Behaviour & Information Technology

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tbit20

Examining the persuasiveness of text and voice agents: prosody aligned with information structure increases human-likeness, perceived personalisation and brand attitude

Hilde Voorveld, Andreas Panteli, Yoni Schirris, Carolin Ischen, Evangelos Kanoulas & Tom Lentz

To cite this article: Hilde Voorveld, Andreas Panteli, Yoni Schirris, Carolin Ischen, Evangelos Kanoulas & Tom Lentz (14 Nov 2024): Examining the persuasiveness of text and voice agents: prosody aligned with information structure increases human-likeness, perceived personalisation and brand attitude, Behaviour & Information Technology, DOI: 10.1080/0144929X.2024.2420871

To link to this article: https://doi.org/10.1080/0144929X.2024.2420871

BEHAVIOUR & INFORMATION TECHNOLOGY
https://doi.org/10.1080/0144929X.2024.2420871

Examining the persuasiveness of text and voice agents: prosody aligned with information structure increases human-likeness, perceived personalisation and brand attitude

Hilde Voorveld[a], Andreas Panteli[b], Yoni Schirris[b], Carolin Ischen[a], Evangelos Kanoulas[c] and Tom Lentz[d]

[a]Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, The Netherlands; [b]Netherlands Cancer Institute (NKI), Amsterdam, The Netherlands; [c]Informatics Institute University of Amsterdam, The Netherlands; [d]Tilburg School of Humanities and Digital Sciences, Tilburg University, Tilburg, The Netherlands

ABSTRACT

To give product and brand recommendations, marketers make use of conversational agents which increasingly communicate via voice rather than text. Existing research comparing the persuasiveness of text and voice agents showed mixed results. The quality of the speech synthesis employed may strongly influence consumers' responses. This study investigates to what extent a voice agent with pragmatically aligned prosody is more persuasive (i.e. yields a more positive brand attitude) than an agent with a standard voice or text, and whether perceived human-likeness and perceived personalisation provide an underlying mechanism to explain these differences. In an experiment ($n = 212$), participants interacted with a conversational agent that recommended a camera. Results showed that a voice agent using prosody aligned to the information state of the user is more persuasive than a text agent. This effect is mediated by perceived human-likeness and perceived personalisation. Hence, aligned prosody can make synthetic speech meet a certain quality threshold to be perceived as more human-like. Theoretically, this study helps to unravel *why* conversational agents with human-like features are more persuasive.

ARTICLE HISTORY
Received 28 November 2023
Accepted 19 October 2024

KEYWORDS
Conversational agents; virtual assistants; voice; speech; prosody; modality

1. Introduction

Conversational agents are increasingly integrated into consumers' lives, as technologies like virtual assistants are more and more often used by marketers to improve consumer experience. Some authors even refer to this as a 'chatbot tsunami' (Grudin and Jacques 2019, 4). Next to assistants that communicate with consumers via text, the use of agents that employ voice or speech has become more popular in recent years (e.g. E Marketer 2022; Reicherts et al. 2022). One of the main reasons why voice agents have been so frequently adopted is that marketers believe that voice agents are more persuasive than text agents and that voice commerce provides 'a great opportunity for their brand'(Ischen et al. 2022; Mari, Mandelli, and Algesheimer 2020, 417). For example, they believe that branded recommendations made by voice agents (vs. text agents) lead to more positive attitudes towards the recommended brands, mainly because they believe that 'a machine that can create speech should be judged as more human-like than a machine that creates text' (Schroeder and Schroeder 2018, 472). Although there is a lot of research on the voice in related fields like social robotics (Xu, Chen, and You 2023), synthesised voices (Gong and Lai 2001; Lee et al. 2006) or avatars (Qui and Bensabat 2009), existing research comparing the persuasiveness of text vs. voice agents is scarce (Rapp, Curti, and Boldi 2021), and shows mixed findings. Cho, Molina, and Wang (2019) find voice, in comparison to text, to evoke more positive attitudes towards the assistant Cortana. In contrast, Ischen et al. (2022) and Voorveld and Araujo (2020) find no differences or find voice to be less persuasive. Our study adds to the limited body of empirical studies comparing the persuasiveness of voice vs. text agents.

These contrasting findings in earlier studies may result from the quality of the speech synthesis employed (also see Im et al. 2023; Schroeder and Epley 2016). The general quality of synthetic speech, especially its

CONTACT Hilde Voorveld h.a.m.voorveld@uva.nl Amsterdam School of Communication Research, University of Amsterdam, P.O. box 15791, Amsterdam, 1001 NG, The Netherlands

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0144929X.2024.2420871

naturalness, has markedly improved recently. Currently, researchers are focusing on its expressivity, and specifically on prosody (Rodero 2017; Torresquintero et al. 2021). Prosody is, for the purposes of this paper, the marking of parts of an utterance using vocal pitch and speech rate, and one of its functions is to mark information structure, the relation between parts of an utterance and the knowledge of listeners. We posit that prosody is especially relevant to the persuasiveness of voice agents. English, like most Germanic languages, uses prosody to express a distinction between knowledge a listener already has and new information (Vallduví and Engdahl 1996). Human interlocutors build a common ground in a dialogue by taking each other's viewpoints and knowledge into account (Brown and Yule 1983), and such alignment may give users the feeling that an agent responds to the input that was given, and/or that it is aware of what the user already knows. Therefore, a user may feel that an agent using aligned prosody understands their needs and that its recommendation is specifically designed for them. Thus, using aligned prosody may be crucial for the quality of speech synthesis employed in voice agents.

Therefore, using an experiment ($n = 212$ US participants) we propose and test whether a conversational agent using voice will be more persuasive than a text agent if (1) it uses current off-the-shelf state-of-the-art synthetic speech (i.e. standard voice) and (2) if, additionally, its prosody is aligned with the context and knowledge of the user (i.e. aligned prosody voice). We also investigate the role of two underlying processes that may explain why this would be the case: perceived human-likeness and perceived personalisation. More specifically, we test whether an agent with standard speech is perceived as more human-like, and more personalised than a text agent. Additionally, as aligned prosody may be necessary to cross the threshold, we compare the perceived human-likeness, and perceived personalisation of a speech agent with aligned prosody to a text agent. This is important to further develop theoretical explanations on *why* adding human-like features to conversational agents may lead to stronger persuasive effects (in our study defined as people's attitudes towards the recommended brand).

Practically, this paper provides evidence-based advice for marketers when considering *whether* and *how* to offer voice versus text agents to give product recommendations. For professionals working on text-to-speech systems, this paper provides advice on a feature of speech – prosody reflecting the information state of speaker and hearer – that is often neither annotated in training data nor part of the loss function used to train deep neural networks behind speech synthesis and therefore generally absent. Finally, empirical knowledge on consumer responses towards specific human-like features in machines may help developers and marketers to further understand the impact of their efforts when designing conversational agents with human-like features.

2. Theoretical background

2.1. Conceptualisation and earlier empirical research on voice vs. text

In this study, persuasion is defined as the formation or change of attitudes, cognitions, and behaviours (McGuire 1989; Perloff 2021). We are specifically interested in attitudinal responses because persuasion attempts are often aimed at changing a recipient's opinion of an object (e.g. Petty, Fabrigar, and Wegener 2003; Vaughan-Johnston et al. 2021). We further specify this as the attitude towards a brand recommended by an agent. The distinction between a text and voice agent is conceptualised as response modality (Schroeder and Schroeder 2018); we focus on how an agent is communicating with the user, not how a user is communicating with an agent (input modality).

There is a large and relevant body of research on related topics like text-based chatbots (see Rapp, Curti, and Boldi 2021), speech interfaces in the field of human–computer interaction (see Clark et al. 2019), and social robots (see Xu, Chen, and You 2023). It goes beyond the scope of this paper to provide a comprehensive literature review of the research in these adjacent fields, but we would like to summarise some key findings most relevant to our paper by discussing some relevant literature reviews and meta-analyses that have recently been conducted. A review on text-based chatbots (Rapp, Curti, and Boldi 2021) synthesises 83 papers along the relevant main themes they have identified: chatbot acceptance, experiencing the chatbot, conversational issues, emotional experience and expression, and humanness. The review gives important insight into the role of humanness in text-based chatbots and the 25 papers that fit within this theme show that people consider different dimensions, and rely on different cues (including linguistics and language style) when ascribing humanness to text-based chatbots and that considering chatbots as humans may have both positive and negative outcomes. The review also concludes that future theory development would benefit from research directly comparing text-based chatbots with other similar conversational technologies like speech-based agents. Finally, as chatbots' persuasiveness was not identified as one of the key

themes, this can be seen as still an under-researched area.

A recent review on speech interfaces reviewing 99 papers (Clark et al. 2019) concluded, amongst others, that most papers focused on usability, and not persuasive outcomes, and are rather diverse, or even fragmented. When it comes to comparing modalities in speech interface research, research mostly focused on input modality; comparing speech vs text/ keyboard and mouse input, rather than output is the focus of our paper.

