Voice in Human–Agent Interaction: A Survey

KATIE SEABORN, Tokyo Institute of Technology and RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
NORIHISA P. MIYAKE, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan
PETER PENNEFATHER, gDial Inc., Toronto, Ontario, Canada
MIHOKO OTAKE-MATSUURA, RIKEN Center for Advanced Intelligence Project (AIP), Tokyo, Japan

Social robots, conversational agents, voice assistants, and other embodied AI are increasingly a feature of everyday life. What connects these various types of intelligent agents is their ability to interact with people through voice. Voice is becoming an essential modality of embodiment, communication, and interaction between computer-based agents and end-users. This survey presents a meta-synthesis on agent voice in the design and experience of agents from a human-centered perspective: voice-based human–agent interaction (vHAI). Findings emphasize the social role of voice in HAI as well as circumscribe a relationship between agent voice and body, corresponding to human models of social psychology and cognition. Additionally, changes in perceptions of and reactions to agent voice over time reveals a generational shift coinciding with the commercial proliferation of mobile voice assistants. The main contributions of this work are a vHAI classification framework for voice across various agent forms, contexts, and user groups, a critical analysis grounded in key theories, and an identification of future directions for the oncoming wave of vocal machines.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; **Sound-based input/output**; *Auditory feedback*; • **Social and professional topics** → *User characteristics*;

Additional Key Words and Phrases: Computer agent, computer voice, synthetic speech, voice perception, vocalics, conversational agents, voice assistants, robots, embodied AI, embodied agents, voice-user interface (VUI), human–agent interaction (HAI), human–computer interaction (HCI), human–robot interaction (HRI), human–machine communication (HMC)

ACM Reference format:
Katie Seaborn, Norihisa P. Miyake, Peter Pennefather, and Mihoko Otake-Matsuura. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.* 54, 4, Article 81 (April 2021), 43 pages.
https://doi.org/10.1145/3386867

Katie Seaborn completed this work while a Postdoctoral Researcher and later Visiting Researcher at RIKEN AIP.

This work was funded in part by the Japanese Society for the Promotion of Science (JSPS) through Grants-in-Aid for Scientific Research (KAKENHI Grants No. 20H05022 and No. 20H05574).

Authors. addresses: K. Seaborn, Tokyo Institute of Technology and RIKEN Center for Advanced Intelligence Project (AIP), W9-51, 2-12-1 Ookayama, Meguru-ku, Tokyo, 152-8552, Japan; email: seaborn.k.aa@m.titech.ac.jp; N. P. Miyake, RIKEN Center for Advanced Intelligence Project (AIP), Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan; email: norihisa.miyake@riken.jp; P. Pennefather, gDial Inc., Toronto, 87 Earl Grey Road, Toronto, Ontario, M4J 3L6, Canada; email: p.pennefather@gmail.com; M. Otake-Matsuura, RIKEN Center for Advanced Intelligence Project (AIP), Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo, 103-0027, Japan; email: mihoko.otake@riken.jp.

0360-0300/2021/04-ART81 $15.00
https://doi.org/10.1145/3386867

1 INTRODUCTION

Voice is an aural communication medium that carries sound-based information from a speaker to one or more receivers. As a natural phenomenon, voice has emerged over the course of human evolution to dominate communication. When we think of voice, we tend to think of *speech*: spoken language in the linguistic sense of words produced by the combination of vowel and consonant sounds, or *phonetics* [56]. Speech can also be non-linguistic, such as when babies coo and cats meow. But speech is only one aspect of voice. Voice can also transmit expressive non-linguistic information: it can signal emotion and affective states; it can portray social characteristics, such as gender and personality; and it can be used as a tool, such as when singers produce music. In nature, voice plays a key role in moderating interactions within and between species [73], including collaboration, such as between humans and animals, like dogs [73], and possibly artificial agents [56]. From this evolutionary perspective, voice is a fundamental feature of social interaction for and beyond people.

The emergence of voice as a factor of experience when interacting with computers, especially ones that use **artificial intelligence (AI)**, may then come as no surprise. In the past few years, a range of voice-based and AI-infused [5] technologies have appeared. From Apple's Siri on the iPhone to Google's Alexa and other smart speakers, these "virtual assistants" that interact with people by way of voice user interfaces have proliferated within everyday life [89]. As computer-based agents, they follow in the footsteps of human–computer interaction work on social robots [17], chatbots [1], and conversational agents [115]. As a form of embodied AI [19], they represent an expected "platform convergence" [67, 48] between the often disparate fields of AI and human–computer interaction. From an engineering communication perspective, the closing gap between such agents and people is represented by the field of human–machine communication [68, 71]. The computer science analogues are human–robot interaction, **human–agent interaction (HAI)** [104] and human-AI interaction [5]. Recent commercial and public strides in consumer voice-based AI systems have pushed these interrelated fields to the forefront of scholarship on computer voice.

Driving research on voice is the idea that the computers of best fit are the ones most like people. Much of social robotics work, for instance, has explored replicating anthropomorphic forms, human mannerisms, and social modes of communication [17, 71]. Bolstering this approach is decades of research showing that people tend to treat computers *as if they are people*, often without realizing it [136]. In this body of work, voice is not a new topic. Recent reviews on conversational agents [52] and smart devices [21] have revealed that aspects of voice are key in the design of agents. Other surveys have focused on detection and conveyance strategies from a technical angle: algorithms and hardware that "perceive" various aspects of human communication [22, 144], such as affect [36, 59, 111]), non-semantic vocalizations [192], non-verbal cues [52], including laughter [31], and other expressive qualities [64]. Specific behavior change outcomes, such as for health [160], and specific populations, such as children [45], have also been surveyed. Yet, there is a gap when it comes to the state of the art on "the voice of the machine" as an *expressive and influential* factor across *different kinds* of agents. We are motivated by a growing recognition that the creators and gatekeepers of these agents—computer scientists, engineers, designers, data providers, researchers, producers, managers, CEOs, and others—have a key role to play in perpetuating or disrupting social and ethical status quos, as well as in conscientiously designing technologies for the ethical and social good [21, 78, 178, 187].

To this end, we mapped the state of affairs in research on the *perception of agent voice and its effects* in **voice-based human–agent interaction (vHAI)** across a *range of agent types*. This work is a vital step toward identifying what we know (and do not know) so far, as well as developing an agenda for research and practice. We asked: *What is the nature of agent voice and its effects in*

human–agent interactions? To provide a multi-faceted answer of interest to a variety of stakeholders, we asked the following sub-questions:

- Q1: What kinds of voice-based agents have been evaluated?
- Q2: What enabling technologies have been used?
- Q3: What characteristics of voices have been explored?
- Q4: What measures have been evaluated?
- Q5: What evaluation methods have been used?
- Q6: What user groups have been included?
- Q7: What application contexts have been explored?
- Q8: What research designs have been used?
- Q9: What theoretical frameworks have been used?
- Q10: What are the major results and findings?

We begin by describing our methods, then report and discuss the results of our search and synthesis. We end with trajectories for future work. The major contributions of this work are threefold: (i) a systematic and multidisciplinary synthesis of the state of the art; (ii) an empirically grounded classification framework; and (iii) a research agenda for addressing the gaps and opportunities for voice in human–agent interaction.

2 METHODS

A meta-synthesis approach [75, 83] was taken to survey the experimental literature on voice in HAI experiences. The goal of meta-synthesis is to extract mixed forms of data, including descriptive statistics, inferential statistics, and qualitative findings to describe the state of the art and uncover consensus, or a lack thereof. It is thus appropriate for drawing generalizations from a body of experimental work that may include one or both of quantitative and qualitative data [75, 83], which is representative of vHAI research. It is also useful for uncovering the nature of the research in a given domain at a meta level (e.g., research designs, theoretical frameworks, etc.). We focused on experimental work to make generalizations about vHAI findings, in line with Kitchenham's guidelines and protocol for engineering surveys [88]. We detail our process in Section 2.2. To start, we next define our key terms and how these affected our process.

2.1 Defining Terms

Some of our key terms can be defined in different ways, depending on the context or the discipline, while others are difficult to define in general, due to reasons such as a lack of consensus. Here, we specify definitions for these terms and discuss any definitional and ontological issues relevant to our survey work.

2.1.1 Voice and Vocalics. We define voice as an *expressive aural medium of communication.* From a technical standpoint, voice refers to the manifestation of sound heard by a recipient and vocalized by a sender through some vocal mechanism [29, 91, 92]. Put simply, it is the "how" of vocalizations [21, 171]. The vocal mechanism has traditionally meant the vibration of human or animal vocal tracts, but since the development of computer-based media and related technologies, it can also include machine-based synthesis of sound. From a social standpoint, voice is how "…speakers project their identity—their "physical, psychological, and social characteristics"—to the world" [92, 115, 97]. Voice can also be defined in other ways: grammatically, as in the relationship between the speaker and verb (e.g., active vs. passive voice); phonetically, as in speech sounds produced by vocal cords (or other mechanisms) that can be above or below the threshold of hearing; as a part of a musical composition; and idiomatically, as in to declare a position (to "give voice to" an idea).

However, we are strictly considering voice as the *sound medium through which communication is expressed.*

Voice is characterized by *vocalics*: nonverbal paralinguistic [147] properties—tone, loudness, pitch (frequency), and timbre (voice quality)—and nonverbal prosodic properties—rhythm, intonation (variation in pitch), stress (emphasis). Since these properties can and do vary, voice is a multi-dimensional, able to represent and relay more than just speech or language information. This can include emotion and affective states (e.g., valence and intensity), social signals and conventions (e.g., dialects and accents), social identities (e.g., personality and gender), and biological states (e.g., age and health) [85]. Voice does not necessarily express linguistic content or speech, i.e., words and sentences. For example, babbling, whistling, and vocal fillers or pauses are non-linguistic or pseudo-linguistic [76, 150]. An important application of this is in the orchestration of dialogue and conversation to cue turn-taking [65]. We will continue to discuss these characterizations below.

For this survey, we included papers that treated voice as the "how" of communication. For instance, studies of voice-based agents that did not evaluate voice and vocalics were excluded, because they could not provide specific insights on human perceptions of or behavioral responses to the voice of the agent.

2.1.2 Speech. The terms "voice" and "speech" are often used interchangeably [131, 145]. Yet, they refer to different, albeit interconnected phenomena. Speech is the *linguistic content* of voice, primarily comprising words (vocabulary, including pseudo-vocabularies and slang), grammar and syntax, and phonetics (the physical measurement of speech). It is the structured *language* of communication [145], or the "what" of vocalizations [21, 171]. Voice, then, is a[1] medium of speech. In this way, speech is the *content* represented by the *medium* of voice. Latinus and Belin [96] provide a clear example in overhearing a conversation on a noisy plane: meaningful information, such as the emotional states of the speakers, can be gleaned from vocal stimuli even without direct access to speech stimuli (i.e., words). However, it can sometimes be difficult to draw a line between voice and speech. This is especially true for *pseudo-linguistic utterances*, such as vocal fillers and gibberish. Our position, in line with Read and Belpaeme [150], is that voice carries semantic meaning while speech carries linguistic content. From this perspective, the use of gibberish is an aspect of voice with content but without speech, while vocal fillers may occupy a liminal space, depending on the "filler" used and its semantic purpose (e.g., "I mean…" may be speech, because its linguistic semantic meaning is preserved, while "um" has no linguistic content).

For this survey, we include work where voice is treated as an expressive medium and exclude work on speech and/or where voice is defined as speech. We specifically include pseudo-linguistic utterances when they are treated as an aspect of voice (or vocalics) and not only as speech (content).

2.1.3 Artificial Intelligence (AI) and Artificial Agents. Human–agent interaction involves people interacting with an embodied AI of some kind in some direct way. We use the term "HAI" to refer to this concept, in line with its historical usage and that of a major international conference.[2] Key to this definition are the concepts of "embodied AI" and "agent," which we will break down here. When it comes to "AI," there are many conceptualizations. According to Nils Nilsson, co-founder of the field of AI, consensus on a definition remains elusive, primarily because consensus on what "intelligence" means remains elusive. Nevertheless, he proposed the following

[1]We use "a" rather than "the" due to recent innovations in brain–computer interfaces that are forecasting an alternative to voice as a medium of speech, with direct translation to text from neural signals [124]. But a usable communication system is still a long way off.

[2]http://hai-conference.net/.

comprehensive definition for "intelligence": "a quality that enables an entity [agent] to function appropriately and with foresight in its environment" [139, 13]. The entity (or agent) can be artificial or otherwise. There are no particular skills or tasks, algorithms or mechanisms, learning methods (e.g., machine learning) or interaction policies, technologies or devices, or form factors or morphologies required. Importantly, intelligence (artificial or otherwise) is a "quality" that is *perceived*, and in HAI, it is perceived *by people* through an agent's *embodiment* in the world. Embodiment is an agent's situatedness within an environment (whether physical, virtual, or mixed) and its interaction with aspects of that environment, including itself and other agents, through its sensors and actuators [143]. This conceptualization is broad enough to encompass the diversity of vHAI varieties: voice assistants and virtual humans (embedded in various forms of computers or "smart" systems and spaces, such as speakers, homes, and motor vehicles), conversational interfaces and agents (including chatbots), intelligent personal assistants, robots (social or otherwise, such as collaborative robots in industrial settings), voice user interfaces and speech interfaces, and virtual agents (including those in video games). At the same time, it is specific enough to exclude typical computer-based systems, non-computer intelligences (i.e., people and other animals), and pseudo-intelligent agents, such as automatons. By taking on this conceptualization of the "agent," we can consider the factor of voice in human-interfacing agents in a *platform-neutral way*. This allows us to extend the work of specialized surveys with more transferable findings about the human factor in agent voice.

