of racial and ethnic minorities. *Journal of gerontological nursing*, *41*(5), 14–23.

McGettigan, C., & Lavan, N. (2023). Investigating the effects of talker age and listener age on trait perception from adult voices. IPA.

Mergler, N. L., & Goldstein, M. D. (1983). Why are there old people: Senescence as biological and cultural preparedness for the transmission of information. *Human Development*, *26*(2), 72–90.

Mileva, M. (2025). Multimodal person evaluation: First impressions from faces, voices, and names. *Journal of Personality and Social Psychology*.

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. (2018). Audiovisual integration in social evaluation. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(1), 128-138. doi: 10.1037/xhp0000439

Mileva, M., Tompkinson, J., Watt, D., & Burton, A. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality and Social Psychology Bulletin*, *46*(4), 617-625. doi: 10.1177/0146167219867965

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*(3), 241–259.

Montepare, J. M., & Zebrowitz, L. A. (1998). Person perception comes of age: The salience and significance of age in social judgments. In (Vol. 30, pp. 93–161). Elsevier.

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*(1), 64–94.

Moreno, R., & Flowerday, T. (2006). Students' choice of animated pedagogical agents in science learning: A test of the similarity-attraction hypothesis on gender and ethnicity. *Contemporary educational psychology*, *31*(2), 186–207.

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 & automation magazine*, *19*(2), 98–100.

Moussalli, S., & Cardoso, W. (2017). Can you understand me? speaking robots and accented speech. *CALL in a climate of change: Adapting to turbulent global conditions-short papers from EuroCALL*, 217–221.

Mulac, A., & Giles, H. (1996). 'your're only as old as you sound': Perceived vocal age and social meanings. *Health Communication*, *8*(3), 199–215.

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. Retrieved from `https://doi.org/10.1145/2559206.2581231` doi: 10.1145/2559206.2581231

Naef, M., & Schupp, J. (2009). Measuring trust: Experiments and surveys in contrast and combination.

Nam, C. S., & Lyons, J. B. (2020). *Trust in human-robot interaction*. Academic Press.

Nass, C., & Lee, K. M. (2000). Does computer-generated speech manifest personality? an experimental test of similarity-attraction. In (pp. 329–336).

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In (pp. 72–78).

Nass, C. I., & Brave, S. (2005). *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge.

Newton, K., & Zmerli, S. (2011). Three forms of trust and their association. *European Political Science Review*, *3*(2), 169–200.

Nick, T. G., & Campbell, K. M. (2007). Logistic regression. *Topics in biostatistics*, 273–301.

Nomura, T., Suzuki, T., Kanda, T., & Kato, K. (2006a). Altered attitudes of people toward robots: Investigation through the negative attitudes toward robots scale. In (Vol. 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*(3), 437–454.

Norval, M., Wang, Z., & Sun, Y. (2023). Synthetic speech data generation using generative adversarial networks. In *International conference on cloud computing and computer networks* (pp. 117–126).

O'Connor, J., & Barclay, P. (2018). High voice pitch mitigates the aversiveness of antisocial cues in men's speech. *British Journal of Psychology*, *109*(4), 812-829. doi: 10.1111/bjop.12310

Ohala, J. J. (1983). Cross-language use of pitch: an ethological view. *Phonetica*, *40*(1), 1–18.

Ohala, J. J. (1995). The frequency code underlies the sound-symbolic use of voice pitch. *Sound symbolism*, 325–347.

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 and Review*, *24*(3), 856-862. doi: 10.3758/s13423-016-1146-y

Ostrom, E., & Walker, J. (2003). *Trust and reciprocity: Interdisciplinary lessons for experimental research*. Russell Sage Foundation.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan–a web and mobile app for systematic reviews. *Systematic reviews*, *5*, 1–10.

Pabon, P., Stallinga, R., Södersten, M., & Ternström, S. (2014). Effects on vocal range and voice quality of singing voice training: the classically trained female voice. *Journal of Voice*, *28*(1), 36–51.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Brennan, S. E. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, *88*, 105906.

Page, M. J., Moher, D., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Brennan, S. E. (2021). Prisma 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *bmj*, *372*.

Pargent, F., Schoedel, R., & Stachl, C. (2023). Best practices in supervised machine learning:

A tutorial for psychologists. *Advances in Methods and Practices in Psychological Science*, *6*(3), 25152459231162559.

Peng, C.-Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The journal of educational research*, *96*(1), 3–14.

Peng, Z., Wang, Y., Meng, L., Liu, H., & Hu, Z. (2019). One's own and similar voices are more attractive than other voices. *Australian Journal of Psychology*, *71*(3), 212–222.

Perrachione, T. K., Chiao, J. Y., & Wong, P. C. M. (2010). Asymmetric cultural effects on perceptual expertise underlie an own-race bias for voices. *Cognition*, *114*(1), 42–55.