A recent meta-analysis on social robots (including virtual assistants) focused on the two dependent variables social presence and trust, of which trust can be related to further attitudinal change (Xu, Chen, and You 2023). The study investigated the effects of different types of social cues (appearance, movements, language and voice) and showed that manipulating social robots' voices has only a small-sized effect on trust and social presence (based on eight effect sizes) (Xu, Chen, and You 2023). The authors here compared (more) human-like voices with synthetic or machine-like voices and expected larger effect sized but argue that a reason might be that synthetic voices have been designed to be increasingly human-like in the past years.

Whereas there is some earlier research focusing on the influence of voice or vocal cues on trust and it is typically assumed that trust can lead to further attitudinal and behavioural change (see Chandra, Shirish, and Srivastava 2022; Xu, Chen, and You 2023), earlier research directly comparing voice vs text hardly focused on persuasiveness in terms of attitudinal responses.

The few studies that empirically compared voice- vs. text-based agents in terms of attitudes towards the recommended brand provided mixed findings. Voorveld and Araujo (2020) showed in a scenario-based study that a text-based recommendation via a smartphone resulted in a more positive brand attitude than a voice-based recommendation via a speaker. Ischen et al. (2022) used conversational agents specifically developed for experimental research and found that a text-based agent was perceived as more human-like than voice-based agents, which in turn resulted in more positive brand attitudes and higher purchase intentions. In contrast, Cho, Molina, and Wang (2019) conducted a study with the assistant Cortana and found a positive effect of voice interaction (in comparison to text) on attitudes towards Cortana which was mediated by perceptions of human-like characteristics, though only for utilitarian tasks. The contrasting findings suggest that differences in persuasiveness do not only depend on *whether* voice is used in comparison to text but also that other factors like voice quality are important.

2.2. Speech synthesis and prosody

2.2.1. Current state-of-the-art

The literature cited above does not concern speech synthesis with the improved naturalness provided by WaveNet technology, which became available for general use in about 2018 (van den Oord et al. 2018). The speech generated using WaveNet is characterised by human-like variation and a lack of disfluencies, making it potentially human-like enough to instil more positive attitudes towards agents using such synthesis (e.g. Im et al. 2023). However, written text does not fully specify how an utterance should sound. Thus, text-to-speech systems cannot always generate the right speech. Prosody is especially hard to predict from text alone (Mohan et al. 2021), in spite of impressive work on predicting appropriate prosody from text (e.g. Hodari et al. 2021; Stephenson et al. 2022; Zou et al. 2021). In addition, interlocuters may adapt their general prosody to their interlocutors (see e.g. Suzuki and Katagiri 2007, and references therein). It may be relatively straightforward to measure a user's general tone of voice and align synthetic speech prosodically in this respect. However, the alignment of information status especially will be hard to predict from text, as explained below. Thus, its prosodic expression will be lacking in current off-the-shelf text-to-speech models and conversational voice agents that use them.

2.2.2. Alignment, information status and prosodic marking

In a dialogue, interlocutors align their lexical choices, references and syntax (Pickering and Garrod 2004). Both interlocutors build models of the other's knowledge. Many linguistic features depend on interlocutor models, e.g. the use of definite articles to refer only to things that have been made 'copresent' in the dialogue (Clark and Marshall 1981). Information structure concerns how information fits the listener's knowledge and is thus part of interlocutor modelling. We follow Vallduví and Engdahl's (1996) tripartite distinction of information as either the focus or ground, with ground subdivided into link and tail. The *focus* is something for which the hearer could entertain multiple alternatives (see Krifka, Féry, and Fanselow 2007). It is helpful to use questions to identify and illustrate focus; the information given in an answer that is not in the question is generally

the focus. For example, when the utterance 'I prefer relaxed vacations' is given as answer to the question 'Which kind of vacations do you prefer?', the focus is 'relaxed', but if the question was 'What do you prefer?', the focus would be 'relaxed vacations'.

The *ground* of an utterance is an already-known entity about which new information is given. *Link* is one form of ground, and contrasts with the other form, *tail*, in prominence; link is the prominent (but known) part of the utterance, while link is neither new nor prominent information. To find a link, one can use a 'What about' question sentence, e.g. 'What about vacations? Which kind do you prefer?', which makes 'vacations' in the answer 'I prefer relaxed vacations' the link and 'I prefer' the tail.

Crucially, phrases obtain information structure status only in a conversation context, in contrast to e.g. syntactic status. Languages vary in how they mark information structure, e.g. German employs word order, while Japanese uses discourse particles (see Vallduví and Engdahl 1996, for more). In English, information status is mainly marked by prosody, not reflected in written text.

It is important to distinguish between information structure statuses, e.g. focus, and their marking. Focus is marked in English with a pitch accent, often called H*, a focus accent or confusingly just 'focus'. The accent is, a steep rise followed by a quick fall. A link is marked by a pitch fall, often annotated as an L* (though Vallduví and Engdahl 1996, refer to it as L + H*), and a tail is unaccented. As information status, nor these accents, are part of the text, text-to-speech synthesis cannot accurately generate accents aligned with information status. This paper investigates the importance of aligned prosody for perceived human-likeness.

2.3. The role of perceived human-likeness

2.3.1. The influence of aligned prosody on perceived human-likeness

Human-likeness or anthropomorphism is a central concept in research on conversational agents and is believed to increase acceptance of such agents (e.g. Letheren et al. 2021; Moussawi and Benbunan-Fich 2021). Early research for example already showed that people generally like a real human voice more than a synthesised voice (Gong and Lai 2001). Related research from the field of social robotics has shown that robots or virtual avatars with a human voice are perceived as more credible, trustworthy, and socially present than a machine- or synthetic voice (Chérif and Lemoine 2019; Xu 2019).

Anthropomorphism can be defined as 'the assignment of human traits and characteristics to computers' (Nass and Moon 2000, 82), and is either mindful or mindless. Mindful anthropomorphism is the sincere or conscious belief that a computer, technology, or in our case conversational agent, is human-like. Mindless anthropomorphism is the indirect attribution of human-like characteristics (e.g. likeable, personal, or sociable) to an agent when exposed to anthropomorphic cues (Kim and Sundar 2012). Even if someone consciously denies the human-likeness of an agent, it is possible they mindlessly attribute human-like characteristics to it (Epley et al. 2008; Kim and Sundar 2012). We refer to both forms of anthropomorphism together as human-likeness, and empirically assess human-likeness with measurements of mindless and mindful anthropomorphism, in line with previous research (Ischen et al. 2022).

The Computers Are Social Actors paradigm (CASA) is used in many recent works on conversational agents and encompasses that people respond socially to computers when they implement social cues (Nass, Steuer, and Tauber 1994; Reeves and Nass 1996). As mentioned above, a recent meta-analysis by Xu, Chen, and You (2023) has shown small-sized positive effects of social cues on social presence with and trust in social robots and related technologies. Additionally, a literature review (Van Pinxteren, Pluymaekers, and Lemmink 2020) on conversational agents in service encounters has investigated the effects of social cues on relational outcomes such as trust, satisfaction and liking, and has found that overall the appearance of conversational agents positively influences such outcomes when it resembles a human (e.g. in physical appearance but also in behaviour such as nodding, eye contact).

Stemming from this paradigm, it is argued that auditory cues (i.e. cues that can be heard, next to words themselves, Feine et al. 2019) can function as social cues that can influence (human-like) perceptions of conversational agents (for an overview see Feine et al. 2019). A simple read of the literature on human-likeness perceptions might suggest that adding *any* human cue, so also a standard synthetic voice, to a machine will induce anthropomorphism (Schroeder and Schroeder 2018). However, results of studies on different types of conversational agents suggest that this is not always the case.

On the one hand, research showed that voice is perceived as more human-like than text, in the context of virtual assistants (Cho, Molina, and Wang 2019), but also in other smart contexts like self-driving cars (Waytz, Heafner, and Epley 2014). On the other hand, Schroeder and Schroeder (2018) showed that listening to an application's voice did not result in higher levels of human-likeness than reading its text, and Ischen

et al. (2022) did not find a voice agent was perceived as more human-like than a text agent.

Therefore, recent studies suggested that the cues added to conversational agents or machines in general need to achieve a certain threshold of 'humanness' before they can affect anthropomorphism (Schroeder and Schroeder 2018). Schroeder and Schroeder (2018) emphasised that 'adding a voice to a machine may not be sufficient for anthropomorphism if the voice does not sound adequately human, even though in theory any voice should be more humanizing than no voice' (478). Current speech synthesis technology, released after Schroeder & Schroeder's article may make a synthetic voice so human-like it reaches this threshold. So, we test whether an agent with state-of-the-art speech (i.e. standard voice) is perceived as more human-like than a text agent. Additionally, as aligned prosody may be necessary to cross the threshold, we compare the perceived human-likeness of a speech agent with aligned prosody to a text agent (see H1a below, after the next section).