2.2 Procedure

We describe the procedure we undertook for our two-phase systematic survey and meta-synthesis of the literature. Our flow diagram in the PRISMA format [120] is shown in Figure 1.

2.2.1 Eligibility Criteria. A set of inclusion and exclusion criteria were developed based on the research question and Kitchenham's [88] guidelines for selecting papers in engineering. Inclusion criteria were: full and complete primary studies of voice with experimental results on vocalics; a focus on perception of agent voice rather than detection and generation voice or control of interfaces through voice; original human subjects research; and peer-reviewed journal papers and proceedings-based conference papers. Exclusion criteria were: qualitative research, which is difficult to analyze and/or generalize in the same way as experimental work [34, 55, 146]; studies in which voice was not a primary factor; research designs and/or reporting of results whereby voice data was inseparable from other data; inaccessible papers; papers not in English; and low quality research designs and papers (notably those lacking the details necessary to evaluate their eligibility).

2.2.2 Selection Process. Scoping searches were conducted in two phases. The selection process for each phase followed the same basic procedure. It began with the first author performing two rounds of review to ensure basic relevancy: the first round on the paper's abstract, and the second on the content of the paper. After this, at least two authors reviewed the full papers and rated the quality of the research reported. The first and second phases differed based on the number of papers and raters involved. The differences will be described in turn. After establishing inter-rater reliability, one author then rated and determined final inclusion for all papers. Papers cited within the included papers but not found by query were assessed and included as found.

In the first phase, two authors rated the papers in the third round of review. To do this, a randomly selected set of 20% of the papers were reviewed independently by each rater. Random.org's Random Integer Set Generator was used to generate a list of random numbers corresponding to rows (one per paper) in the spreadsheet. Raters used an adapted version of Kitchenham's [88] study design hierarchy to evaluate the papers. Specifically, it included experimental work,

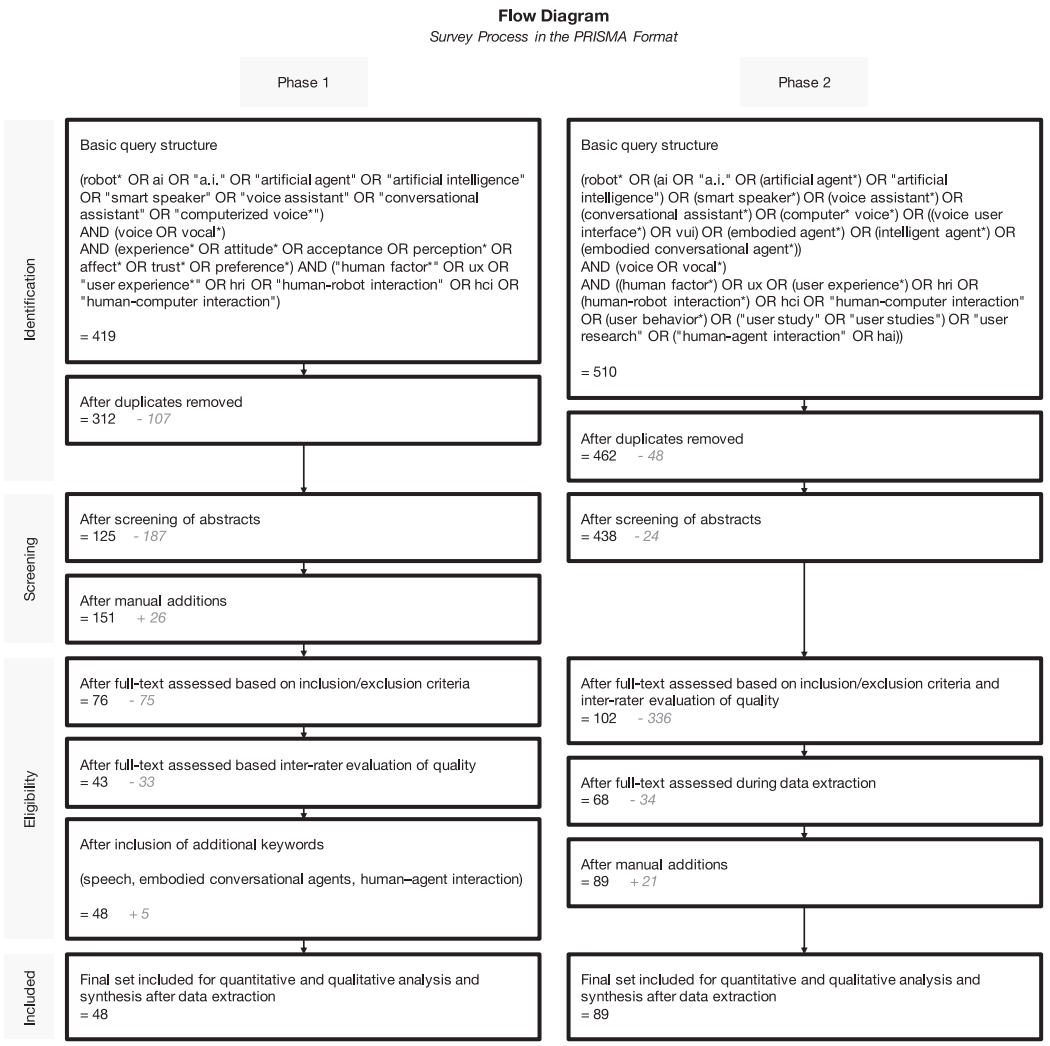

Fig. 1. Flow diagram for the survey process using the PRISMA format.

with true experiments or **randomized controlled trials (RCTs)** ranked highest, imperfect (not randomized, or low n) experiments placed second, experimental field studies placed third, and exploratory lab or field studies or classification studies last. Non-experimental studies, such as case studies, interviews, expert reviews, and survey papers, were not included. Cohen's Kappa statistic was used to assess inter-rater reliability. In uncertain cases, raters discussed after completing their ratings. Values of 90% or greater were achieved, indicating strong agreement, according to Cohen [30]. In the second phase, the initial set of papers was much higher, so three authors were involved in the third round of review. Due to this, percentage agreement was used instead of Cohen's (two-rater) Kappa for assessing inter-rater reliability. An agreement value of 85.7% was achieved across all three raters.

2.2.3 Search Query Keywords and Totals. All subject areas of the Ovid Inspect, Web of Science, the ACM Digital Library, IEEE Xplore, and Google Scholar databases were searched. The

queries differed slightly based on each database's requirements. Alternative term forms, such as pluralization, were considered. In the first phase (9th August 2019), the query keywords were: *robot, AI, artificial agent, artificial intelligence, smart speaker, voice assistant, conversational assistant, computerized voice, voice, vocal, experience, attitude, acceptance, perception, affect, trust, preference, human factor, user experience, UX, human–robot interaction, HRI, human–computer interaction, HCI.* A total of 419 papers were returned. In the first round of review, 151 papers (with 26 manual additions from references) were selected. In the second round, 76 papers were selected. After the inter-rater reliability stage and third round of review, 43 papers were selected. In the second phase (18th December 2019), keywords were added based on external review of the paper: *speech, embodied conversational agents, human–agent interaction.* A final 48 papers were selected. In the second phase (4th August 2020), additional keywords were included by recommendation of peer reviewers: *voice user interface, VUI, embodied agent, intelligent agent, embodied conversational agent, user behavior, user study, user research.* This led to a total of 510 papers. After two rounds of review by the first author, a total of 438 were selected. After the third round, three raters selected a total of 102 papers. After data extraction, which involved removals and manual additions, a final 89 papers were selected.

2.2.4 Data Extraction and Analysis. Data extraction for descriptive synthesis [88] was carried out and organized by sub-question. This involved generating descriptive statistics for the quantitative data and identifying patterns in the qualitative data, specifically the research findings. For the quantitative analysis, the first author generated descriptive statistics for the metadata categories per sub-question, specifically counts and percentages comparing to the entire corpus of papers. Mean, standard deviation, median, interquartile ranges were generated for the participant demographics data. For the qualitative research findings data, patterns were extracted individually by three researchers and then collated and prioritized for inclusion by the first author.

3 RESULTS

The results before synthesis are grouped by sub-question. Q1–Q9 are summarized in Figure 2 (frequencies) and Figure 3 (chronology), with additional details described below (e.g., what constituted "Other" categories). Since Q10 generated qualitative results, these are summarized in-text below. All papers are listed in Table 1.

3.1 Agent Types and Enabling Technologies (Q1, Q2)

Computer voice agents (38%) and robots (48%) made up the majority of agent types, although the use of robots is a more recent shift (see Figure 3). Virtual characters, virtual humans, screen-based conversational agents, and avatars made up 25%. Smart "things" (speakers, vehicles, etc.) made up 6%. Controls included humans (19%) and text (10%). A range of enabling technologies were used: robots, screen-based characters, and **text-to-speech (TTS)** systems. In terms of robots, 13% used the Aldebaran Nao robot; the rest were other commercial robots and prototypes (29%) or custom robots (12%). Seven used one system for comparison of different embodiment forms: Nao, Flobi, Valerie, KOBIAN, MobileRobots Pioneer P3AT, FACE, and Eva. In terms of virtual characters (15%), 31% used a virtual version of a robot and 23% created custom characters. For voice, 69% used a TTS, 22% relied on a human actor/recording, and one study used sounds as a comparison. The most common TTS's were the CSLU ToolKit (6%) and Amazon Polly (6%). The DECTalk Express V2.4C, MacOS TTS, and Festival Speech Synthesis were used three times each. IBM Watson, Voicery Speech Synthesis, Microsoft Mary, Neospeech Katie, and Loquendo TTSwere used twice each. 19% used a custom TTS. For 15% of papers, the nature of the voice used was not reported, and for 9%,

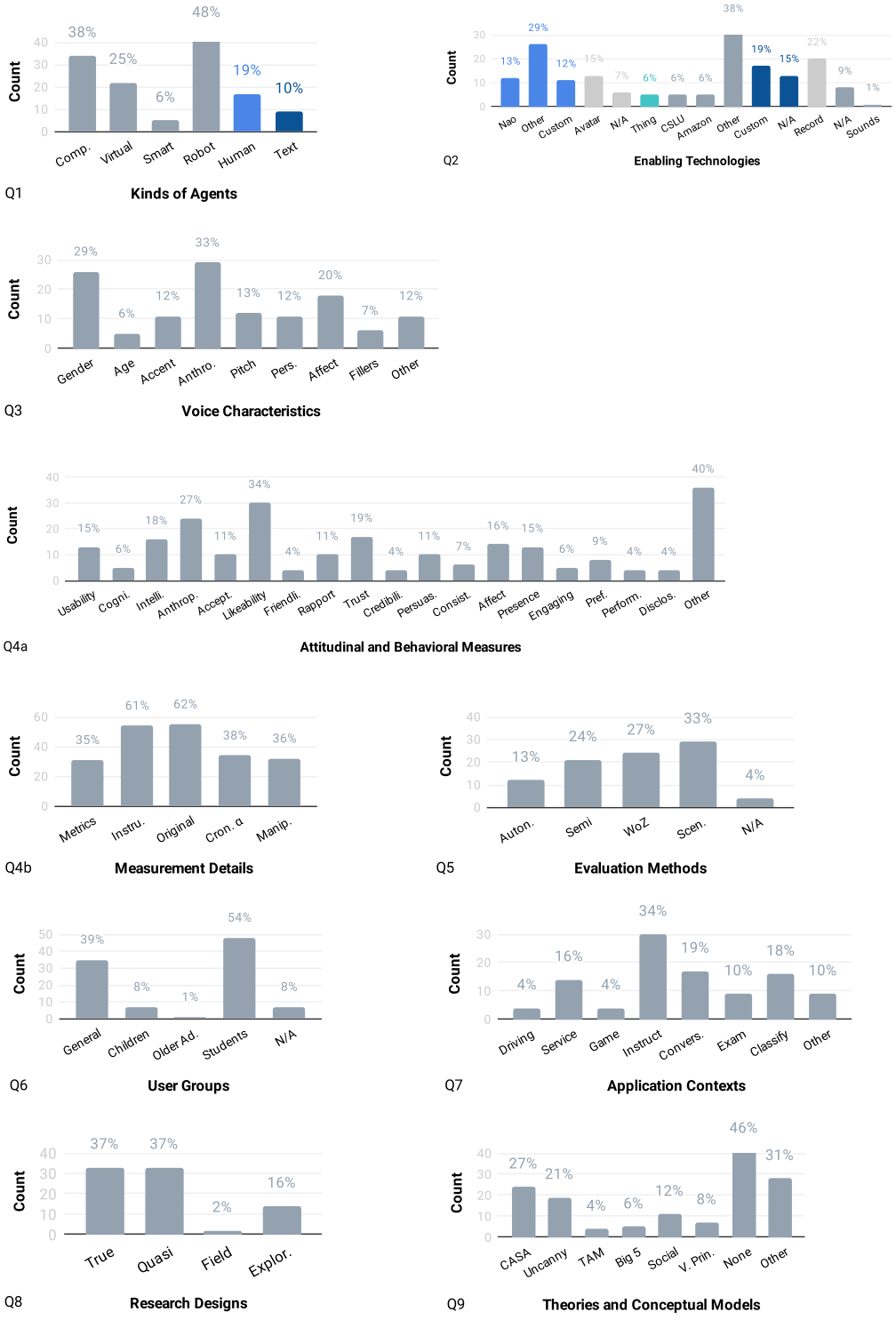

Fig. 2. Bar graphs showing quantitative summary of results for Q1–Q9.