Ponsot, E., Burred, J., Belin, P., & Aucouturier, J.-J. (2018). Cracking the social code of speech prosody using reverse correlation. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(15), 3972-3977. doi: 10.1073/pnas.1716090115

Prolific. (2014). *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. C. (2007). Men's voices as dominance signals: vocal fundamental and formant frequencies influence dominance attributions among men. *Evolution and Human Behavior*, *28*(5), 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*(3), 397–419.

Razin, Y. S., & Feigh, K. M. (2023). Converging measures and an emergent model: A meta-analysis of human-automation trust questionnaires. *arXiv preprint arXiv:2303.13799*.

Reetz, H., & Jongman, A. (2020). *Phonetics: Transcription, production, acoustics, and perception*. John Wiley & Sons.

Rehman, M. U., Shafique, A., Jamal, S. S., Gheraibia, Y., & Usman, A. B. (2024). Voice disorder detection using machine learning algorithms: An application in speech and language pathology. *Engineering Applications of Artificial Intelligence*, *133*, 108047.

Riek, L. D. (2012). Wizard of oz studies in hri: a systematic review and new reporting guidelines. *Journal of Human-Robot Interaction*, *1*(1), 119–136.

Riener, A., Jeon, M., & Alvarez, I. (2022). *User experience design in the era of automated driving*. Springer.

Rodero, E., Mas, L., & Blanco, M. (2014). The influence of prosody on politicians' credibility. *Journal of applied linguistics & professional practice*, *11*(1).

Rojas, S., Kefalianos, E., & Vogel, A. (2020). How does our voice change as we age? a systematic review and meta-analysis of acoustic and perceptual voice data from healthy adults over 50 years of age. *Journal of Speech, Language, and Hearing Research*, *63*(2), 533–551.

Rotter, J. B. (1967). A new scale for the measurement of interpersonal trust. *Journal of personality*.

Scheibe, S., Kunzmann, U., & Baltes, P. B. (2009). 16 new territories of positive life-span development: Wisdom and life longings. *The Oxford handbook of positive psychology*, 171.

Scherer, S., Stratou, G., Gratch, J., & Morency, L.-P. (2013). Investigating voice quality as a speaker-independent indicator of depression and ptsd. In (pp. 847–851).

Schild, C., Stern, J., & Zettler, I. (2020). Linking men's voice pitch to actual and perceived trustworthiness across domains. *Behavioral Ecology*, *31*(1), 164–175.

Schilke, O., Reimann, M., & Cook, K. S. (2021). Trust in social relations. *Annual Review of Sociology*, *47*, 239–259.

Schirmer, A., Chiu, M., Lo, C., Feng, Y.-J., & Penney, T. (2020). Angry, old, male – and trustworthy? how expressive and person voice characteristics shape listener trust. *PLoS ONE*, *15*(5). doi: 10.1371/journal.pone.0232431

Schweinberger, S. R., Kawahara, H., Simpson, A. P., Skuk, V. G., & Zäske, R. (2014). Speaker perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 15–25.

Seaborn, K., Miyake, N. P., Pennefather, P., & Otake-Matsuura, M. (2021). Voice in human-agent interaction: A survey. *ACM Computing Surveys (CSUR)*, *54*(4), 1–43.

Sebastian, R. J., & Ryan, E. B. (2018). Speech cues and social evaluation: Markers of ethnicity, social class, and age. In (pp. 112–143). Routledge.

Seligman, A. B. (2000). *The problem of trust*. Princeton University Press.

Sharma, D., Levon, E., & Ye, Y. (2022). 50 years of british accent bias: Stability and lifespan change in attitudes to accents. *English World-Wide*, *43*(2), 135–166.

Shen, Z., Elibol, A., & Chong, N. Y. (2020). Understanding nonverbal communication cues of human personality traits in human-robot interaction. *IEEE/CAA Journal of Automatica Sinica*, *7*(6), 1465–1477.

Shigemi, S., Goswami, A., & Vadakkepat, P. (2018). Asimo and humanoid robot research at honda. *Humanoid robotics: A reference*, *55*, 90.

Simpson, J. A. (2007). Foundations of interpersonal trust. *Social psychology: Handbook of basic principles*, *2*, 587–607.

Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in psychology*, *6*, 978.

Smith, S. M., & Shaffer, D. R. (1995). Speed of speech and persuasion: Evidence for multiple effects. *Personality and Social Psychology Bulletin*, *21*(10), 1051–1060.

Smith, S. S. (2010). Race and trust. *Annual review of sociology*, *36*, 453–475.

Soroka, S., Helliwell, J. F., & Johnston, R. (2003). Measuring and modelling trust. *Diversity, social capital and the welfare state*, 279–303.