2.3.2. The relationship between perceived human-likeness and brand attitude

Subsequently, we are interested in whether the perceived human-likeness of a voice agent results in more positive attitudes towards a brand it recommends. There is little empirical work investigating the relationship between these variables in the context of conversational agents, which is surprising as the relationship between perceived human-likeness and persuasion has often been assumed by marketers. For example, Ischen et al. (2020) showed that specific human-like cues such as visual, identity, or conversational cues can increase the perceived human-likeness of a virtual text assistant; this, in turn, positively influences affective and behavioural persuasive outcomes. Also, in the context of text vs. voice agents, Ischen et al. (2022) showed that human-likeness positively influences affective and behavioural persuasive outcomes, but note that in that study a text agent was perceived as more human-like than a voice agent. Research on other machines showed that human-likeness can influence people's trust in a machine (Schroeder and Schroeder 2018). For example, Waytz, Heafner, and Epley (2014) added a name, gender, and human voice to a self-driving car simulator. Findings suggest that these cues lead to human-likeness, in turn leading to trust. But also here, there are mixed findings, as Schroeder and Schroeder (2018) showed that listening to an application's voice did not affect anthropomorphism or trust compared to reading its text.

Just like the empirical substantiation for a positive relationship between human-likeness and persuasion is scattered, there is also no one clear theory that explains *why* human-likeness is related to more positive affective responses or acceptance of a recommendation. Recent empirical research has shown that perceived competence, authenticity, or social pressure may play a role (Nguyen et al. 2023; Park et al. 2023). A relevant theoretical model is the MAIN model (Sundar 2008), which posits that different technological affordances (i.e. Modality (M), Agency (A), Interactivity (I), and Navigability (N)) can embed cues that may serve as a trigger for the operation of a heuristic or judgement rule that ultimately influences the credibility of a message (Sundar 2008). A conversational agent using a human-like voice may trigger heuristic-based judgements about the content and therefore positive evaluations. Relevant heuristics are the realism heuristic (i.e. 'reminding users of face-to-face communication, which positively influences information credibility by providing the illusion of a realistic natural interaction' Kim and Sundar 2012, 242), or the social presence heuristic (i.e. 'the idea that the user is communicating with a social entity rather than an inanimate object' Sundar 2008, 84). In line with this model, it is likely that perceived human-likeness will impact consumers' attitudes by functioning as a heuristic.

Persuasion knowledge theory (Friestad and Wright 1994) can also be used to explain why human-likeness influences persuasion (i.e. brand attitude). Its basic premise is that knowledge of the agent, tactic and topic of a persuasive attempt helps to protect people from persuasion. Perceptions of human-likeness induced by an agent's voice might lead to lower persuasion knowledge because people are more likely to infer social motives rather than commercial motives (Ischen et al. 2022).

The present study tests the following hypothesis:

> H1: Participants who are exposed to a conversational agent that communicates via voice with prosody that is aligned to their information state, vs. an agent that communicates via standard voice, or via text, (a) perceive this agent as more human-like (b) which subsequently leads to a more positive brand attitude.

2.4. The role of perceived personalisation

2.4.1. The influence of aligned prosody on perceived personalisation

Perceived personalisation can be defined as 'a customer's perception of a [recommendation agent's (RA's] personalisation (i.e. the extent to which the RA understands and represents his or her personal needs)', Komiak and Benbasat (2006, 944). More recently, it is described as whether a particular message recipient perceives a message fitting into his or her preferences (Li 2016, 27).

Aligning prosody with the information state might increase users' perception that the agent understands their needs and that the recommendation given by the agent is specifically or individually made for them. Feine et al.'s (2019) conceptualisation of cues, signals (and social reactions to signals), and the CASA paradigm helps to explain this. Aligning prosody is the cue 'any design feature of a CA salient to the user that presents a source of information' (Smith and Harper 2003). Such a cue can evolve into a social signal through the attribution of socialness towards the agent (Smith and Harper 2003 in Feine et al. 2019). The social signal, in this case, is that people interpret the cue as that the conversational agent and its communication is 'made for me'. This interpretation is likely because aligned prosody signals new and important information to the hearer. So even if an ultimate recommendation does not change (e.g. because the agent asks better questions to articulate users' needs; Komiak and Benbasat 2006), aligned prosody may give users the feeling that the agent takes them into account. Ultimately, a social signal may trigger a social reaction (i.e. 'an emotional, cognitive, or behavioural reaction of the user towards a conversational agent that is considered appropriate when directed at other human beings', Feine et al. 2019, 141). Aligned prosody, with its social signal of 'made for me', may trigger the social reaction of acceptance of the agent and its recommendations. H2a tests whether an agent with aligned prosody is perceived as being more personalised than an agent that communicated via standard voice or via text (see H2a below, after the next section). At the same time, we also explore potential differences between a text agent and a standard voice agent because the voice in itself is also a social cue (Feine et al. 2019).

2.4.2. The influence of perceived personalisation on brand attitude

The impact of perceived personalisation on consumer responses can explained by the extensive literature on personalised communication. Although potential negative consequences (e.g. privacy concerns, or reactance) are mentioned too, most marketers emphasise the opportunities that personalised messages offer (for an overview see Boerman, Kruikemeier, and Zuiderveen Borgesius 2017). Personalised advertising is generally believed to be more effective than non-personalised advertising because consumers perceive messages as more relevant, or useful (De Keyzer, Dens, and De Pelsmacker 2015), pay more attention to personalised ads (Maslowska, Smit, and Van den Putte 2016); and develop stronger behavioural intentions (Bleier and Eisenbeiss 2015). With regard to brand attitude, multiple studies have shown that persuasive messages that are perceived as more personalised are more effective in influencing brand attitude than messages that are perceived as less personalised (Segijn and Voorveld 2021). Here it is important to note that personalisation perceived by the user (whether the user perceives that a persuasive message fits him or her) is usually more predictive of consumer responses than the actual level of personalisation (i.e. personalisation features included in the communication) (De Keyzer, Dens, and De Pelsmacker 2015; Maslowska, Smit, and Van den Putte 2016).

Also in early work on recommendation agents, it was found that perceived personalisation was able to influence users' cognitive and emotional trust in the agent which then influenced the intention to adopt the agent (Komiak and Benbasat 2006), and recently, it was shown that personalisation in a conversational agent in a voice shopping context had a positive influence on attitudes towards the product (Rhee and Choi 2020). Although the perceived personalisation that users may experience when using a conversational agent that is aligning its prosody to the input given by the user is focusing on a totally different characteristic of the message, (i.e. the voice qualities), we assume the impact of perceived personalisation on user's attitude towards the recommended brand is similar, because by having the feeling that the conversational agent is understanding the user's need, they may also be more open to the recommendation made by the agent and thus evaluate the recommended brand more positively. We formulated the following hypothesis:

> H2: Participants who are exposed to a conversational agent that communicates via voice with aligned prosody vs. an agent that communicates via standard voice, or via text, (a) perceive this agent as being more personalized (b) which subsequently leads to a more positive brand attitude.

2.5. The relationship between perceived human-likeness and perceived personalisation

We will also explore the relation between perceived human-likeness and perceived personalisation to provide further theoretical grounding to the question *why* perceived human-likeness would influence brand attitude. Both perceived human-likeness and perceived personalisation can be interpreted as social signals in the framework of Feine et al. (2019). While human-likeness is a direct perception of the social cue implemented (in this case voice), personalisation as a perception of the content might further be influenced by the immediate social cue perception. This is in line with Ischen et al. (2020) showing that human-likeness (as a direct perception of a social

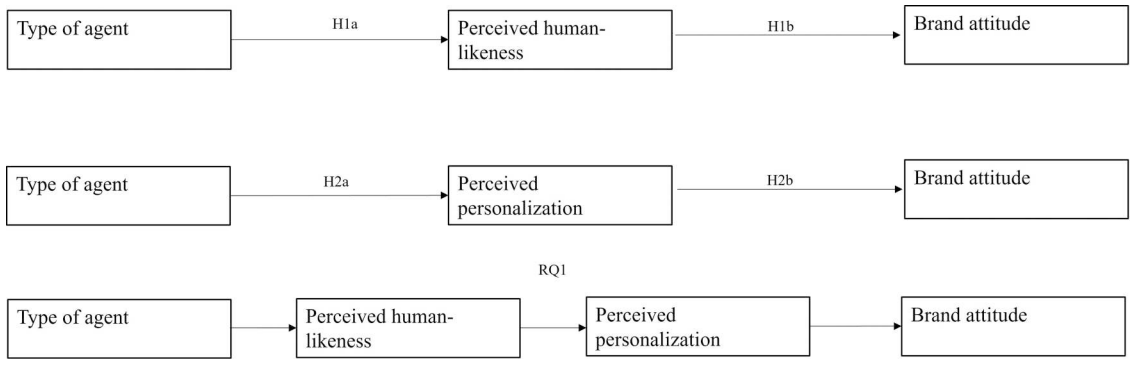

Figure 1. Overview of the hypotheses and research question.

cue) subsequently influences privacy concerns and recommendation adherence. It might thus be possible that perceived human-likeness would lead to higher levels of perceived personalisation which subsequently influences brand attitude. The concept of perceived personalisation involves users feeling that a conversational agent understands their needs, knows what they want, and takes their needs as its own preferences, and it could be that these perceptions are stronger when people perceive a conversational agent as human-like.