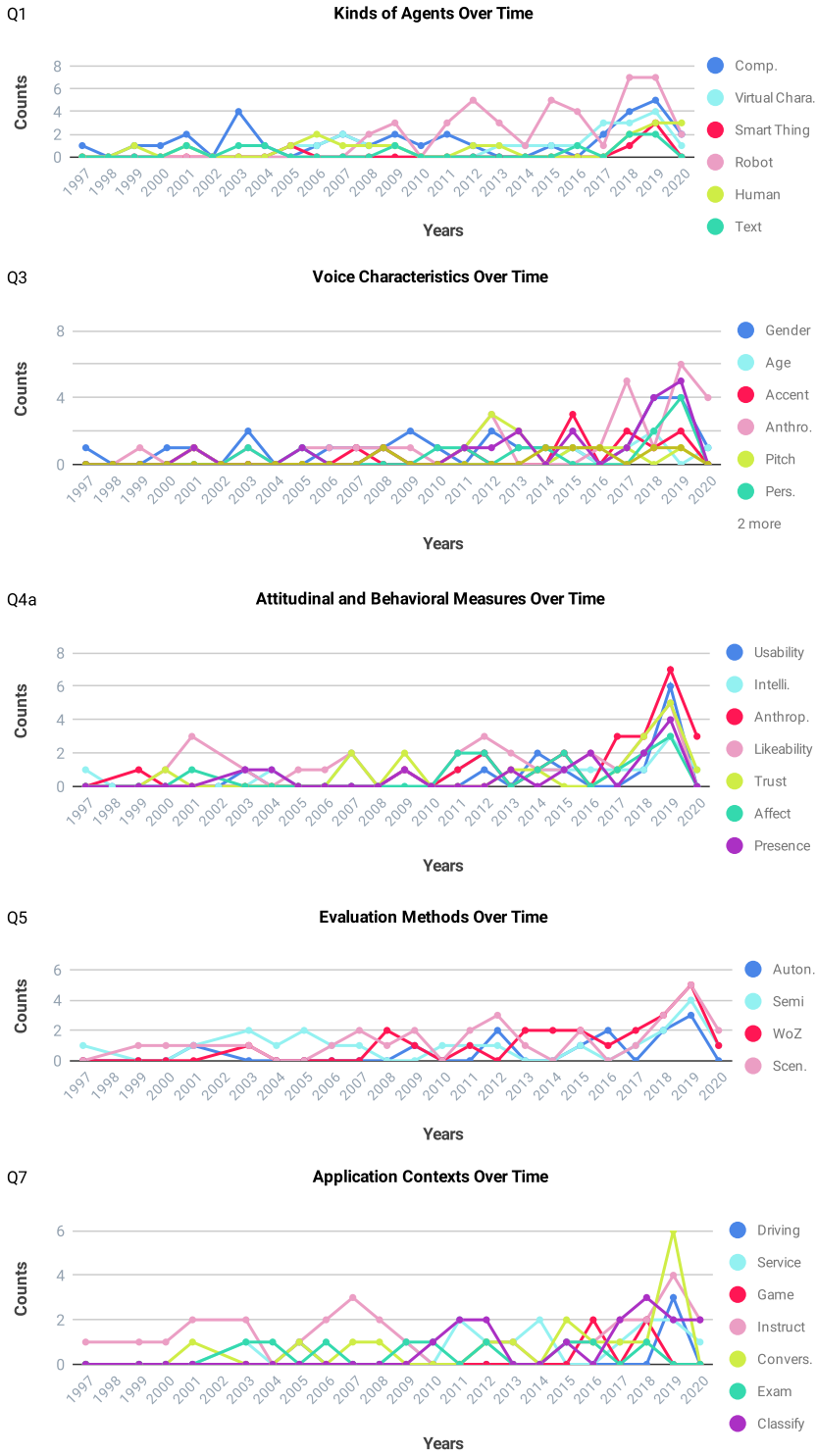

Fig. 3. Line graphs showing chronological quantitative summary of results for Q1, Q3, Q4a, Q5, and Q7.

Table 1. The Surveyed Papers at a Glance, Listed in Alphabetical Order, then by Date

		Year	# of Exp.	Total n	Types of Agents			Comparison					Characteristics					
ID	Citation				Com. Voice	Virt. Char.	Smart Thing	Robot	Hum.	Text	Gend.	Age	Acce.	Anth.	Pitch	Perso.	Affect	Filler
49	Abdulrahman et al. [2]	2019	1	118		•								•				
1	Andrist et al. [6]	2015	2	32				•					•					
50	Atkinson et al. [8]	2005	2	90		•			•					•		•		
51	Baird et al. [10]	2017	1	23	•						•	•	•	•				
52	Baird et al. [9]	2018	1	23	•						•	•		•				
53	Behrens et al. [13]	2018	1	80				•			•							
54	Bhagya et al. [14]	2019	1	32	•												•	
2	Bracken et al. [15]	2004	1	134	•					•								
3	Braun et al. [16]	2019	1	55			•									•		
55	Cambre et al. [20]	2019	1	1090	•				•	•	•			•				
4	Cao et al. [23]	2019	1	32				•					•					
5	Chang, Lu, and Yang [24]	2018	1	226	•						•					•		
56	Chérif and Lemoine [25]	2019	1	640		•								•				
6	Chidambaram et al. [26]	2012	1	32				•							•			
57	Chiou, Schroeder and Craig [27]	2020	1	89		•			•					•				
7	Chita-Tegmark and Scheutz [28]	2019	3	494				•	•		•						•	
58	Craig and Schroeder [32]	2017	1	140		•								•				
8	Creed and Beale [33]	2008	1	68		•									•		•	
9	Crowell et al. [35]	2009	1	44	•			•			•							
59	Dahlbäck et al. [38]	2001	1	79	•								•					
60	Dahlbäck et al. [39]	2007	1	96	•								•					
61	Davis, Vincent and Park [42]	2019	1	172		•								•				
10	Donahue and Scheutz [43]	2015	1	55		•		•									•	
62	Dou et al. [44]	2020	1	36				•			•	•		•				
63	Elkins and Derrick [46]	2013	1	88		•					•					•		
64	Evans and Kortum [49]	2010	2	303	•						•					•		
11	Eyssel et al. [50]	2012	1	58				•	•		•			•				
12	Eyssel et al. [51]	2012	1	58				•			•			•				
13	Ghazali et al. [60]	2019	1	18				•									•	
14	Goble and Edwards [61]	2018	1	67				•										•
65	Gong and Lai [62]	2003	1	24	•													
15	Gong and Nass [63]	2007	2	160		•			•					•				
16	Hennig and Chellali [74]	2012	1	63	•									•			•	
17	Hoegen et al. [77]	2019	1	30			•								•			
18	James, Watson and MacDonald [82]	2018	1	120				•									•	

(Continued)

Table 1. Continued

ID	Citation	Year	# of Exp.	Total n	Types of Agents			Comparison					Characteristics					
					Com. Voice	Virt. Char.	Smart Thing	Robot	Hum.	Text	Gend.	Age	Acce.	Anth.	Pitch	Perso.	Affect	Filler
66	Jeong, Lee and Kang [84]	2019	1	26	•													•
67	Kim, Goh and Jun [87]	2018	1	20		•				•								
68	Komatsu et al. [90]	2011	1	20	•			•									•	
19	Krenn, Schreitter, and Neubarth [93]	2017	1	91	•								•	•				
20	Lazzeri et al. [98]	2018	1	25		•		•	•								•	
21	Lee and Nass [100]	2003	2	152	•											•		
22	Lee, Nass, and Brave [99]	2000	1	48	•						•							
23	Lee, Ratan, and Park [102]	2019	1	158			•				•					•		
69	Lee, Sanghavi, Ko and Jeon [103]	2019	1	12	•			•								•		
24	Lubold, Walker and Pon-Barry [106]	2016	1	48				•							•			
70	Mara, Schreibelmayr and Berger [107]	2020	1	165	•				•					•				
25	McGinn and Torre [114]	2019	1	90	•			•	•		•		•	•				
71	Mendelson and Aylett [116]	2017	2	147		•								•			•	
72	Moreno et al. [122]	2001	4	235		•				•								
26	Mullennix et al. [126]	2003	1	195	•						•			•				
27	Nass et al. [130]	2001	1	56	•				•		•			•	•		•	
28	Nass et al. [135]	2003	1	100	•					•	•							
73	Nass et al. [132]	2005	1	40			•										•	
29	Nass, Moon, and Green [134]	1997	1	40	•						•							
74	Niculescu et al. [138]	2011	1	28				•							•	•		
30	Niculescu et al. [137]	2013	1	28				•							•		•	
75	Ohshima et al. [140]	2015	1	27				•										•
76	Ohta et al. [141]	2014	1	24		•												•
77	Qiu and Benbasat [148]	2009	1	168	•	•				•								
78	Read and Belpaeme [149]	2012	1	42				•							•			
31	Read and Belpaeme [150]	2015	1	29				•							•		•	
32	Rosenberg-Kima et al. [153]	2007	1	89	•	•					•							
33	Rosenthal-von der Pütten et al. [154]	2016	1	130		•		•		•				•				
34	Sandygulova and O'Hare [156]	2015	1	64				•			•	•	•					
35	Sandygulova and O'Hare [157]	2018	1	107				•			•	•						
79	Sarigul et al. [158]	2020	1	17	•			•	•					•				
36	Shamekhi et al. [161]	2018	1	40	•													

(Continued)

Table 1. Continued

ID	Citation	Year	# of Exp.	Total *n*	Types of Agents			Comparison					Characteristics					
					Com. Voice	Virt. Char.	Smart Thing	Robot	Hum.	Text	Gend.	Age	Acce.	Anth.	Pitch	Perso.	Affect	Filler
80	Shi et al. [162]	2018	1	24		•				•							•	
81	Shibata et al. [163]	2012	2	64				•							•			
82	Shiwa et al. [164]	2008	1	38	•			•										•
83	Siegal et al. [165]	2009	1	134				•			•							
84	Sims et al. [166]	2009	1	96				•						•				
37	Stern et al. [169]	1999	1	193	•				•					•				
38	Stern et al. [170]	2006	2	400	•				•					•				
39	Tamagawa et al. [174]	2011	2	111	•			•					•	•				
85	Tay, Jung and Park [176]	2014	1	164				•			•					•		
40	Torre et al. [179]	2018	3	274	•			•					•					
86	Torrey, Fussell and Kiesler [180]	2013	1	77				•	•									
41	Trovato et al. [182]	2017	1	60		•		•						•	•			
42	Tsiourti et al. [183]	2019	1	30				•									•	
43	Vannucci et al. [184]	2018	2	30				•	•								•	
44	Walters et al. [186]	2008	1	58				•	•		•			•				
87	Wigdor et al. [189]	2016	1	26				•										•
45	Xu [190]	2019	1	110				•						•				
46	Yarosh et al. [191]	2018	1	114			•									•		
47	Yilmazyildiz, Verhelst and Sahli [193]	2015	5	99	•			•					•		•		•	
88	Yu et al. [195]	2019	1	30		•				•								
89	Zaga et al. [196]	2016	1	18				•										
48	Zanbaka et al. [197]	2006	1	138		•			•		•							
	Range, Mean, or Count	**24 yrs**	**110**	**7888**	**34**	**22**	**5**	**43**	**17**	**9**	**26**	**5**	**11**	**29**	**12**	**11**	**18**	**6**
	Percentage				**38%**	**25%**	**6%**	**48%**	**19%**	**10%**	**29%**	**6&**	**12%**	**33%**	**13%**	**12%**	**20%**	**7%**

the nature of the human voice used (e.g., a live actor or a recording, or a human-recording-based TTS) was not disclosed.

In short, there is little diversity in agent type, but great diversity in the enabling technologies. Notably, few projects included smart "things," such as smart speakers, smart TVs, and smart vehicles. Of these, three were vehicles and two were speakers, pointing to a limited focus on the kinds of "things" possible. None involved the most common options on the market today: Siri, Alexa, and Google Assistant. However, there is also a lot of information missing (in total, a gap of 30%), limiting our ability to draw firm conclusions.