Stewart, M. A., & Ryan, E. B. (1982). Attitudes toward younger and older adult speakers: Effects of varying speech rates. *Journal of language and social psychology*, *1*(2), 91–109.

Stumpf, L., Kadirvelu, B., Waibel, S., & Faisal, A. A. (2024). Speaker-independent dysarthria severity classification using self-supervised transformers and multi-task learning. *arXiv preprint arXiv:2403.00854*.

Sun, T., Liu, Y., Zhang, Q., Jiang, J., Xiong, X., Wen, J., … Jeung, J. (2024). Cross-cultural differences in recruiting older adults in design research activities: Insights from china and the uk. *Innovation in Aging*, *8*($Supplement_1$), 720–720.

Sundberg, J., Patel, S., Bjorkner, E., & Scherer, K. R. (2011). Interdependencies among voice source parameters in emotional speech. *IEEE Transactions on Affective Computing*, *2*(3), 162–174.

Tamagawa, R., Watson, C. I., Kuo, I. H., MacDonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, *3*, 253–262.

Tanis, M., & Postmes, T. (2005). A social identity approach to trust: Interpersonal perception, group membership and trusting behaviour. *European journal of social psychology*, *35*(3), 413–424.

Taylor, L., & Rommelfanger, K. S. (2022). Mitigating white western individualistic bias and creating more inclusive neuroscience. *Nature Reviews Neuroscience*, *23*(7), 389–390.

Ter Kuile, H., Kluwer, E. S., Finkenauer, C., & Van der Lippe, T. (2017). Predicting adaptation to parenthood: The role of responsiveness, gratitude, and trust. *Personal Relationships*, *24*(3), 663–682.

Thomas, W. H. (2004). *What are old people for?: How elders will save the world*. Publisher: VanderWyk & Burnham.

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., & Bernstein, A. (2021). Female by default? – exploring the effect of voice assistant gender and pitch on trait and trust attribution - extended abstracts of the 2021 chi conference on human factors in computing systems. Retrieved from `https://doi.org/10.1145/3411763.3451623` doi: 10.1145/3411763.3451623

Torre, I., Goslin, J., & White, L. (2020, 4). If your device could smile: People trust happy-sounding artificial agents more. *Comput. Hum. Behav.*, *105*. Retrieved from `https://doi.org/10.1016/j.chb.2019.106215` doi: 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. In (pp. 1–6).

Torre, I., White, L., & Goslin, J. (2016). Behavioural mediation of prosodic cues to implicit judgements of trustworthiness. *Speech Prosody 2016*.

Tsantani, M., Belin, P., Paterson, H., & McAleer, P. (2016). Low vocal pitch preference drives first impressions irrespective of context in male voices but not in female voices. *Perception*, *45*(8), 946-963. doi: 10.1177/0301006616643675

Tschannen-Moran, M., & Hoy, W. K. (2000). A multidisciplinary analysis of the nature, meaning, and measurement of trust. *Review of educational research*, *70*(4), 547–593.

Tuckman, B. W. (1965). Developmental sequence in small groups. *Psychological bulletin*, *63*(6), 384.

Tuomela, R., & Tuomela, M. (2005). Cooperation and trust in group context. *Mind & Society*, *4*, 49–84.

Turner, J. C. (2010). Towards a cognitive redefinition of the social group. Psychology Press.

Uittenhove, K., Jeanneret, S., & Vergauwe, E. (2023). From lab-testing to web-testing in cognitive research: Who you test is more important than how you test. *Journal of Cognition*, *6*(1).

Uslaner, E. M. (1999). Democracy and social capital. *Democracy and trust*, *121*, 150.

Uslaner, E. M. (2002). The moral foundations of trust. *Available at SSRN 824504*.

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*(6), 977–983.

Weinschenk, S., & Barker, D. T. (2000). *Designing effective speech interfaces*. John Wiley & Sons, Inc.

Weinstein, N., Zougkou, K., & Paulmann, S. (2018). You 'have'to hear this: Using tone of voice to motivate others. *Journal of Experimental Psychology: Human Perception and Performance*, *44*(6), 898.

Yamagishi, T. (2003). Cross-societal experimentation on trust: A comparison of the united states and japan.

Yamagishi, T., & Yamagishi, M. (1994). Trust and commitment in the united states and japan. *Motivation and emotion*, *18*, 129–166.

Yokoyama, H., & Daibo, I. (2012). Effects of gaze and speech rate on receivers' evaluations of persuasive speech. *Psychological Reports*, *110*(2), 663-676. doi: 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.

Zebrowitz, L. A., & Montepare, J. (2013). The ecological approach to person perception: Evolutionary roots and contemporary offshoots. In (pp. 81–113). Psychology Press.

Zheng, F., Zhang, G., & Song, Z. (2001). Comparison of different implementations of mfcc. *Journal of Computer science and Technology*, *16*, 582–589.