This line of reasoning would fit the cue-based route proposed in the MAIN model (see our argumentation leading up to H1 and H2) and earlier empirical research on the influence of human-likeness. Earlier research on textual chatbots suggests that human-likeness may trigger persuasion via the social presence heuristic (Sun, Chen, and Sundar 2024), rather than via systematic processing. In addition, a meta-analysis showed that social cues in social robots (like voice) impact social presence (Xu, Chen, and You 2023). We argue that perceived personalisation may serve as another heuristic that is triggered by human-likeness. The heuristic would constitute something like the idea that communication is 'made for me' rather than generic. Given the dearth of research examining the relationship between human-likeness and perceived personalisation, we pose the following open research question:

> RQ1: To what extent are human-likeness and perceived personalization related and serially explain brand attitude?

An overview of our hypotheses and our research question is given in Figure 1.

3. Methods

3.1. Participants and design

An online experiment with a between-subjects design was conducted in the autumn of 2021. We compared consumer responses towards three versions of a conversational agent (in this specific study a virtual assistant making a recommendation); an assistant that communicates via text, an assistant that communicates with a standard voice, and an assistant that communicates with a voice in which prosody was aligned with information structure (see details on our stimuli below). Participants were recruited through an ISO-certified research company in the United States using quotas for age, gender, and region. Sampling in the United States was deemed most appropriate given that we used Google's Text-to-Speech engine with the US English voice for our stimuli. In turn, this voice was selected as it is expected to offer the highest quality of all available options (specifically, other languages). A total of 212 US participants successfully completed the experiment; they completed the full conversation, entered a correct conversation code (to proof that they completed the full interaction), and passed the attention check in the questionnaire. Participants were nearly equally distributed among conditions (text: 73, standard voice: 71, voice with aligned prosody: 68). Participants were between 18 and 83 years old (M = 53.28; SD = 16.79), 51.4% female. In terms of education, 35.8% indicated they had a high (Master's or doctoral programme) educational level (middle: 40.7% (undergraduate programme); low: 23.5% high school or middle school).

3.2. Stimuli

Three versions (text, standard voice, and voice with aligned prosody) of a virtual assistant were designed for this study using and extending the conversational agent research toolkit (CART) developed by Araujo (2020). We used https://botonic.io/ for the front end, where we customised it to allow for audio being played. Similar to Rhee and Choi (2020), the experiment consisted of a

preformatted conversation between the participant and the assistant and the assistant's recommendation message for a specific product (a non-existing digital camera brand) brand. In line with Go and Sundar (2019), we chose a digital camera because it is a product with relatively complex product features which makes it likely that people would consult a virtual assistant to select the product that best matches their needs. In the recommendation given by the assistant, consumers' preferences for product attributes were reflected in the content of the persuasive message (so you find it important a camera has feature X, Y & Z, 'based on your answers I found a camera that best matches your needs and preferences') but the ultimate brand that was recommended was the same in every condition (the fictional VisionPro T5x). The complete interaction can be found in Appendix 1. The interaction is interactive and as such differs per user, but the same questions were asked to every participant so it crucially does not differ depending on condition.

In the text condition, all elements (questions, and answer options) of the conversation were displayed in text. Participants responded via clicking on one of the answers options or typing in their answers. In both voice conditions, the assistant generated the questions via voice, but the participants still responded by clicking or typing, so we only manipulated the response modality and not the input modality (Schroeder and Schroeder 2018), which is in line with Im et al. (2023). The main reason for not having the participants respond with spoken words is because this would very much complicate the interaction as the agent would probably need to ask participants to repeat their answers in case the agent did not 'understand' the participants' input which would impact users' evaluations (see Jiang, Jeng, and He 2013 for an extensive discussion of voice input errors and their implications). We believe this could be a serious confounding factor which would make the comparison between the response modality of voice and text less 'clean'.

Spoken versions of the dialogue were generated using Google's Text-to-Speech engine, with the male US English voice (en-US-Wavenet-0, so using the WaveNet type of synthesis, with the default audio device profile and settings). For each sound file, a version with aligned prosody was created manually using the Praat software (Boersma and Weenink 2020).

Aligning of prosody took place as follows. First, existing prosodic markings of information structure (focus and link), summation and/or continuation, and question intonation were evaluated. If they were not in accordance with the information status of the utterance in the context of the dialogue, they were removed by smoothing the pitch contour so no peak remained. Second, new pitch peaks were placed where one was required by the information status of phrases, specifically on the phrases containing new information, which were given a pitch peak aligned with the stressed syllable (*H), and contrastive information, which was given a fall-rise (L + H*) accent. Question intonation (final rise), or summation intonation (LH%) was added when appropriate. Prosody alignment also included moving accents to the correct (lexically stressed) position within a phrase. The duration was adapted if necessary, which was rarely.

An example of the adaptation of the pitch contour is given in Figure 2.

The conversation comprised 20 exchanges of one to eight sentences, divided over one to six turns (only counting the agent's turns). An agent turn is for the purposes of this description defined as everything said to the user until the user gets the opportunity to respond; a turn usually contains one question and generally includes some explanation of the options, as well as confirmation of the answer given by the user. In addition, 24 turns were explanations and error messages, produced by the agent when prompted by the user or when an answer could not be processed.

The content of each turn does not depend on the condition, the two different speech versions and the text version are based on the same text. Each participant is exposed to a subset of the turns, based on their choices in the conversation (see procedure below).

The total amount of different turns was 75, and of those 61 (81%) needed at least one correction. Focus prosody had to be added 18 times and removed 40 times, and made contrastive six times. Ground (i.e. link) prosody had to be added 17 times and removed once. Word stress was corrected 15 times, question intonation 26 times and list or summation intonation 69 times. There were two corrections that did not fall into only one category. All changes included pitch changes, and changes to duration (speech rate) were made nineteen times.

3.3. Procedure

The study was approved by the Ethical Review Board. Participants were invited by the panel company to participate in a study on their perceptions of communication with a virtual assistant. Participants were also informed they could only participate if they used a Chrome browser on a desktop or laptop (because of technical restrictions of the developed agent and our

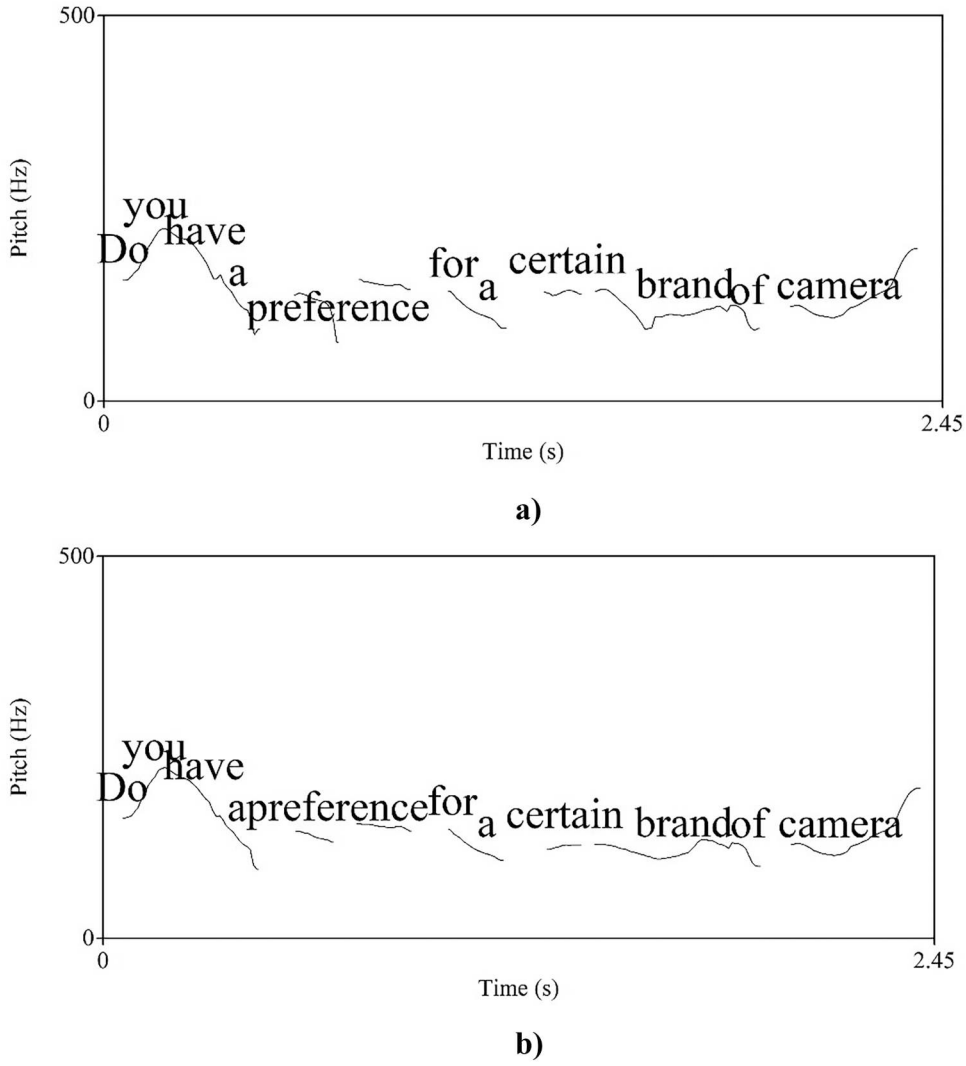

Figure 2. Example of pitch alignment. (a) Speech generated from text, without manipulation; (b) Speech generated from text, with pitch alignment.