3.2 Voice Characteristics, Measures, and Measurement (Q3, Q4)

The most common characteristics studied were anthropomorphism, i.e., "natural" human-like versus synthetic (33%), gender (29%), and affect (20%). About 12% of studies considered each of pitch, accent, and personality. 12% considered other characteristics, such as quality of synthetic voices, personalization of the content, and conversational style. Six percent of studies considered age and 7% considered vocal fillers. In the past five years, anthropomorphism, affect, gender, and personality have emerged as key factors (see Figure 3).

A variety of measures were used to evaluate these characteristics and their effects, which can be roughly divided into *perceptions of and attitudes* and *effects on the user*. Most measures (40%) were unique, i.e., not found in other studies. The most common were likeability (34%), in which interest has fluctuated over time (see Figure 3), anthropomorphism (24%), trust (19%), intelligence or competence (18%), affect (16%), social presence (15%), and usability, which includes understandability (15%). All have surged in popularity after 2016 (see Figure 3). Examples of unique measures include stress, leadership, source of the voice, similarity to the user, gender-stereotyped attributes (e.g., assertiveness), compliance, social skillfulness, honesty, and self-confidence. Most of these were associated with the type of agent, task, and/or context, such as improving the user's self-efficacy, approach distances for social robots, willingness in the user to disclose personal information, perceived obligation to speak, and performance in cooperative tasks. In short, there were a vast range of measures uncovered across papers, making it difficult to summarize or draw consensus.

Almost all studies used questionnaires to gather self-reports and other subjective data from participants. 62% used original instruments developed by the authors, while 63% used validated instruments; 19% used a combination. There was a mean Cronbach's α of 0.84 (SD = 0.07); the lowest α was 0.53 [170]. Fifty-two percent of papers did not provide this or another measure of internal consistency. Metrics based on system records or quantified observations were collected in 35% of cases. Manipulation checks were performed in 36% of papers.

3.3 Evaluation Methods (Q5)

Four key evaluation methods were identified. Thirteen percent of studies used an autonomous setup, involving a fully working system with no experimenter involvement and the interaction left up to the participant; use has been steadily increasing since 2014 (see Figure 3), in line with advances in technology. Twenty-four percent of studies used a semi-autonomous setup, involving a working but limited system, such as script-based, pre-recorded, or interactive simulations. Twenty-seven percent used a Wizard of Oz setup [37], where an actor controlled the agent and/or spoke for the agent in realtime. Thirty-three percent used scenarios, which could be video, audio, or static image of a scene or sample, often used for classification studies. Four percent did not specify the evaluation method or provide enough detail for it to be determined. While there has been some fluctuation, all have experienced increase use since about 2016 (see Figure 3).

3.4 User Groups and Application Contexts (Q6, Q7)

Out of a grand *n* of 7,888, the largest user group was college or university students (54%), with the next being adults in the general population (39%). Eight percent included children and 1% included older adults. Eight studies featured dyadic user groups, specifically adults and children or adults and students. The average age was 25.4 (SD = 5.5). 3,025 men and 3,331 women were identified, with 1,532 unidentified, and 19 classified as "other"; a t-test did not find a significant difference between the total number of men and women. In summary, while there appears to be gender parity across studies, there was an overreliance on young adult students.

Application contexts were instruction (34%), conversation (19%), classification (18%), and service tasks (16%). The uses of instruction and classification contexts have fluctuated over time but have experienced a resurgence since 2016 alongside most other contexts (see Figure 3). Conversation, in particular, has recently experienced as surge in use since 2017, in line with a renewed focus on **conversational user interfaces (CUIs)** at large. Other contexts included watching videos, being persuaded, experiencing different life scenarios, approaching an agent in a space, and exploration of an environment. From an ecological validity perspective, most studies involved some kind of realistic context, e.g., education, conversation, driving. A variety were employed, representing different situations in which voice may be an important factor. Yet, this diversity does not

make for easy generalizations. One exception is the instruction context (comprising 40% of studies surveyed), where the voice-based agent provided some kind of description or explanation.

3.5 Research Designs (Q8)

Thirty-seven percent of studies used a true experiment or **randomized controlled trial (RCT)** design involving hypotheses, one or more controls, and randomized allocation of participants to groups. 37% used imperfect or quasi-experimental designs, such as those without random assignment or with low participant numbers. There were two field experiments and 14 (16%) exploratory lab or field experiments or classification studies.

3.6 Theories and Conceptual Models (Q9)

The majority did not reference any theory or conceptual models (46%), making it is unclear how to position the findings of most studies. This was true even for experiments that had hypotheses: about 30% did not reference theory. Most theories related to the social perception of agents, especially their anthropomorphism: Computers Are Social Actors (27%) followed by the Uncanny Valley theory [123] (21%), and social agency theory (12%) [109]. Others related to voice: the voice principle/effect (8%) [8], Tannen's theory of conversational style [175], Brennan and Hulteen's conversational feedback model for voice agents [18], and the theory of doubly disembodied language [101]. Others were specific to the voice characteristics and measures. For personality, the "Big Five" or five-factor model of personality [112]. For acceptance, the **Technology Acceptance Model (TAM)** [41], Social Identity Theory [173], and similarity-attraction theory [133]. For trust, Mayer's model [110]. For anthropomorphism, the Three-Factor Theory of Anthropomorphism [48]. For affect, Russel's model of valence–arousal [155]. For cognition, cognitive load theory [172]. For behavior, the Theory of Planned Behavior [4], and Hall's social zone distances [72]. The theories and conceptual models cited do not necessarily represent all models or even the key models in the extant literature. While there is great diversity, most of these are based on human models developed in other disciplines and applied to the case of agent voice.

3.7 Findings (Q10)

We group the findings by voice characteristic or factor of vHAI experience.

3.7.1 Gender. People tended to react to voice gender in line with stereotypes. When vocal gender markers were present, Chita-Tegmark et al. [28] found that participants applied gender stereotypes to robots. Similarly, McGinn and Torre [114] found that participants perceived mechanical voices as masculine and "round" voices as feminine. Lee et al. [99] found that masculine voices were rated as more persuasive, socially attractive, and trustworthy than feminine voices, confirming stereotypes. Similarly, Nass et al. [134] found that masculine synthetic voices were rated more friendly and competent than feminine ones. They also found that the voice was better received when its gender matched stereotypical expectations for gendered speech content. Stereotypes were also confirmed by Lee et al. [102] with "informative" masculine and "social" feminine voice agents. Masculine synthetic voices were rated as more favorable and persuasive than feminine synthetic voices in a study by Mullennix et al. [126]. Lubold et al. [106] found that social presence was rated higher when the perceived gender of the robot and its voice matched. In contrast, Lee and Nass [100] found that perceived gender did not affect ratings of social presence. Tay et al. [176] found that a healthcare robot with a feminine, extraverted voice and a security robot with a masculine, introverted voice were rated highly, although gender was not as strong as personality in terms of role matching. In contrast, Dou et al. [44] found that masculine agents were considered more

suitable for shopping and the home, while feminine and masculine voices were rated appropriate for education.

In terms of assigning gender, there appears to be a male/masculine bias across age groups, as well as an effect of providing limited gender categorization options. Walters et al. [186] found that a robot was labeled either neutral or male. Similarly, Sandygulova and O'Hare [157] found that children, in particular younger children aged 5-8, misattributed a male gender to a feminine-voiced robot. Baird et al. [10] found that most voices were accurately categorized by gender, except for a male German voice (rated 73.9% masculine). They also found that it was least humanlike, suggesting a link between gender and anthropomorphism. In a later study, Baird et al. [9] found that raters did not agree on assignments of gender after they revised their assignment options to include a non-binary gender. This was especially true for the feminine Japanese voice, which was also rated as least humanlike, reinforcing the link between gender and anthropomorphism.

Preference for voice varied. In several studies [50, 51, 99, 157], same gender robots were preferred. Although there were individual differences, Chang et al. [24] found that robots with feminine, extroverted voices were most liked and preferred in general. Niculescu et al. [138] found that higher-pitched voices were more likeable and perceived as extraverted, especially for men. Behrens et al. [13] found that the masculine voice was perceived as friendlier and more trustworthy, and a better fit for Nao. Cambre et al. [21] found that masculine voices received the highest ratings regardless of participant gender.

The role of participant gender on behavioral outcomes varied. It was not a factor in the experience of a persuasive robot [26]. However, Zanbaka et al. [197] found that men were persuaded by female-voiced virtual characters and vice-versa. Similarly, Siegal et al. [165] found that men were more likely to donate to a robot with a feminine voice, which was seen as more credible, trustworthy, and engaging; for women, there was no difference. Behrens et al. [13] found that the masculine voice for Nao led to participants being more willing to ask "him" for help. In contrast, Evans and Kortum [49] found that disclosures were not affected by voice gender. Donahue and Scheutz [43] found that women who experienced an affective voice or physical version of the agent in a cooperative situation switched attention more than men. Lubold et al. [106] found that women felt greater rapport with and persisted in helping a robot more than men, regardless of the robot's perceived gender. Although perceptions varied based on perceived gender, Mullennix et al. [126] found that persuasiveness did not. Nass et al. [135] found that a feminine voice elicited more disclosures than a masculine one, regardless of participant gender. Yu et al. [195] found that masculine voices led to longer disclosures, while feminine voices reduced the question skip rate. Stern et al. [169] found that men perceived humanoid voices as softer, and women rated them as more convincing.

In summary, stereotypes and biases appear to have been at play. Matching gender of voice and participant was best, in general. Yet, methodological and conceptual issues related to gender categories appear to have affected results. Even so, over time, a preference for feminine, extraverted voices has emerged.

3.7.2 Anthropomorphism, Humanlikeness, and Natural vs. Synthetic Voices. Human voices were rated more likeable than synthetic voices overall [50, 63, 130, 170, 182, 190, 191], but there was some nuance. Feelings of trust were greater for a human voice in Reference [190], and if participants were not familiar with robots, social presence and attraction rates were also higher. Eyssel et al. [51] found that the effect was increased when the gender of the agent and participant matched. Chérif and Lemoine [25] found that human voices elicited greater social presence compared to synthetic voices in a website, but there was no effect on perceived trustworthiness. Chiou et al. [27] found that the human voice was the most trusted and engaging, and human as well

as high-quality synthetic voices were most credible. Craig and Schroeder [32] found that human and modern TTSs were superior to older TTSs in terms of perceived credibility and engagement. In contrast, Davis, Vincent, and Park [42] found that weak-prosodic human voices were superior to a modern TTS in terms of credibility and engagement, while a strongly prosodic human voice was less engaging, perhaps because it was perceived as exaggerated. Qiu and Benbasat [148] found that a human voice elicited greater feelings of social presence, trust, and enjoyment. McGinn and Torre [114] found that humanoid voices were preferred for all robots except G5. Atkinson et al. [8] also found that human voices were rated more positively. Mullennix et al. [126] found that feminine synthetic voices were less pleasing and persuasive than their human counterpoints. In Tsiourti et al. [183], mismatching the affective vocal expression of a robot and its physical reaction and the socio-emotional context lowered its likeability and believability as human-like. In a comparison study, Baird et al. [10] found that different humanoid voices were rated at different levels of humanlikeness; the German male voice was rated the least humanlike. Cambre et al. [20] similarly discovered that while non-TTS voices were preferred, the relative humanlikeness of the voice could not be used to determine the best voice for long-form content. They concluded that "the variation…between how voices ranked…underscores that no single metric is sufficient for evaluating long-form speech" [9, 20].

Matching rather than mixing factors of voice and body appears to matter. McGinn and Torre [114] found that participants were only able to match the voice of one robot, the PR2, to its body. Sarigul et al. [158] found that people were quicker to assign a robot voice to a robot image, rather than a human voice to a robot image. Gong and Lai [62] found that mixing a human voice with a TTS at the same time led to poorer performance, even though people thought that they had performed better and found that version of the system easier to use. In a novel approach to exploring perceived embodiment, Mara et al. [107] had participants draw the body based on the voice, finding that humanlike voices were given ears, eyes, and noses, and synthetic voices were given wheels.

Humanlike and highly anthropomorphic voices sometimes had behavioral effects. Walters et al. [186] found that a robot's humanoid voice drew people in closer compared to a synthetic one, and in fact drew people closer than a real human did. Atkinson et al. [8] found that in lab and in the field student performance improved with a human voice, although understanding and perceived difficulty were not affected. In contrast, Stern et al. [169] found there was no effect on persuasiveness of humanoid or synthetic speech. Similarly, Abdulrahman et al. [2] found that human and synthetic voices were equally good at reducing feelings of stress. In their respective studies, Davis, Vincent, and Park [42] and Chiou et al. [27] found that there was no difference for learning measures. Sims et al. [166] found that when viewing a robot needing assistance, people were more likely to give commands when it had a synthetic voice rather than a human voice, suggesting that the voice could mediate whether people treat a robot as a capable human or a machine that needs direction. Other work indicates that agent having a visible body was key. Nass et al. [135] found that a synthetic voice performed worse than the text-based version at eliciting disclosures from participants. Similarly, Rosenberg-Kima et al. [153] found that a visual avatar was more effective at influencing young female students' views of engineering than a humanoid voice alone. Craig and Schroeder [32] found no differences between human and modern TTSs in terms of learning outcomes, but both were superior to older TTSs.