Note: The pitch is shown over time. Image (a) shows a focus accent on 'preference', shown as a pitch peak. However, as preferences are already a topic of conversation, the phrase does not correspond to new information. In the image (b), the pitch contour on 'preference' has been removed. In addition, a pitch peak has been added on the syllable 'brand', as the phrase 'a certain brand of camera' is new information (focus); it is new to the listener that the question is about camera brands. Note that focus accents near the end of an utterance tend to be lower.

need to keep factors like layout as constant as possible). They were informed to be compensated by the panel company in a regular manner if they completed the full conversation with the assistant and completed the questionnaire. After giving consent, participants were asked to test whether the speakers of their computer were working. Next, they were asked to imagine that they intend to buy a digital camera and to interact with a virtual assistant that can help them to find a digital camera that best matches their needs. They were again informed that it is important to fully complete the dialogue with the assistant and that, at the end of the conversation, they will receive a test code needed to proceed. Participants could then start the conversation by typing 'hi'. The chatbot asked input from the user and replied with appropriate information and a new question. After completing the full conversation (duration was about six minutes on average for the voice conditions and four and a half minute for the text condition), and typing in the test code, they were asked whether the conversation took place without any problems, Next, they filled out the questionnaire. Participation took about 20 min in total.

3.4. Measures

3.4.1. Perceived human-likeness

Perceived human-likeness was assessed with self-reported measurements of mindless and mindful anthropomorphism (Kim and Sundar 2012). Mindless

anthropomorphism was measured by asking participants to rate how well each adjective described the assistant using four 10-point scales (1 describes very poorly to 10 describes very well): 'likeable', 'sociable', 'friendly', and 'personal'. This measure was originally developed by Kim and Sundar (2012) (Cronbach's α = 0.94; M = 7.97; SD = 2.01).

Mindful anthropomorphism was measured by asking participants whether they perceived the assistant as being 'human-like/ machinelike', 'natural/unnatural', 'lifelike/artificial', and 'unconscious/ conscious', on 7-point semantic differential scales Bartneck et al. (2009) (Cronbach's α = 0.94; M = 4.75; SD = 1.72).

3.4.2. Perceived personalisation

Perceived personalisation was measured with five items based on the items used by Li (2016) and Komiak and Benbasat (2006). Participants were asked to what extent they agree with the following statements on a 7-point scale ranging from 1 (totally disagree) to 7 (totally agree): 'The recommendation seems to be designed specifically for me'; 'The assistant targets me as a unique individual'; 'The assistant understands my needs'; 'This virtual assistant knows what I want'; 'This assistant takes my needs as its own preferences' (Cronbach's α = 0.93; M = 5.14; SD = 1.33).

3.4.3. Brand attitude

The brand attitude was measured with four items on a 7-point semantic differential scale (Chang and Thorson 2004). Participants evaluated the brand by answering the following: 'The virtual assistant recommended a camera from the brand VisionPro. How do you evaluate the brand VisionPro?' Answer options ranged from 'bad-good'/ 'dislike-like'/ 'unfavorable-favorable'/ 'not interesting- interesting' (Cronbach's α = 0.98; M = 5.09; SD = 1.33).

3.4.4. Familiarity with the assistant

Familiarity with the assistant was measured with two items adapted from Steenkamp, Batra, and Alden (2003). Participants were asked to what extent they agreed with the statements using a 7-point scale ranging from 1 (totally disagree) to 7 (totally agree): 'Interacting with a virtual assistant like this one is very familiar to me', and 'I am very knowledgeable about interacting with a virtual assistant like this one'. (Pearson's r = .89, p = .00; M = 4.13; SD = 1.81).

4. Results

All analyses were conducted using SPSS. After performing randomisation checks and checking for potential control variables to be included, we used ANCOVA's to test for the impact of the type of agent on perceived human-likeness and perceived personalisation (H1a and H2a), followed by a formal mediation test using Hayes (2017) PROCESS macro (model 4). We ended with a single serial mediation model (Hayes, model 6) to answer RQ1.

4.1. Randomisation checks and control variables

To check whether the random assignment to the different groups was successful, we conducted a chi square test for gender and ethnic background, and ANOVA's for age, education, camera ownership, brand familiarity, and familiarity with the type of assistant. There were no significant differences for gender (X^2 = 0.19, p = .911), age (F (2,209) = 2.09, p = .126) ethnic background (X^2 = 9.73, p = .464), education (F (2,201) = 84, p = .434), camera ownership (X^2 = 1.02, p = .600), nor brand familiarity (F(2,209) = 1.64, p = .197). Only familiarity with the type of assistant differed significantly between conditions (F(2,209) = 2.46, p = .044, tested one-sided given strong expectations about the direction; people being more familiar with text assistants than voice assistants). Participants were most familiar with text assistants (M = 4.45, SD = 1.79) and less familiar with voice assistants (assistant with standard voice: M = 3.79, SD = 1.76; assistant with correct prosody: M = 4.14, SD = 1.84). Since this variable was also significantly correlated with all outcome variables (mindless anthropomorphism, r = .30, p = .000, perceived personalisation r = .36, p = .000, and brand attitude r = .54, p = .000) it was included as covariate in the subsequent analyses.

4.2. Direct effects on brand attitude

Whereas we did not hypothesise a direct effect of assistant type on brand attitude, we tested whether such a direct effect exist. Results of an ANCOVA with the type of assistant as an independent variable, brand attitude as a dependent variable and familiarity with the type of assistant as a covariate showed no significant direct effect (F(2,208) = .45, p = .639; aligned prosody: M = 5.15; SD = 1.28; standard voice: M = 5.00; SD = 1.30, text: M = 5.12; SD = 1.42).

4.3. The role of perceived human-likeness

To test whether the different assistants were perceived differently in terms of their perceived mindless anthropomorphism (H1a), an ANCOVA with the type of assistant as an independent variable, mindless

anthropomorphism as a dependent variable and familiarity with the type of assistant as a covariate was conducted. Results showed a significant difference between the three conditions. ($F(2,208) = 3.77$, $p = .025$). The voice assistant with aligned prosody scored highest with regards to mindless anthropomorphism $M = 8.39$; SD = 1.60, followed by the standard voice assistant ($M = 7.92$; SD = 2.09), and the text assistant ($M = 7.62$; SD = 2.23). Simple contrast tests with the text condition as a reference category showed a significant difference between the text assistant and the voice assistant with aligned prosody ($p = .007$), but not between the text assistant and the standard voice assistant ($p = .102$). An additional simple contrast between the standard voice and aligned prosody conditions showed no significant difference ($p = .282$). A similar ANCOVA with mindful anthropomorphism as a dependent variable did not show a significant effect ($F(2,208) = .18$, $p = .834$). Therefore, we did not proceed with mediation analyses for this variable.

4.3.1. The mediating role of human-likeness

To formally test the mediating role of mindless anthropomorphism (H2b), we used the Hayes (2017) PROCESS macro (v 4.54, model 4; Hayes) with the type of assistant as an independent variable, mindless anthropomorphism as mediator, brand attitude as dependent variable and familiarity as a control variable. Because the type of assistant is a multicategorical variable with three values (text, voice and voice with aligned prosody), we used indicator coding with the text as the reference group based on the results of the ANCOVA.[1] In line with the ANCOVA's results, there was a significant effect of the type of assistant on mindless anthropomorphism when comparing aligned prosody with text assistant (effect: 0.88, SE = .32, $p = .007$), but not when comparing standard voice to text (effect = .53, SE = .32, $p = .10$). Crucially, there is a significant mediation effect when comparing the voice assistant with aligned prosody to the text assistant (effect = 0.22, SE = .09, bias corrected 95% confidence interval (CI) [.0675, .4014]), but not when comparing the standard voice version with the text assistant (effect = 0.13, SE = .09, CI [−.0310, .3246]). So, an assistant that is enhanced with aligned prosody is perceived as more human-like compared to a text assistant, which subsequently leads to a more positive brand attitude.

4.4. Perceived personalisation

To test whether the assistant types were perceived differently in terms of personalisation (H2a), an ANCOVA with type of assistant as independent variable, mindless anthropomorphism as a dependent variable and familiarity with the type of assistant as control variable was conducted. Results showed that the assistant types did not differ with regards to perceived personalisation ($F(2,208) = 1.07$ $p = .346$; aligned prosody: $M = 5.30$; SD = 1.27; standard voice: $M = 5.07$ SD = 1.29; text: $M = 5.07$; SD = 1.44). This means there is no direct effect of the type of assistant on perceived personalisation.

4.4.1. The mediating role of perceived personalisation

Since there was no significant difference between the three conditions with regard to perceived personalisation, we refrain from testing for mediation.