In summary, humanoid voices are preferred and more effective than synthetic voices. There is also some evidence that modern synthetic voices are reaching the level of human voices, but more work is needed.

3.7.3 Vocal Fillers. The use of vocal fillers appeared to be influential at times. Wigdor et al. [189] found that vocal fillers increased children's perceptions of a robot's humanlikeness,

responsiveness, and agency. They also found that children's heart rates increased when vocal fillers were used, a possible measure of engagement. Torrey et al. [180] found that people responded the same to humans and robots when robots used hedges and discourse markers like vocal fillers. It also increased perceptions of politeness, but only for the robot. Ohshima et al. [140] found that vocal fillers influenced speaking obligation, reduced feelings of awkwardness, and increased perceptions of sincerity. Additionally, less socially skilled people perceived greater levels of sincerity when fillers were present. Shiwa et al. [164] found that vocal fillers were preferred for a robot. Ohta et al. [141] found that pauses and silences added at natural breaks within sentences improved comprehension of the information presented by an agent as well as the naturalness of the agent's voice. In contrast, Jeong et al. [84] found that likeability when vocal fillers were used depended on the context, specifically better for social situations rather than in a service context. In general, it seems that vocal fillers improve experience with agents.

3.7.4 Affect and Emotion. Findings suggest that expectations about human vocal affect apply to agents as well. Chita-Tegmark et al. [28] found that participants were able to rate the emotional intelligence of vocal robots with the same accuracy as when rating humans. Yilmazyildiz et al. [193] found that this depended on the quality of the TTS. Age may also be a factor; Read and Belpaeme [149] found that while children perceived emotion in non-linguistic utterances, there was little consensus on what emotion was perceived. Additionally, pitch played no role in affect assignment. James et al. [82] found that empathetic voices led to considerations of the agent as empathetic and was preferred. Niculescu et al. [137] found that affect was more appealing. Yilmazydiz et al. [193] found the highest scores when voice affect and facial expression matched. Creed and Beale [33] found that audio expressions were harder to rate than visual ones for a conversational therapist agent. Expressive voices were perceived as more effective communicators in a study by Hennig and Chellali [74], but neutral voices were more socially relatable. Nass et al. [130] found that voice affect needs to match the emotional tone of the content, unless credibility is important, in which case a mismatch is best. An exception is the work of Mendelson and Aylett [116], which showed a contradiction between voice-only and avatar conditions. The correlations between irritated voices and feelings of positivity and calmness disappeared for the voice-only condition. Relatedly, Komatsu et al. [90] found that people were better at assigning positive and negative affect to sounds attributed to a computer rather than the AIBO dog robot and a Lego Mindstorms robot.

Some behavioral effects were found. Donahue and Scheutz [43] found that affect increased the amount of participant utterances in a cooperative task, which continued after affect was removed. Niculescu et al. [137] found that the use of empathy increased participants' feelings of confidence. Bhagya et al. [14] found a relationship between proxemics and affect similar to human patterns, e.g., a happy voice drew people in. Nass et al. [132] found that when a smart car voice matched the participant's mood, the participant had fewer accidents and paid more attention to the road, and spoke more with the car. Shi et al. [162] found that a waveform embodiment of voice produced fewer facial expressions in the participant, although there was no difference in cognitive engagement compared to avatar and text versions.

In summary, affect is a key factor that can impact perceptions of the agent and human behavior. Matching the content or speech appears to be important. The humanlikeness of the agent's body may also be a factor in perceiving voice affect and its relevance, but more work is needed on a variety of embodiments.

3.7.5 Accent and Dialect. People seem to be able to accurately perceive accents and dialects through agent voice, and this can have effects on perceptions of the agent in general and the participant's behavior. In general, this aligns with human models and stereotypes. Krenn et al. [93]

found that an agent with the Austrian standard accent was perceived as more educated, trustworthy, competent, polite, and serious compared to one with the dialectal accent, which was perceived as more natural, emotional, relaxed, open-minded, humorous, and aggressive. Moreover, the colloquial accent fell somewhere in between on ratings. Similarly, Dahlbäck et al. [38] found that matched accents led to more disclosures of a socially undesirable nature as well as perceptions of sociality in the agent. This was confirmed in a later study [39], where they also found that participants rated the information quality higher when accents matched. Sandygulova and O'Hare [156] found that Irish children preferred the voice agent with a UK accent. Tamagawa et al. [174] found that US accents were perceived as more robotic by New Zealanders, who preferred UK accents and rated New Zealand accents highest in terms of perceived robot ability and higher affect. Stereotypically trustworthy-sounding accents in the UK were rated as more trustworthy in an agent, and greater disappointment resulted for these voices compared to those with stereotypically untrustworthy-sounding accents when the trust was broken in a game [179]. Yilmazyildiz et al. [193] found that the use of gibberish did not obscure the origin of the language. In two studies, Andrist et al. [6] showed that the amount of speech was linked to preference for the standard or local dialect: when the robot had more to say, participants preferred the local dialect, and vice versa with less speech.

3.7.6 Paralinguistic Cues, Prosody, and Speech Style. Paralanguage can provide important cues that reinforce or disrupt other voice characteristics. Ghazali et al. [60] found that facial expressions resulted in higher social agency scores compared to voice alone. Lazzeri et al. [98] found that, regardless of virtual avatar or robot, "muted" facial expressions without vocalizations were confusing. Read and Belpaeme [150] found that nonverbal utterances conveyed classifiable emotion. The use of vocal fillers increased social presence in a study by Goble and Edwards [61]. However, Cao et al. [23] found that vocal fillers were perceived as a sign of lack of fluency rather than human-like hesitation. Read and Belpaeme [150] found a "magnet effect," where small differences in expression of nonverbal utterances were unconsciously ignored by participants in favor of classifying the affective quality of the utterances in line with basic emotions.

Pitch was a key paralinguistic feature. Lubold et al. [106] found that social presence was rated higher when the body and voice matched in terms of pitch. Niculescu et al. [137] found that high pitch in a robot voice was perceived as appealing, extroverted, socially skillful, and similar to the user. Chidambaram et al. [26] found that modulating the pitch of the robot's voice in tandem with behavioral cues, such as gestures, resulted in increased compliance with its suggestions. Elkins and Derrick [46] found that higher pitched voices lowered perceptions of trust in the agent, but only at the beginning; the effect flattened out over time. Shibata et al. [163] found that participants did not change their own pitch when speaking with a higher or lower pitched robot, although they liked the higher-pitched robot more. Niculescu et al. [138] found that higher-pitched robots were better liked and seen as extraverted, though there was no effect on perceived usefulness.

The style of speech as expressed through voice can also influence perceptions of the agent. Bracken et al. [15] found that praise from an agent led to higher niceness scores, but a text-only version led to higher perceived intelligence. Yarosh et al. [191] found no effect of personalizing (i.e., referring to the user by name) for children. Yilmazydiz et al. [193] found that using gibberish, rather than semantic content, is acceptable. Similarly, Zaga et al. [196] found that intent expressed through gibberish was more recognizable to children than non-linguistic utterances (85% versus 20%). Children also complied less to non-linguistic utterances, while there was no difference between gibberish and a soundless condition.

In summary, higher-pitched voices appear best, but matching voice to other features, such as affective content, as well as personalities, and avoiding too highly pitched voices, is important.

3.7.7 Personality. Matching personality to the participant as well as features of the voice and agent role were best. Braun et al. [16] found that matching participant and driving assistant personalities was more liked and trusted compared to mismatched personalities. Hoegen et al. [77] found that participants with "high consideration" conversation styles perceived a matching agent to be more trustworthy, while those with "high involvement" conversation styles were indifferent. Lee and Nass [100] found that social presence was higher in introvert/extrovert-matched voice agents, though extroverted participants gave higher social presence ratings regardless. Niculescu et al. [138] also found that introverted participants liked a robot with a lower-pitched voice better than one with a higher pitch. Similarly, Tay et al. [176] found that personality matches led to better outcomes than gender matches. In contrast, Creed and Beale [33] found that a happy conversational therapist agent was most engaging, and mismatching personalities did not cause confusion or increase cognitive load. Niculescu et al. [137] found that the use of humor was more appealing and increased perceptions of the robot's social skills, while not affecting trustworthiness. In contrast, Evans and Kortum [49] found that voice personality did not affect rate of disclosures. Since there appear to be no ill effects, matching personalities is safest.

3.7.8 Morphology and Medium. The body matters for the voice of the agent. Yilmazyildiz et al. [193] found that matching the "size" implied by pitch and the actual size of the body mattered, such as by matching a low pitched voiced with a large body. Xu [190] found that the use of gestures increased feelings of intimacy and interest. Crowell et al. [35] found that the voice having a body increased perceptions of its friendliness, while Shamekhi et al. [161] found the same for social presence. Stern et al. [170] found that voice source mattered, with human-as-source of a humanoid voice rated more highly than one with a synthetic voice; however, there was no difference for computer-as-source. Similarly, Trovato et al. [182] found that non-humanoid voices were judged inappropriate for non-humanoid robots. However, Gong and Nass [63] found that matching human-human and humanoid-humanoid in voice and face (through an avatar) mattered, especially for women. Zanbaka et al. [197] found that virtual human and non-human characters given a human voice were rated more bold than real humans and their real voices. Lee et al. [103] found that a conversational robot elicited greater impressions of intimacy, competence, and warmth than a voice agent. In contrast, Kim, Goh, and Jun [87] found that voice was superior to text in terms of helpfulness and self-validation for a banking chatbot. Tsiourti et al. [183] found that mismatching affective expressions in body and voice led participants to perceive the robot as less intelligent or unable to express emotions properly.

Behavioral impacts were also found when the voice had certain bodies. Donahue, and Scheutz [43] found that the presence of a robot body increased the amount of utterances in a cooperative task. Bracken et al. [15] found that the text condition led to increases in perceived ability and intrinsic motivation compared to voice alone. Moreno et al. [122] found that a voice agent increased learning transfer, retention, and interest in learning materials for middle schoolers and undergraduate students compared to text alone. Evans and Kortum [49] found that people disclosed more to a voice agent than in a web-based survey. Vannucci et al. [184] found that the voice-only condition resulted in quicker reactions to the agent's calls to action. Sometimes both positive and negative results were found. Shamekhi et al. [161] found that group interaction was improved with the use of a body, but group decision making was not. Similarly, Rosenthal-von der Pütten, Strasmann and Kramer [154] found that varying the medium—robot, avatar, voice agent—and naturalness of the voice—recorded or synthetic—had no effects on performance measures in a second language learning context. Tsiourti et al. [183] found that mismatching voice and body led to confusion and lower accuracy scores for evaluating the robot's emotion.

In summary, adding voice to body conscientiously can be effective, or at least have no ill effects. Still, the variety of bodies explored for agent voice is limited, a point we will explore below.

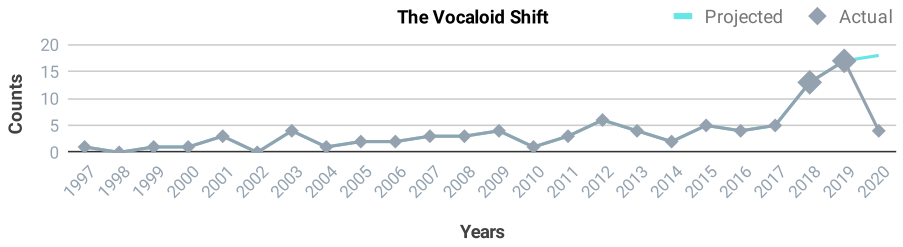

Fig. 4. Recent interest in agent voice along with a surge in commercial voice assistants indicates a "vocaloid shift" starting in 2018 and projected to continue from 2020 onwards.

4 SYNTHESIS AND DISCUSSION

In 89 papers over 24 years, interest in voice has waxed and waned, remaining an important but understudied factor in HAI. Following the emergence of commercial voice assistants, like Siri and Alexa, there has been a recent sharp uptick in published work. We call this the "vocaloid shift" (Figure 4) and use it as a tool for contextualizing our findings over time. We begin by presenting a classification framework for vHAI to guide discussion of our synthesis and findings. We then present and discuss our synthesis of results across sub-questions and quantitative and qualitative data [75], illuminating foci, patterns of findings, and limitations. We structure our findings in terms of theory, method, technology, design, and knowledge. We then consider the link between voice and body. We end with a research agenda for advancing vHAI work.

4.1 A Classification Framework for vHAI

The surveyed papers provide a rich tapestry of the key factors in vHAI. We begin by distilling these factors into a high-level classification framework for vHAI experiences, illustrated in Figure 5.

The agent and the person are situated within a context, comprised of activities as well as the space in which those activities occur. The ability of the agent to interact here renders it embodied [11]. The person perceives the agent through its voice and the body housing its voice. The agent's body is not necessarily humanoid or physical. The body can be further broken down into medium (i.e., the enabling technology, such as a device) and morphology (i.e., form factor, appearance, shape, etc.). This distinction is in line with Miller and Feil-Seifer's [118] social robotics framework, wherein they hint at its applicability beyond robots (i.e., virtual agents). As the surveyed literature indicates, the voice and body may not share the same medium (e.g., the voice is coming from a speaker rather from the computer screen). Also, the person may not perceive the voice and body as originating in the same medium. Voice and body morphologies may also differ.

The voice conveys information that is comprised of vocalics (non-linguistic information with semantic meaning) and speech (linguistic content). Vocalics features have psychosocial and sociolinguistic characteristics. As such, perception of the agent's voice and body involves the elicitation and processing of social information between the person and the agent; in short, social embodiment [11]. According to the surveyed literature, the most common psychosocial characteristics are anthropomorphism (humanlikeness), gender, age, personality, and affect, whereas the most common sociolinguistic characteristics are paralanguage (i.e., pitch, prosody, rhythm, etc.), accent and dialect, style of speech (i.e., manner of speaking irrespective of the content or intended message), and vocal fillers (such as pauses and "umms").

Now that we have described the classification framework, we turn to the synthesized findings.

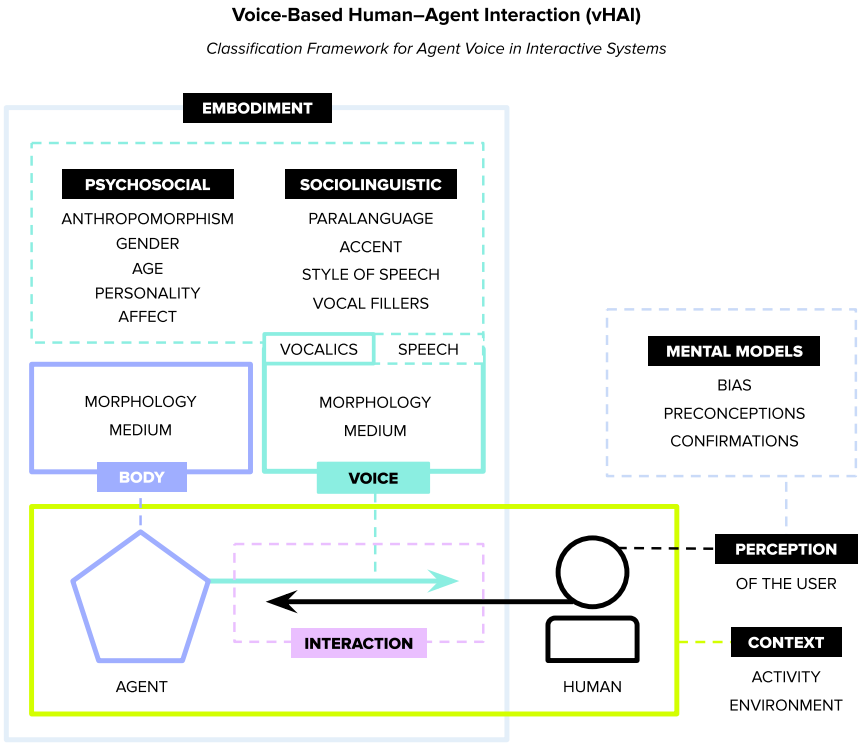

Fig. 5. Classification framework for voice-based human–agent interactions (vHAI).

4.2 Theoretical

The use of theoretical frameworks and conceptual models in the research to date has been limited, with most papers not reporting or referencing theory. Yet one theory appears to be gaining consensus: the Computers Are Social Actors theory [136]. As an overarching theory, its premise is that people apply human social heuristics to computers without realizing it [131], effectively treating computers as people, at least to some extent. This idea is grounded in *anthropocentrism*, or the tendency for us to unconsciously read everything we perceive as acting with intelligent agency as human-like. This can have both positive and negative repercussions. Agents designed this way—using "anthropomorphic metaphors" based on human models—can improve understanding, fit, and user experience, but this comes with the risk of people over-ascribing intelligence, intentions, and abilities to these creations [108]. Notably, our tendency to anthropomorphize occurs even without such anthropomorphic metaphors being (consciously) designed for. Over the years, without a specific focus on voice or speech, research on the Computers Are Social Actors theory has revealed how several human–human phenomena emerge in human–computer interactions, including gender stereotyping (based on assumptions and/or cues in the design), reciprocity (based on feelings of obligation applied to the computer), and personification (based on personality cues designed into the computer). Whether or not the papers surveyed here referenced this theory, most did go some way to support this model of HAI in the context of voice-based experiences.

They also contribute to it. Taking a longitudinal perspective on the survey data, there is a recent preference for high-pitched feminine voices. Yet, older studies showed a preference for deeper-pitched masculine voices. A defining moment in voice-based HAI in the intervening two decades is, of course, the emergence of voice assistants, especially Apple's Siri (2011), Amazon Alexa (2014),

and Google Assistant (2016), all of which feature a feminine voice. This shift, part of the vocaloid shift, may be attributed to the proliferation of these consumer devices. This reflects the notion that gender is socially perceived [105]. For the Computers Are Social Actors theory, the implication is that higher-order social phenomena that are dynamic, contextual, and may change over time at a societal level are at play. Findings from 10 or 20 years ago may be too limited or no longer apply. Indeed, most of the work involved young, college-educated people in Western societies, the majority of which were probably white, given these demographics. The Computers Are Social Actors theory seems to represent the findings of voice-based HAI well, but it may need to be expanded and older research conducted again with different groups of people to become truly comprehensive.

Aside from this, most theories and conceptual models cited relate to social phenomena derived from research on people. Given that voice is considered a basic and natural mode of human communication, it is not unexpected that these human models would apply to computer voices. We should explore to what extent human models of social interaction and communication apply in vHAI. Indeed, those working in human–machine communication have recently suggested as much [71]. To do this, we need to ground our research in theory. Most likely, this will need to be done in a multi- or supradisciplinary way: through a mix of theories from specific technical fields with those from social disciplines to cover both disciplinary framings of HAI and also move beyond them. The survey results suggest that this is already being done; however, most theories and models are used in a one-off fashion, so consensus is difficult to achieve. We can safely use the Computers Are Social Actors theory as a starting point for our questions and hypotheses: determining what particular feature of voice (or voice in combination with other modalities or attributes of the agent) we wish to focus on, and then manipulating that feature based on established theories and models.

4.3 Methodological

Findings reveal several methodological trends and shortcomings. One issue is a common and long-standing problem in human subjects research: a limited participant pool and overreliance on young students from the West [58]. Some work has started to consider voice for elders and children, as well as in non-Western cultures, but work in these areas is nascent. There are also many other demographics and combinations of factors to consider, as calls for intersectionality in human–computer interaction have recently pointed out [95, 159]. Non-humans, such as companion animals [73], are also viable subjects, especially where voice is concerned. Dogs, for instance, can be trained to understand some speech but cannot speak themselves. Yet they can convey semantic meaning through non-linguistic utterances and other forms of non-speech-based vocalizations that can be understood by people and potentially computers [121]. Indeed, there is great need and opportunity to incorporate different user groups in studies of voice-based HAI.

While most studies used comparisons, many did not use experimental controls or were limited in the number and kind of comparisons (see Table 1 for a quick review). This makes it difficult to draw conclusions about the relative effects of voice in HAI versus other "vocal" agents, particularly humans, but also "non-vocal" options, such as written text (e.g., chatbots). Also, multimodal voices, e.g., text + verbal, could be explored to better understand the differences among voice and speech factors, as well as explore novel configurations of output modalities. For instance, some video games feature non-linguistic but affective vocalizations, including short exclamations to augment speech (e.g., 2019's Pokémon Sword/Shield). Others use long patterns of gibberish presented at the same time as linguistic speech using text (e.g., 2004's Katamari Damacy). This may have practical current or near-future implications, such as in the design of text-to-speech interfaces for allowing people with speech disabilities to freely express themselves in a variety of ways at various times and in various contexts.

A related methodological issue is the use of Wizard of Oz, found in one-third of studies. This method involves a human operator acting in place of computer functionality, unbeknownst to the end-user. It is a classic human–computer interaction and human–robot interaction evaluation method [37]. It is often used during the early stages of design, when there are technical limitations, scoping issues, and delays. It is also used as a quick way of evaluating ideas during rapid prototyping. In human–robot interaction, it allows for the exploration of future forms of robots that may not be possible now, or may require a huge investment of time, effort, resources to make. Indeed, it is a powerful method and one that should be kept in the vHAI toolkit. Nevertheless, it has limitations. Findings from studies that use Wizard of Oz with commercially available technologies, such as the Nao, do not reflect the technologies' actual functionality, usability, or user experience. Taking the Nao as an example, only one study used the Nao TTS, while the rest used other TTS's or Wizard of Oz. It is thus difficult to draw conclusions about the actual effects of Nao's "natural" voice, let alone the importance of other intersecting factors, like Nao's robot or virtual bodies.

While many studies involved a realistic application, many of these were not evaluated under "truly" realistic conditions. Most were conducted in a controlled lab situation. It is thus hard to conclude how the effects reported would play out in the wild. This is important, because vHAI systems are already in the wild and are likely to stay there. More experimental ethnographic field studies like the one by Braun et al. [16] will need to be conducted for us to get a better sense of real world impacts. This will involve a careful consideration of ethics and a special setup wherein there is a delineation of the real environment but also potentially new instruments and facilitative technologies to enact the research procedures. Confounding effects will need to be accounted for; as Braun et al. [16] discovered, the novelty effect and novices can still impact field work. For voice, other sounds in the environment will need to be considered—or discovered through the research, such as the levels and kinds of noise that impact physical perception of voice.

The measures evaluated were diverse, and so was their measurement, a problem that needs to be addressed Many original instruments were used, which can increase the utility of the research design to address specific and unique questions [113]. But original instruments also limit transferability and generalizability. Moreover, many measures appear to address the same underlying factor; for instance, rapport in Lubold et al. [106] and psychological closeness in Eyssel et al. [50, 51]. Additionally, one-fifth of the surveyed papers did not report Cronbach's alpha or a similar measure of internal reliability, thus limiting understanding of the strength of these original instruments and efforts towards construct validity [113]. This was the case even for studies that used the most common technologies; of the studies with Nao, for instance, only two reported Cronbach's alpha, even when similar measures were used. Additionally, even when the Nao was used, the voice was often not the Nao TTS, limiting generalizability even if Cronbach's alpha was reported for the same measures across studies. Finally, the reliance on self-reports makes it difficult to come to objective conclusions.

A range of characteristics for agent voice were evaluated, with projects often involving the design of different interfaces and embodiments to determine the design of best fit and impact on people. Yet, less than half conducted a pre-test or manipulation check, i.e., whether or not the design features and manipulations were perceived in the intended way. For instance, the robot in Crowell et al. [35] was presented as gender-neutral. Not ensuring that user perceptions match designer and researcher expectations, yet including the given characteristic as a key factor, is dangerous. In the case of gender, consensus is not guaranteed [114], and there is a tendency to perceive the "default" as male [12]. Additionally, the measures related to voice characteristics need to be carefully considered to allow for diverse and emergent options. For instance, in Sandygulova and O'Hare [157], children were only able to respond with male or female gender categories, leaving

out the possibility of a "mechanical" or robotic gender [167], no gender, uncertain gender, multiple genders, and so on. This raises a serious question of reliability given that people will tend to answer a question even when they have no answer [94]. Others did not consider the impact of gender, e.g., References [33, 74, 137, 190]. In such cases, design decisions maintained the status quo. For example, a therapist agent was gendered female, because at the time of the study and in that cultural context most therapists were women [33]. Others used a male voice based on the assumption that robots are read as male [190]. Lack of awareness, lack of interest, or understandable limitations in technical, research, or budget scope may have been at play. However, the findings from this survey strongly suggest that major social factors like these need to be accounted for as influencing variables, at the least, but also considered as opportunities for design innovation.

4.4 Technological

Most work was limited to computer-based voices and robots. We now need to explore the varieties of smart objects, devices, and spaces sphere, e.g., smart speakers, vehicles, phones. Some of this technology, e.g., the iPhone for Siri, the Echo smart speaker for Alexa, and so on, have been around for a while and critiqued [69], often colloquially [53], for their social depictions of gender. A lot of custom technology was also used, making it difficult to compare studies. This is important, because the surveyed findings suggest that "body" may not matter, yet it is unclear why or when. For instance, voices associated with virtual characters were sometimes preferred over people (e.g., Reference [197]) and in other cases not (e.g., Reference [126]), even within the same context of use. Wizard of Oz, in particular, represents an ideal form of the technology that is not currently possible. It is useful for deciding what trajectories to pursue, but it does not represent ecologically valid knowledge.

The Nao suite, comprised of a robot, a virtual (screen-based) avatar, and TTS, is a one-stop morphology comparison research instrument. It was also the most common choice for robot studies. Yet, only one study [154] used a combination of options. This may be due to technical limitations (indeed, many Nao studies use Wizard of Oz for this reason). It may also be due to nascent work on the question of morphology [118], which we expect will continue to gain attention. Additionally, the surveyed work indicates that the platform may be applicable to different generational groups: 40% included adults, 30% included students, and 40% included children. Yet, not all voice characteristics were explored, in particular, personality and anthropomorphism. Technically, explorations of these factors through the Nao TTS and comparison to human voices are possible.