4.5. The role of human-likeness and perceived personalisation: serial mediation

To test the significance of the mediators in relation to each other (RQ1), we entered both mediators into a single serial mediation model (Hayes, model 6, again with text as the reference group). All coefficients for the full model are displayed in Table 1. When comparing the voice assistant with aligned prosody to the text assistant, the serial mediation effect via mindless anthropomorphism and perceived personalisation was significant (effect = 0.14, SE = .07, CI [.0335, .2938]). The serial mediation for the comparison between the standard voice and the text assistant was not significant (effect = 0.08, SE = .06, CI [−.0133, .2155]), which is not surprising given the non-significant difference between the two conditions with regard to human-likeness reported above. The separate indirect effects via anthropomorphism and perceived personalisation respectively were no longer significant (anthropomorphism: standard voice vs text: effect = 0.05, SE = .05, CI [−.0160, .1657]; aligned prosody vs text: effect = 0.08, SE = .06,

Table 1. Path coefficients sequential mediation explaining brand attitude.

	Standard voice vs. text	Aligned prosody voice vs. text
Type of agent → perceived human-likeness	.53 (.32)	.88** (.32)
Type of agent → perceived personalisation	−.06 (.16)	−.07 (.16)
Perceived human-likeness → perceived personalisation	.44*** (.03)	.44*** (.03)
Perceived human-likeness → brand attitude	. 09† (.05)	. 09† (.05)
Perceived personalisation → brand attitude	.36*** (.07)	.36*** (.07)
Type of agent → brand attitude (c′)	.04 (.17)	−.03 (.17)

Note: †$p < .10$; *$p < .05$; **$p < .01$; ***$p < .001$; controlled for familiarity with medium. Hayes, model 6.

CI [−.0152, .2068]; perceived personalisation: standard voice vs text: effect = −0. 020, SE = .05, CI [−.1256, .0901]; aligned prosody vs text: effect = −0. 03, SE = .06, CI [−.1362, .0964]). This means that a voice assistant with aligned prosody leads to higher levels of mindless anthropomorphism, which then leads to higher levels of perceived personalisation, which subsequently results in more positive brand attitude.

5. Conclusion and discussion

This study showed that perceived human-likeness and perceived personalisation together explain why voice assistants with aligned prosody are more persuasive than text assistants. So, aligning the prosody of a voice agent to the information state of the user makes it more human-like, and therefore makes the assistant seem to understand the user better, which makes its recommendations seem more individually tailored. In turn, this perceived personalisation or tailoring makes the attitude towards the recommended brand more positive. The absence of evidence that these mechanisms also cause differences between standard voice agents and text agents tentatively suggests that the quality threshold needed for these mechanisms to apply may be that an agent uses prosody to show it is attending to the users' information state.

5.1. Theoretical implications

The present study makes at least four contributions to literature and theory. It is one of the first that provided evidence for the idea proposed by Schroeder and Schroeder (2018) that a voice needs to meet a quality threshold to be perceived as human-like. More specifically, it shows one possible implementation of the required quality, namely the social signal 'made for me', provided by the cue-aligned prosody, which belongs to the 'voice qualities' of Feine et al. (2019). Aligned prosody in a voice agent can affect perceived human-likeness compared to a text agent, whereas there is no evidence that a conversational agent with a standard voice has the same effect. This difference in evidence may not reflect a true difference between the two voice conditions, but it may explain why earlier studies directly comparing conversational text and voice assistants sometimes did (Cho, Molina, and Wang 2019) and sometimes did not (e.g. Ischen et al. 2022) find differences in terms of human-likeness. There were already indications (Qiu and Benbasat 2009) or authors suggesting (Ischen et al. 2022) that voices sounding too unnatural might not be able to influence perceptions of human-likeness, although the CASA paradigm suggests that voice itself is already a social cue that should be able to trigger social reactions.

Second, a conversational assistant using aligned prosody not only impacts perceived human-likeness but also influences participants' attitudes towards a recommended brand. Whereas the downstream consequences of perceived human-likeness have already been shown multiple times, empirical evidence in the context of persuasion and more specifically brand attitude was still limited. Our study shows that the cue 'aligned prosody' did not only function as a social signal but also that it leads to a social reaction which is in agreement with the recommendation made as reflected in participants' brand attitude towards the recommended brand.

Third, this study contributed to the further theoretical substantiation of the link between perceived human-likeness and persuasive outcomes such as brand attitude. Although this relationship had already been tested a few times, the theoretical explanation for a link between the two constructs was limited. By not only empirically showing a link between perceived human-likeness and brand attitude but also unravelling the underlying mechanism via perceived personalisation, our study contributes to the theories about perceived human-likeness in the context of persuasion. By doing so, the study opens possibilities to further explore perceived personalisation in the CASA paradigm and empirical research on conversational agents.

Finally, an additional theoretical finding, not from the experiment but rather from stimulus creation, is that current state-of-the-art text-to-speech quite often contains errors in prosodic marking of information structure, and hence is often not pragmatically felicitous. Although information structure is crucially determined by the context, for many utterances the default option may be right. The amount of corrections necessary suggests that either the default option is not generally right, or that the text-to-speech software is not tapping into the semantics of the sentence-to-become-utterance to predict its pragmatic status. As the technology behind Wavenet, deep learning, is notoriously hard to interpret, this question is hard to answer directly.

5.2. Practical implications

For marketers considering implementing a conversational agent for persuasion purposes, for example, to recommend brands or products, our results can inform the decision whether to implement either a text agent or a voice agent. Voice agents do not necessarily lead to a more positive brand attitude than text agents. If marketers decide to implement a voice agent, they should consider that the quality of the synthetic voice employed by

the agent should pass a certain threshold, and our study shows that one way to achieve this goal is to use an agent with aligned prosody. Our research also implies that it is important to empirically test users' responses towards developed agents. In their search for implementing human-like features, it may be easy for developers to forget that user expectations play an important role in shaping user responses. Whereas a synthetic voice as such is usually considered as a human-like cue, our results showed only for a voice with aligned prosody that it is perceived as more human-like than a text agent. Users' expectations about human-like conversational agents may change and it may be harder to meet them for a voice agent than for a text agent.

5.3. Limitations and future research

Our study obviously has some limitations and yields some suggestions for future research. As explained before, we only manipulated the response modality and not the input modality. Future researchers might also want to take input modality into account. Whereas we expect that perceived human-likeness and perceived personalisation also play a role when people talk (vs. type or write) to an assistant, there is some research suggesting that there will be unique dynamics for input (e.g. Schroeder and Schroeder 2018). Next to the impact of possible voice input errors (see Jiang, Jeng, and He 2013) other underlying mechanisms may play a role too, for example, because emotions are more often expressed in speaking than writing (Berger, Rocklage, and Packard 2022). Investigating response modality and input modality at the same time may further complicate the design and execution of experiments, but will also make them more externally valid.

Furthermore, our study used conversational agents specifically developed for the purpose of the study. While this approach has the advantage of being able to manipulate prosodic elements in a fine-grained manner and furthermore ensure that all participants were exposed to the same interaction in a controlled environment (for an overview of methodological considerations see Greussing et al. 2022), it also comes with certain limitations regarding ecological validity. Recent developments in generative AI provide advanced ways of personalising the language as well as the content of conversational agents and furthermore allows for more open-ended conversations.

In the present study, the assistant asked the same questions to all participants, and ultimately all participants were recommended the same brand, only pretending that it took into account their input. In future studies, it may be interesting to investigate a potential moderation effect of aligned prosody and actual personalisation, to see if even stronger persuasive effects occur if both the social cue that signals personalisation (aligned prosody) and the actual personalisation of interaction is implemented at the same time.

Furthermore, there are many efforts underway to improve prosody in speech generation. For example, Hodari et al. (2021), Mohan et al. (2021), Stephenson et al. (2022), Torresquintero et al. (2021) and Zou et al. (2021) show promising results for prosody improvement, possibly enough that it will be, or feel, aligned to the user's information state, and/or allow developers of voice agents to tweak the prosody of the synthetic speech used by the agent. The paradigm used in the current study may serve to test the effects of such improvements, replacing the manual alignments by semi-automatic ones, and therefore may serve as a stepping stone for empirical research in this promising, interdisciplinary area.

6. Conclusion

In conclusion, this study shows that a voice agent with pragmatically aligned prosody leads to users having a more positive brand attitude towards the recommended brand than a text agent, and that perceived human-likeness and perceived personalisation together provide an underlying mechanism to explain these differences. Hence, aligned prosody can make synthetic speech meet a certain quality threshold to be perceived as more human-like, and subsequently more personalised, and more persuasive.

Note

1. As a mediation analysis with 3 group requires to set a reference category (in line with having 2 dummies when conducting a regression analysis with a categorical IV with 3 groups ($n-1$)), we were not able to test the mediation effect for the aligned prosody vs. the standard voice agent condition. However, as the ANCOVA showed no significant different in mindless antropomorphism between the two conditions, a mediation analysis would never show a significant mediation via that variable.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This study was funded by a grant from the Interdisciplinary Research Priority Area 'Humane AI' at the University of Amsterdam.

Data available statement

Data are available from the first author upon request.

References

Araujo, T. 2020. "Conversational Agent Research Toolkit: An Alternative for Creating and Managing Chatbots for Experimental Research." *Computational Communication Research* 2 (1): 35–51. https://doi.org/10.5117/CCR2020.1.002.ARAU.