TTS technologies continue to improve over time. Advances in technical methods using machine learning, especially neural networks and deep learning [7], are pushing the quality of TTSs forward at a rapid pace. As the surveyed work and other surveys of technical work (e.g., Reference [36]) suggest, the quality of the TTS is important for voice. Yet, a lot of work (40%) that involved comparisons of computer voice was carried out more than 10 years ago, when TTS quality was lower. Previous research may need to be replicated with newer TTSs, especially given recent work on higher quality TTSs that still suggest an effect of quality [193].

From a theoretical perspective, the most referenced theory is the Computers Are Social Actors theory. Yet, the technology distribution indicates a reliance on TTS, with robots and virtual agents making up only 54% of work. This may be due to the slightly different historical trajectories across fields of study. Work on the Computers Are Social Actors theory is based in psychology and human–computer interaction, rather than human–robot interaction and other areas. Yet, as our findings show, this theory is applicable to and valuable for studies of robot and virtual character voice (and likely other morphologies, such smart "things"). Going forward, a unification of work on the Computers Are Social Actors theory with work in human–robot interaction and other disciplines will be fruitful, if not necessary.

4.5 Empirical

Despite the limitations and criticisms noted above, we can draw out several higher-order findings based on consensus across the surveyed work. We will discuss the state of the art and empirical knowledge gained overall, considering the findings at a high level and the intersections of factors uncovered by Q10.

4.5.1 Meta-Findings and Meta-Criticisms. The major takeaway across studies and categories of findings is that human (anthropomorphic or "natural"), happy, empathetic voices with a high (rather than low) pitch seem best overall, regardless of the kind of agent, enabling technology, user group, and application. In general, preference for gender changed over time, with older studies showing a preference for masculine voices, and recent studies showing a preference for feminine voices. There are some exceptions, especially [154], which found no differences between a robot and virtual character tutor. Future work will need to tease out why.

While the focus in terms of voice characteristics has been gender, anthropomorphism (i.e., "natural" or humanoid versus synthetic), and affective qualities, there is some evidence to suggest that explorations of voice in HAI are beginning to expand. Recent work on semantic and paralinguistic factors, especially accents (e.g., Reference [179]), gibberish (e.g., Reference [193]), and vocal fillers (e.g., Reference [61]) are indicating a maturation of the design thinking involved in the study of voice in HAI experiences. However, a drawback of much of the work surveyed is a lack of detail in the design of the agents; it is possible that such features have already been included in the designs and evaluated to a degree, but these details were not reported. Additionally, characteristics may be overlooked or assumptions may be made without double-checking. Grounding these decisions in social theories with exploratory design and verification research using human-centered methods, as well as pre-tests and manipulation checks, would increase the design and research rigor.

Bias may also have been a limiting factor in a significant portion of the work. As the survey shows, people perceive and react to voice characteristics in different ways, often premised on surface-level judgments and stereotypes in line with the Computers Are Social Actors theory. But bias may also be present in the design of the technologies (or their selection), the selection of participants (especially with respect to age, which was all but absent), and the design of the research (especially the selection of measures and instruments). We can take this seriously as we actively develop a new chapter in vHAI for the "social good" [199]: We can confirm or disrupt the status quo with respect to gender, age, cultural background, social role, job, and so on in the way we design and research vHAI. To do this, we can draw on reflexive [151] and intersectional [159] frameworks known and practiced with human–computer interaction and human–robot interaction.

4.5.2 Psychosocial Factors: Gender, Anthropomorphism, Personality, and Affect. People tend to perceive gender in voice-based HAI, whether voice gender was intended by the designers or not. Higher-pitched voices tend to be read as feminine and lower-pitched voices tend to be read as masculine, indicating an intersection between the psychosocial factor of gender and pitch as an aspect of prosody, a sociolinguistic factor. These findings must be carefully considered in light of the methodological limitations particularly how evaluations of gender have been conducted (e.g., limited categories) and the inclination for most people to read masculine when no gender cues are present [12]. Whether or not gender matters appears to depend on when the study was conducted, in line with the vocaloid shift, as well as other factors that we could not determine in this survey. For example, early research showed gender stereotyping and preferences based on human social norms (e.g., Reference [134]), while recent research suggests an influence of the feminine-gendered voice assistants (e.g., Reference [24]) as well as the emergence of other influential factors, like personality [176], and changing expectations about gender (e.g., Reference [44]). The influence of

commercial voice assistants is further bolstered by the results of research on very young children (e.g., Reference [157]).

Voice anthropomorphism, or humanlikeness, was a major factor, and appears to supersede gender in terms of perceptions and behavioral influence (e.g., References [9, 126, 186]). This is good tidings; even though most societies rely on a gender binary, i.e., women and men [129], notions of gender are beginning to expand. A "human" voice gendered "neutral" or "nonbinary" or even "robotic" has yet to be developed or evaluated and may not work in practice at a general level until the majority of society moves beyond the gender binary, but is a real option that should be considered for the future of voice in HAI.

Human voices were best overall. However, there is some evidence that modern synthetic human voices are reaching the same performance as live or recorded human voices. This can be explained by improvements in TTS systems and the public's exposure to TTS systems in daily life, in line with the vocaloid shift.

In general, matching people and agent personalities appears to be ideal. For example, introverted agents for introverts [100]. Personality stereotypes based on gender were also at play, such as a preference for informative masculine voices and social feminine voices [102]. An open question, from research based on the Big-5 model, is whether gender matters and under what circumstances. For instance, Lee and Nass [100] found that matching extroverted agents to extroverted people supersedes any effect of gender. However, Chang et al. [24] found that feminine extroverted voices are preferred. Given the fifteen years that separate these studies, this may be another example of the influence of modern voice assistants like Siri.

Findings show that people can perceive emotion through voice, if certain conditions are met. The inclusion of voice affect to visual expressions of affect appears key (e.g., References [74, 98]), with some work suggesting that voice alone is more difficult to interpret (e.g., Reference [33]). For instance, if there is a mismatch between bodily affective cues and vocal affect, then people are confused (e.g., Reference [183]). Indeed, non-humanoid agent bodies appear to increase the difficulty of perceiving affect, even when voice affect exists. Perfection is not necessary, however, due to reports of a "magnet effect," where people tend to gravitate to the most likely affective expression (e.g., Reference [150]). To be safe, a manipulation check should be conducted with the target user groups. More radically, it appears that speech content and linguistic expressions of emotion may not be needed at all, according to the findings on non-linguistic utterances and gibberish (e.g., References [150, 193]). The extent to which this is true needs to be investigated.

4.5.3 Sociolinguistic Factors: Accents, Paralanguage, and Style of Speech. People can perceive accents and dialects in voice-based HAI, even in gibberish [193]. This has effects that match human models and stereotypes, such as standard dialects being preferred for instruction contexts. Paralanguage, especially pitch and loudness, appears to play an important role in perceptions of gender, emotion (e.g., higher pitch and happiness), personality (e.g., high pitch and extroversion), and form factor (e.g., higher pitch for smaller robots). Paralanguage can thus be manipulated to generate different kinds of user experiences and effects. As discussed, voice and speech are closely related and sometimes difficult to disentangle. One example that emerged across studies was the combination of paralanguage and speech content as higher-level behavior in social interaction. Cases include praise (e.g., Reference [15]), empathy (e.g., References [82, 137]), and humor (e.g., Reference [137]). The use of such voice-expressed behaviors appears to especially impact people's perception of the agent's social ability and social presence (e.g., References [106, 137]). An accent or dialect expressed through voice can indicate identity as well as the relationship with the recipients and the agent's personality and emotion. Given that the "building blocks" of such behaviors can be designed into voice, the extent to which these hold true, their impacts, and other varieties

of features can be explored in future research. Importantly, the quality of the voice (e.g., a high-quality TTS) is important to ensure that the intended impression is conveyed [193]. This can be confirmed through a manipulation check with representative members of the target audience. It is unclear from the surveyed papers whether or not speech content is necessary; for instance, praise may be conveyed through a combination of paralinguistics, agent facial expressions, and perhaps even external stimuli, such as happy sounds, lights, and exploding stars, such as in cartoons and video games. It may also be that a multimodal combination of voice and body is sufficient. All of these need to be explored with respect to voice in HAI.

4.5.4 Embodiment Factors: Morphology, Medium, and Situatedness. Voice should be considered with respect to matching the "body" (the morphology or form factor of the medium or source through which the voice is expressed or associated), task, context, and participant. For example, matching voice affect, body affect, and content, such as by pairing happy vocalizations and facial expressions with positive news, and harmonizing voice and, such as synthetic voices with robotic forms, appears ideal. However, one question that remains is whether to use human-like voices with non-humans. Some research surveyed (e.g., Reference [182]) suggests that participants will perceive a mismatch, even though the human-like voices tend to be preferred and more effective. Such findings may be a matter of certain aspects of the morphology. For example, in the case of Trovato et al. [182], the robot had a human form but it was not realistic, with blocky shapes and bright pink stripes over metal. Replicating the human form with great precision while falling just short of accurate representation can lead to the unwanted phenomenon of the Uncanny Valley [123]. Still, a variety of bodily forms are possible without eliciting negative results.

Comparative studies suggest that a lack of a recognizable body can have detrimental effects on social perception of the agent, including lower perceptions of friendliness [35], poorer social presence and group interaction [161], and lower persuasiveness scores [153]. Even when hypothesized benefits did not occur, as Shamekhi et al. [161] found with respect to group task performance, the addition of a body did no (measured) harm. Furthermore, it seems best to include other communication modalities in the mix, especially body-based nonverbal behaviors, such as gestures, facial expressions, in line with social models of human communication [56]. But some exceptions exist. For instance, the morphology and medium did not have an effect on feelings of trustworthiness in Torre et al. [179], where stereotypic vocal cues superseded other factors to the extent that when trust was broken during a manipulated research situation, participants reported an even greater sense of disappointment. In Stern et al. [170], the source of human-like or synthetic voices mattered when the source was human, but not when the source was a computer. Context seems key.

Given the limitations discussed, particularly around the use of different technologies as well as a reliance on original measures and instruments, it is difficult to generalize about morphology. Indeed, this work adds to the recent call on furthering our understanding of mental models on morphology, embodiment, and situatedness in HAI experiences [70, 118]. It is unclear what participants perceived to be the "body" of the voice in previous research, despite the body apparently being a key factor. For instance, is Siri's body the iPhone? The app on the iPhone? The rotating eclipse of rainbow lights in the app? The speaker on the phone (that is not visible)? Something else or a combination of these? If it does matter, then how and when? Indeed, as we discuss in the research agenda section, the gaps and incongruities in this "body" of work highlight areas of opportunity in design and research practice at the intersections of voice, morphology, and medium.

4.5.5 Contextual Factors: Environment and Activity. Human voices were superior in most contexts, and matching expectations, such as emotional expression to the context and speech content, was ideal. However, there were subtle differences across contexts with voice and body that may relate to the particularities of the context. In the case of Nao, for instance, vocal fillers

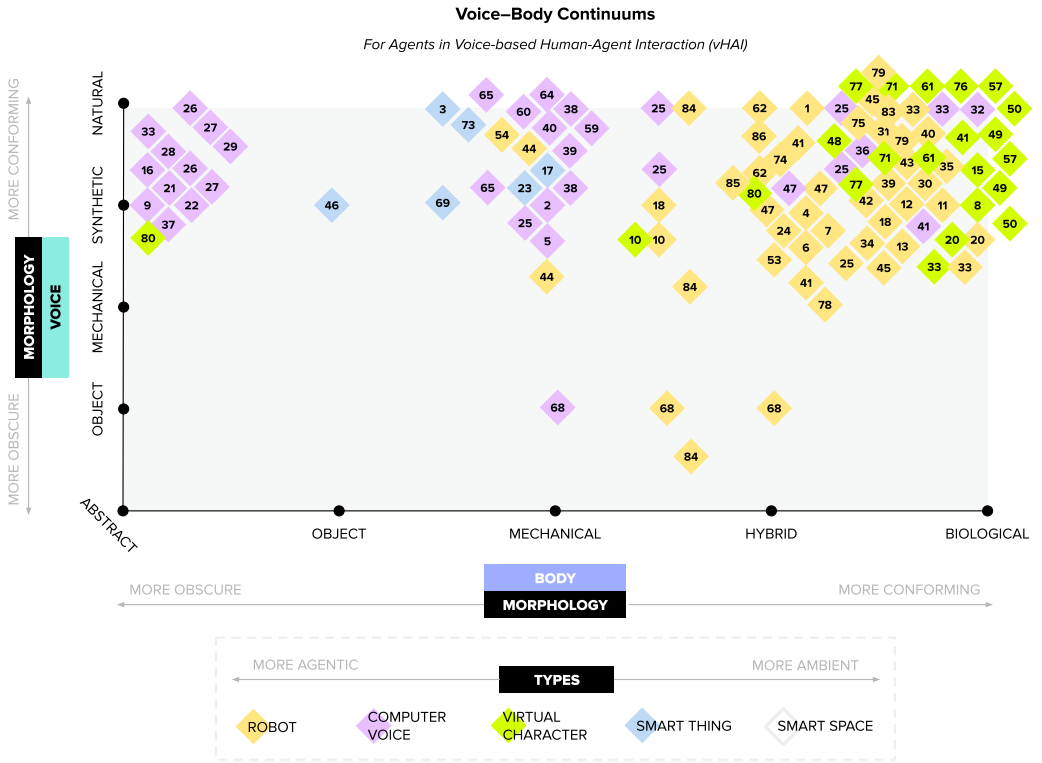

Fig. 6. Voice–body continuums in vHAI.

in conversation increased social presence [61], but for instruction, they were seen as disruptive, especially in terms of agent intelligence [23]. This may relate to the idea that conversation is dynamic and fluid, while instruction is structured and predetermined. This is important for deciding when to use vocal fillers: the activity as well as the social environment. Additionally, perceived gender emerged as a complicated factor: sometimes feminine voices were superior, sometimes masculine voices, sometimes voices not matching the gender of participants, and sometimes no effect. Given the dates of this work, the vocaloid shift could explain some of these results. We suggest first understanding typical perceptions about voice characteristics within the specific context at hand, and then designing the voice (and body) around those perceptions.