Bartneck, C., D. Kulić, E. Croft, and S. Zoghbi. 2009. "Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots." *International Journal of Social Robotics* 1 (1): 71–81. https://doi.org/10.1007/s12369-008-0001-3.

Berger, J., M. D. Rocklage, and G. Packard. 2022. "Expression Modalities: How Speaking versus Writing Shapes Word of Mouth." *Journal of Consumer Research* 49 (3): 389–408. https://doi.org/10.1093/jcr/ucab076.

Bleier, A., and M. Eisenbeiss. 2015. "The Importance of Trust for Personalized Online Advertising." *Journal of Retailing* 91 (3): 390–409. https://doi.org/10.1016/j.jretai.2015.04.001.

Boerman, S. C., S. Kruikemeier, and F. J. Zuiderveen Borgesius. 2017. "Online Behavioral Advertising: A Literature Review and Research Agenda." *Journal of Advertising* 46 (3): 363–376. https://doi.org/10.1080/00913367.2017.1339368.

Boersma, P. and D. Weenink. 2020. PRAAT: Doing Phonetics by Computer [Computer program]. Version 6.1.03. Accessed September 1, 2020. http://www.praat.org/.

Brown, G., and G. Yule. 1983. *Discourse Analysis (Ser. Cambridge Textbooks in Linguistics)*. Cambridge: Cambridge University Press.

Chandra, S., A. Shirish, and S. C. Srivastava. 2022. "To Be or Not to Be … Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents." *Journal of Management Information Systems* 39 (4): 969–1005. https://doi.org/10.1080/07421222.2022.2127441.

Chang, Y., and E. Thorson. 2004. "Television and Web Advertising Synergies." *Journal of Advertising* 33 (2): 75–84. https://doi.org/10.1080/00913367.2004.10639161.

Chérif, E., and J. F. Lemoine. 2019. "Anthropomorphic Virtual Assistants and the Reactions of Internet Users: An Experiment on the Assistant's Voice." *Recherche et Applications en Marketing (English Edition)* 34 (1): 28–47. https://doi.org/10.1177/2051570719829432.

Cho, E., M. D. Molina, and J. Wang. 2019. "The Effects of Modality, Device, and Task Differences on Perceived Human Likeness of Voice-Activated Virtual Assistants." *Cyberpsychology, Behavior, and Social Networking* 22 (8): 515–520. https://doi.org/10.1089/cyber.2018.0571.

Clark, L., P. Doyle, D. Garaialde, E. Gilmartin, S. Schlögl, J. Edlund, M. Aylett, et al. 2019. "The State of Speech in HCI: Trends, Themes and Challenges." *Interacting with computers* 31 (4): 349–371. https://doi.org/10.1093/iwc/iwz016.

Clark, H. H., and C. R. Marshall. 1981. "Definite Reference and Mutual Knowledge." In *Elements of Discourse Understanding*, edited by A. K. Joshi, B. L. Webber, and I. A. Sag, 10–63. Cambridge UK: Cambridge University Press.

De Keyzer, F., N. Dens, and P. De Pelsmacker. 2015. "Is This for Me? How Consumers Respond to Personalized Advertising on Social Network Sites." *Journal of Interactive Advertising* 15 (2): 124–134. https://doi.org/10.1080/15252019.2015.1082450.

E Marketer. 2022. Conversational AI Use Expands, Presents Opportunities - Insider Intelligence Trends, Forecasts & Statistics. emarketer.com.

Epley, N., A. Waytz, S. Akalis, and J. T. Cacioppo. 2008. "When We Need a Human: Motivational Determinants of Anthropomorphism." *Social Cognition* 26 (2): 143–155. https://doi.org/10.1521/soco.2008.26.2.143.

Feine, J., U. Gnewuch, S. Morana, and A. Maedche. 2019. "A Taxonomy of Social Cues for Conversational Agents." *International Journal of Human-Computer Studies* 132:138–161. https://doi.org/10.1016/j.ijhcs.2019.07.009.

Friestad, M., and P. Wright. 1994. "The Persuasion Knowledge Model: How People Cope with Persuasion Attempts." *Journal of Consumer Research* 21 (1): 1–31. https://doi.org/10.1086/209380.

Go, E., and S. S. Sundar. 2019. "Humanizing Chatbots: The Effects of Visual, Identity and Conversational Cues on Humanness Perceptions." *Computers in Human Behavior* 97:304–316. https://doi.org/10.1016/j.chb.2019.01.020.

Gong, L., and J. Lai. 2001. "Shall We Mix Synthetic Speech and Human Speech? Impact on Users' Performance, Perception, and Attitude." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 158–165. https://doi.org/10.1145/365024.365090.

Greussing, E., F. Gaiser, S. H. Klein, C. Straßmann, C. Ischen, S. Eimler, K. Freemann, et al. 2022. "Researching Interactions between Humans and Machines: Methodological Challenges." *Publizistik* 67 (4): 531–554. https://doi.org/10.1007/s11616-022-00759-3.

Grudin, J., and R. Jacques. 2019. "Chatbots, Humbots, and the Quest for Artificial General Intelligence." In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–11.

Hayes, A. F. 2017. *Introduction to Mediation, Moderation, and Conditional Process Analysis: A Regression-Based Approach*. New York: Guilford Publications.

Hodari, Z., A. Moinet, S. Karlapati, J. Lorenzo-Trueba, T. Merritt, A. Joly, A. Abbas, P. Karanasou, and T. Drugman. 2021. "CAMP: A Two-Stage Approach to Modelling Prosody in Context." In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6578–6582. toronto, ON, Canada: IEEE. https://doi.org/10.1109/ICASSP39728.2021.9414413.

Im, H., B. Sung, G. Lee, and K. Q. X. Kok. 2023. "Let Voice Assistants Sound Like a Machine: Voice and Task Type Effects on Perceived Fluency, Competence, and Consumer Attitude." *Computers in Human Behavior* 145:107791. https://doi.org/10.1016/j.chb.2023.107791.

Ischen, C., T. B. Araujo, H. A. M. Voorveld, G. van Noort, and E. G. Smit. 2020. "Privacy Concerns in Chatbot Interactions." In *Chatbot Research and Design: Third International Workshop, CONVERSATIONS 2019, Amsterdam, The Netherlands, November 19–20, 2019: Revised Selected Papers*, edited by A. Følstad, T. Araujo, S.

Papadopoulos, E. L.-C. Law, O.-C. Granmo, E. Luger, and P. B. Brandtzaeg, 34–48. (Lecture Notes in Computer Science; Vol. 11970). Springer.

Ischen, C., T. B. Araujo, H. A. M. Voorveld, G. Van Noort, and E. G. Smit. 2022. "Is Voice Really Persuasive? The Influence of Modality in Virtual Assistant Interactions and Two Alternative Explanations." *Internet Research* 32 (7): 402–425. https://doi.org/10.1108/INTR-03-2022-0160.

Jiang, J., W. Jeng, and D. He. 2013. "How Do Users Respond to Voice Input Errors? Lexical and Phonetic Query Reformulation in Voice Search." In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 143–152. New York, NY, USA: Association for Computing Machinery.

Kim, Y., and S. S. Sundar. 2012. "Anthropomorphism of Computers: Is It Mindful or Mindless?" *Computers in Human Behavior* 28 (1): 241–250. https://doi.org/10.1016/j.chb.2011.09.006.

Komiak, S. Y., and I. Benbasat. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents." *MIS Quarterly* 30:941–960. https://doi.org/10.2307/25148760.

Krifka, M., C. Féry, and G. Fanselow. 2007. *Interdisciplinary Studies on Information Structure 6: The Notions of Information Structure*. Potsdam: Universität.

Lee, K. M., W. Peng, S. A. Jin, and C. Yan. 2006. "Can Robots Manifest Personality?: An Empirical Test of Personality Recognition, Social Responses, and Social Presence in Human–Robot Interaction." *Journal of Communication* 56 (4): 754–772. https://doi.org/10.1111/j.1460-2466.2006.00318.x.

Letheren, K., J. Jetten, J. Roberts, and J. Donovan. 2021. "Robots Should Be Seen and Not heard … sometimes: Anthropomorphism and AI Service Robot Interactions." *Psychology & Marketing* 38 (12): 2393–2406. https://doi.org/10.1002/mar.21575.

Li, C. 2016. "When Does Web-Based Personalization Really Work? The Distinction between Actual Personalization and Perceived Personalization." *Computers in Human Behavior* 54:25–33. https://doi.org/10.1016/j.chb.2015.07.049.

Mari, A., A. Mandelli, and R. Algesheimer. 2020. "The Evolution of Marketing in the Context of Voice Commerce: A Managerial Perspective." In *HCI in Business, Government and Organizations: 7th International Conference, HCIBGO 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings 22*, edited by F. H. Nah and K. Siau, 405–425. Copenhagen, Denmark: Springer International Publishing.

Maslowska, E., E. G. Smit, and B. Van den Putte. 2016. "It Is all in the Name: A Study of Consumers' Responses to Personalized Communication." *Journal of Interactive Advertising* 16 (1): 74–85. https://doi.org/10.1080/15252019.2016.1161568.

McGuire, W. J. 1989. "Theoretical Foundations of Campaigns." In *Public Communication Campaigns*, 2nd ed, edited by R. E. Rice and C. K. Atkin, 43–54. Los Angeles: Sage.

Mohan, D. S. R., V. Hu, T. H. Teh, A. Torresquintero, C. G. R. Wallis, M. Staib, L. Foglianti, J. Gao, and S. King. 2021. "Ctrl-P: Temporal Control of Prosodic Variation for Speech Synthesis." In *Proceedings Interspeech 2021*, 3875–3879. https://doi.org/10.21437/Interspeech.2021-1583.

Moussawi, S., and R. Benbunan-Fich. 2021. "The Effect of Voice and Humour on Users' Perceptions of Personal Intelligent Agents." *Behaviour & Information Technology* 40 (15): 1603–1626. https://doi.org/10.1080/0144929X.2020.1772368.

Nass, C., and Y. Moon. 2000. "Machines and Mindlessness: Social Responses to Computers." *Journal of Social Issues* 56 (1): 81–103. https://doi.org/10.1111/0022-4537.00153.

Nass, C., J. Steuer, and E. R. Tauber. 1994. "Computers Are Social Actors." In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems April 24–28, 1994*, 72–28. Boston, USA.

Nguyen, M., L. E. Casper Ferm, S. Quach, N. Pontes, and P. Thaichon. 2023. "Chatbots in Frontline Services and Customer Experience: An Anthropomorphism Perspective." *Psychology & Marketing* 40 (11): 2201–2225. https://doi.org/10.1002/mar.21882.

Park, G., M. C. Yim, J. Chung, and S. Lee. 2023. "Effect of AI Chatbot Empathy and Identity Disclosure on Willingness to Donate: The Mediation of Humanness and Social Presence." *Behaviour & Information Technology* 42 (12): 1998–2010. https://doi.org/10.1080/0144929X.2022.2105746.

Perloff, R. M. 2021. *The Dynamics of Persuasion: Communication and Attitudes in the 21st Century*. 7th ed. New York: Routledge.

Petty, R. E., L. R. Fabrigar, and D. T. Wegener. 2003. "Emotional Factors in Attitudes and Persuasion." In *Handbook of Affective Sciences*, edited by R. J. Davidson, K. Scherer, and H. H. Goldsmith, 752–772. Oxford: Oxford University Press.

Pickering, M. J., and S. Garrod. 2004. "Toward a Mechanistic Psychology of Dialogue." *Behavioral and brain sciences* 27 (2): 169–190.

Qiu, L., and I. Benbasat. 2009. "Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems." *Journal of management information systems* 25 (4): 145–182. https://doi.org/10.2753/MIS0742-1222250405.

Rapp, A., L. Curti, and A. Boldi. 2021. "The Human Side of Human-Chatbot Interaction: A Systematic Literature Review of Ten Years of Research on Text-Based Chatbots." *International Journal of Human-Computer Studies* 151:102630. https://doi.org/10.1016/j.ijhcs.2021.102630.

Reeves, B., and C. Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge UK: Cambridge University Press.

Reicherts, L., Y. Rogers, L. Capra, E. Wood, T. D. Duong, and N. Sebire. 2022. "It's Good to Talk: A Comparison of Using Voice versus Screen-Based Interactions for Agent-Assisted Tasks." *ACM Transactions on Computer-Human Interaction* 29 (3): 1–41. https://doi.org/10.1145/3484221.

Rhee, C. E., and J. Choi. 2020. "Effects of Personalization and Social Role in Voice Shopping: An Experimental Study on Product Recommendation by a Conversational Voice Agent." *Computers in Human Behavior* 109:106359. https://doi.org/10.1016/j.chb.2020.106359.

Rodero, E. 2017. "Effectiveness, Attention, and Recall of Human and Artificial Voices in an Advertising Story. Prosody Influence and Functions of Voices." *Computers in Human Behavior* 77:336–346. https://doi.org/10.1016/j.chb.2017.08.044.

Schroeder, J., and N. Epley. 2016. "Mistaking Minds and Machines: How Speech Affects Dehumanization and Anthropomorphism." *Journal of Experimental Psychology: General* 145:1427–1437. https://doi.org/10.1037/xge0000214.

Schroeder, J., and M. Schroeder. 2018. "Trusting in Machines: How Mode of Interaction Affects Willingness to Share Personal Information with Machines." In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 472–480. https://pdfs.semanticscholar.org/38ab/c0e1686da72ac4b2947add463682d1aba107.pdf.

Segijn, C. M., and H. A. Voorveld. 2021. "A First Step in Unraveling Synced Advertising Effectiveness." *International Journal of Advertising* 40 (1): 124–143. https://doi.org/10.1080/02650487.2020.1778279.

Smith, J. M., and D. Harper. 2003. *Animal Signals*. Oxford, UK: Oxford University Press.

Steenkamp, J. B., R. Batra, and D. L. Alden. 2003. "How Perceived Brand Globalness Creates Brand Value." *Journal of International Business Studies* 34 (1): 53–65. https://doi.org/10.1057/palgrave.jibs.8400002.

Stephenson, B., L. Besacier, L. Girin, and T. Hueber. 2022. "BERT, Can HE Predict Contrastive Focus? Predicting and Controlling Prominence in Neural TTS Using a Language Model." In *Proceedings Interspeech 2022*, 3383–3387. https://doi.org/10.21437/Interspeech.2022-10116.

Sun, Y., J. Chen, and S. S. Sundar. 2024. "Chatbot ads with a Human Touch: A Test of Anthropomorphism, Interactivity, and Narrativity." *Journal of Business Research* 172:114403. https://doi.org/10.1016/j.jbusres.2023.114403.

Sundar, S. S. 2008. "The MAIN Model: A Heuristic Approach to Understanding Technology Effects on Credibility." In *Digital Media, Youth, and Credibility*, edited by Miriam J. Metzger and Andrew J. Flanagin, 73–100. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning. Cambridge, MA: The MIT Press. https://doi.org/10.1162/dmal.9780262562324.073.

Suzuki, N., and Y. Katagiri. 2007. "Prosodic Alignment in Human–Computer Interaction." *Connection Science* 19 (2): 131–141. https://doi.org/10.1080/09540090701369125

Torresquintero, A., T. H. Teh, C. G. R. Wallis, M. Staib, D. S. R. Mohan, V. Hu, L. Foglianti, J. Gao, and S. King. 2021. "ADEPT: A Dataset for Evaluating Prosody Transfer." In *Proceedings Interspeech 2021*, 3880–3884. https://doi.org/10.1080/09540090701369125.

Vallduví, E., and E. Engdahl. 1996. "The Linguistic Realization of Information Packaging." *Linguistics* 34 (3): 459–520. https://doi.org/10.1515/ling.1996.34.3.459.

van den Oord, A., Y. Li, I. Babuschkin, K. Simonyan, O. Vinyals, K. Kavukcuoglu, G. Driessche, et al. 2018. "Parallel WaveNet: Fast High-Fidelity Speech Synthesis." *Proceedings of the 35th International Conference on Machine Learning* 80:3918–3926.

Van Pinxteren, M. M., M. Pluymaekers, and J. G. Lemmink. 2020. "Human-Like Communication in Conversational Agents: A Literature Review and Research Agenda." *Journal of Service Management* 31 (2): 203–225. https://doi.org/10.1108/JOSM-06-2019-0175.

Vaughan-Johnston, T. I., J. J. Guyer, L. R. Fabrigar, and C. Shen. 2021. "The Role of Vocal Affect in Persuasion: The CIVA Model." *Journal of Nonverbal Behavior* 45 (4): 455–477. https://doi.org/10.1007/s10919-021-00373-3.

Voorveld, H. A. M., and T. Araujo. 2020. "How Social Cues in Virtual Assistants Influence Concerns and Persuasion: The Role of Voice and a Human Name." *Cyberpsychology, Behavior, and Social Networking* 23 (10): 689–696. https://doi.org/10.1089/cyber.2019.0205.

Waytz, A., J. Heafner, and N. Epley. 2014. "The Mind in the Machine: Anthropomorphism Increases Trust in an Autonomous Vehicle." *Journal of Experimental Social Psychology* 52:113–117. https://doi.org/10.1016/j.jesp.2014.01.005.

Xu, K. 2019. "First Encounter with Robot Alpha: How Individual Differences Interact with Vocal and Kinetic Cues in Users' Social Responses." *New Media & Society* 21 (11-12): 2522–2547. https://doi.org/10.1177/1461444819851479.

Xu, K., M. Chen, and L. You. 2023. "The Hitchhiker's Guide to a Credible and Socially Present Robot: Two Meta-Analyses of the Power of Social Cues in Human–Robot Interaction." *International Journal of Social Robotics* 15 (2): 269–295. https://doi.org/10.1007/s12369-022-00961-3.

Zou, Y., S. Liu, X. Yin, H. Lin, C. Wang, H. Zhang, and Z. Ma. 2021. "Fine-Grained Prosody Modeling in Neural Speech Synthesis Using ToBI Representation." In *Proceedings Interspeech 2021*, 3146–3150. https://doi.org/10.21437/Interspeech.2021-883.