4.6 Extending the Framework with Voice–Body Continuums

The connection between voice and body can be expressed as a set of morphology continuums. In Figure 6, the body is the x-axis and its voice is the y-axis. These axes range from obscure to conforming to human expectations. Each are made up of actual and potential categories. All surveyed papers have been placed, unless there was not enough information to make a placement. Papers are color-coded by type of agent. Some papers have been placed multiple times, because multiple agent types were reported in the paper.

There are clear clusters per axis. In terms of voice morphology and type of agent, most are at the top, indicating a majority of human and humanoid voices. In terms of body morphology and type of agent, there is a clear cluster to the far left (i.e., abstract morphologies), in the middle (i.e., mechanical morphologies), and at the far right (i.e., humanoid and hybrid forms). The overlapping

clusters of agent types in the top right corner are expected: We are social animals that have evolved to interact with agents that resemble us. But due to the computer revolution and onset of the information age, particularly the widespread take-up of personal computers and smart devices, we are starting to see extensions into other areas of the continuums and their intersections. In particular, work on "smart things" and "smart spaces" that is emerging. As yet, many of these areas remain under-researched or uncharted. We discuss this in the research agenda section below.

The voice–body continuums reveal that many explorations of computer voice abstracted the morphology of the body such that the voice was "disembodied." A review of the findings presents a complex picture on the results of this. The presence or absence of a perceivable body does not seem to have a general effect. However, only three papers featured such a comparison (e.g., to a robot with a perceivable body). Of these, one found no difference [154]. The other found subtle differences that are difficult to draw general conclusions from, though the voice having a perceivable body was considered more friendly [35]. The third found that animated faces increased feelings of social presence compared to a "disembodied" voice [148]. Notably, much of this work (10 of 13, or 77%) was conducted a decade ago or earlier, where the focus was on comparing synthetic versus human voices. Indeed, this cluster appears to be another form of marker for the vocaloid shift, delimiting the arrival of voice assistants in our everyday technoscape.

As a classification tool, the voice–body continuums framework may be used to describe current work as well as identify future research questions. For instance, what kind of voice should an agent with abstract morphology have? It can also suggest congruencies and tensions at the intersections of the two continuums. For example, what kind of voice should objects have, given that we cannot refer to any natural experience of hearing or talking with objects? Perhaps "human/oid" is too limiting: What kind of voice should a "biological" object, such as a plant, have? A situation where there is a continuum misalignment may lead to tension between artificiality and intuitive modulation. In this way, it has both practical and exploratory use.

4.7 Research Agenda

Our findings show that the area of vHAI is maturing and evolving. We suggest several trajectories for future research, particularly ones that consider the strengths and weaknesses of the present body of work as well as the opportunities, gaps, and (in)congruities suggested by the survey results.

4.7.1 Extending Theory in Light of the Vocaloid Shift and the Voice–Body Continuums. Our findings provide further support for the Computers Are Social Actors theory: that, as a general rule, people seem to unthinkingly treat voice-based agents as people, to a certain extent. This is sometimes predictable (e.g., based on prevailing gender, accent, and cultural roles and stereotypes) and therefore designable. Human voices that are extroverted, higher pitched, and empathetic elicit the strongest results. This means that social cues and constructs, attitudes, and behaviors come into play during vHAI experiences. The effects are dynamic and contextual: they may vary or change over time, among age groups, across cultures, and so on. But due to the limited user groups and locations of study, more research is needed to confirm generalizability. Indeed, as Sutton and colleagues [21, 171] recently argued, we need to more deeply consider the sociophonetic factors of vHAI to diversify our knowledge as well as the voices available through—and understood by—the machine. Furthermore, as Guzman notes [71], we cannot always map known models of human–human communication onto human–agent communication. Additionally, we should continue conducting specialized surveys to unearth particularities in terms of agent type, morphology, and technology. Further research can advance the Computers Are Social Actors theory for vHAI, especially its validity and replicability, as well as unearth new models. Replication will be an integral part of this work, and, as others have pointed out for specific kinds of voice-based

agents [21], demographic, social, and longitudinal factors should also be considered. For instance, the extent to which personalities (such as those based on the Big 5 model) can be replicated through voice when there is only voice to go on.

A longitudinal view also indicates a generational divide resulting from the emergence and proliferation of modern voice assistants (i.e., Siri, Alexa) that are equipped by default with feminine voices and ambiguous "bodies." We have proposed to call this the *vocaloid shift*. Over time, this shift seems to have influenced a preference for feminine voices in general as well as changes in stereotyped responses to perceived voice gender. In one sense, this highlights the social construction of gender and how design choices in technologies that become massively taken up can influence social perceptions. We can replicate older research to solidify this apparent effect, particularly those that had "opposite" results in terms of voice gender with certain user groups, tasks and activities, and contexts. We can also begin to compare more nuances questions about voice and body. Previous research on "bodyless" agents, especially voice source-ambiguous studies of computer voice, may have lacked the additional cues that "bodied" agents, especially social robots, provide. Yet, the new wave of smart speakers and other smart devices, with some focus given to the typically non-humanoid "body" of the device, provide an opportunity to see how much voice cues versus physical cues of cross-modal characteristics such as gender apply. There may be a shift over time that retrospective surveys and longitudinal work could illuminate.

Mapping the surveyed papers onto the voice–body continuums indicates some disciplinary disconnects that have resulted in clusters of particular combinations of voice morphologies, body morphologies, and types of agents. Specifically: voice with abstract bodies and technologies (Computers Are Social Actors), human and humanoid voices in humanoid and hybrid bodies (human–robot interaction), and an emerging "smart things" space (human–computer interaction and human–machine communication). Due to the methodological issues noted, the relationships between different forms of voice and combinations of voice and body morphologies is still unclear. Yet, the overarching findings suggest more similarity between these academic trajectories than differences. For instance, Computers Are Social Actors work, which largely relies on TTSs with ambiguous bodies, could be inspired by morphologies from human–robot interaction.

4.7.2 Consideration of Bias in vHAI Design, Research, and Beyond. Bias in AI has recently emerged as a key issue with widespread effects across the social, political, and legal spheres. As our findings show, it is also a *practical* and *scholarly* issue for technology researchers. Further, it is not only a matter of the AI as the "brain" of the agent (i.e., algorithms, pattern recognition libraries, and so on) but also the rest of the agent's embodiment, particularly voice. There are many areas technically and research-wise (e.g., gender) as well as gaps (e.g., can people discriminate between "young" and "old" voices? What effect does voice "age" have? And so on.) to address. On the flip side, human attitudes and conceptual models come into play. Indeed, the survey results go some way to suggest that people—designers and users both—are *constructing* the identities of vocal agents, even when voice is the sole modality of engagement. As the AI underlying agents become better at learning and acting within the world, this may become a process of *co-construction* [81] akin to human–human social embodiment, in which bias plays a key role.[3] At the moment, those creating and/or studying vocal agents can draw on the work of design reflexivity

[3]At the risk of sounding speculative, but with a pre-emptive view based on known evolutionary processes [181], we suggest to prepare now for the potential future of conscious AI by explicitly managing how we design, interact with, and "grow" the "primordial species" of AI in the present day. Indeed, there are promising machine learning approaches built around concepts of conscious agent networks [54]. There is also the "never-ending learning" paradigm [119], which suggests that private and personalized AI could be co-created much like a baby learns how to communicate with their parents. In fact, studies around baby babbling [47] may provide insights into how to engage people in such a process.

(e.g., Reference [151]) and social justice (e.g., Reference [57]) to reduce, if not avoid, bias and stereotyping in both the design of systems and research. This will likely mean redesigning agents and/or replicating previous research after correcting for instances of bias. Notably, minimizing bias may not lead to the discovery of "pure" truths in the world [188]; rather, it is more a matter of orientation, ethics, and sparking creativity in work on voice-based agents.

The last few years has seen a surge in mandates and societies for an ethics of AI: the Global Data Ethics Pledge [40], the Partnership on AI [177], and the Dagsthul Declaration [199], to name a few. The two major organizing bodies in computer science and engineering—ACM [3] and IEEE [80]—have also started to address ethics. While a unification of these ideas is beyond the scope of this article, we argue that it is essential. Furthermore, as this survey suggests, voice needs to be included as a major factor. The starting point can be exploratory work on ethical issues in design and research practice: from simple awareness at the design stage to critical discussions of research methods and results. We can do comparative and ethnographic studies of designers who begin with a certain ethical perspective, and see whether this influences the resulting vHAI experiences, and how. We can evaluate the relative role of voice and body more concretely by designing one or the other in a bias-conscious way and see which is more influential (i.e., exploring complementary and disruptive voice and body stereotypes through manipulating morphology and designers of varying bias awareness). We can draw from more established areas: a clear starting point would be the algorithm and machine learning side of agents. For instance, a report by the Nuffield Foundation [187] maps out three key directions in this area: term ambiguity, tensions emerging from conflicting values, and weaknesses in the research to date. We can then inform curricula on how to best design these agents [127, 128]. As this survey shows, vHAI have the power to change attitudes and actions. How we choose to design and study these agents will have ethical impacts beyond usability and UX. Even if we do not have control, we can assess the effects.

4.7.3 Incorporation of Rigorous and Relevant Theory and Methods for Voice. Almost half of studies did not reference theory. Yet, their findings contribute to one or more theories. This is a disconnect that needs to be rectified by grounding research within established theories, including general theories, such as the Computers Are Social Actors theory and the Big 5, as well as specific and even niche theories, which have largely been explored in isolation. Theories specific to voice, such as Tannen's theory of conversational styles [175] and the theory of doubly disembodied language [101], may be especially relevant. We should work towards finding consensus among theoretical and experiential constructs and then developing standard measures and instruments, including the creation of new standardized instruments for vHAI.

Methodologically, common issues in human subjects research apply to studies of vHAI as well. Most participant pools were comprised of young Westerners taking undergraduate studies at the institution. Often, it was unclear what ethnic or cultural background these participants had. While gender parity was achieved across the studies surveyed, many studies had uneven gender distributions. Experimental controls are needed. Related to this, when designing contextual features and social cues, such as gender and age, into voice-based agents, we need to be sure that our intended mental models are replicated faithfully. Manipulation and perception checks of designed features should become a staple of research designs. Explorations of voice and body can use audio-visual matching checks or even having participants draw the body from the voice [107]. Further, these should be done before the final study or prototype version as well as after other measures have been collected to avoid unintended effects, such as stereotype threats [168]. This is especially important given the prevalence of subjective measures in the body of work so far. Standard measures and statistics (e.g., Cronbach's α) need to be used and reported for ease of comparison and estimating consensus.

The measures themselves and the instruments designed to capture these measures need to be more carefully considered. While briefly touched in in the results, the specific instruments used were varied, even when the measures were the same or similar, and future work should seek to untangle this. One measure that we can address now is that of perceived gender, which was evaluated in a limited and biased way in most of the work surveyed. In self-report instruments, response categories should consider biological sex vs. social gender [79], social perceptions of gender, and the pitfalls and potentials of human imagination. While previous research has established frequency ranges for male and female voices, there is overlap and ambiguity [125]. Moreover, studies on transgender voice have implicated the role of social perception in assigning gender by voice [198]. This means not only allowing for the categories of men, women, nonbinary, and fluid but also "agent" genders and the possibility of no gender. Self-report instruments should also allow participants to "fill in the blank" to suggest gender categories. At the same time, it may not be possible to achieve gender-less or "gender-neutral" designs: as the survey finding suggest, people are inclined to attribute gender even when not intended by the designers. Moreover, these attributions tend to be stereotyped, based on the role, task, and application. A more open and comprehensive framing of measures and their interpretation can be grounded in existing and emerging theories in social psychology and other domains.

Technologically, many studies used a variety of TTS systems to evaluate agent voice. But TTS systems have advanced since the start of this work over 20 years ago. As the findings suggest, the "human" quality of the voice is a key factor, which may have impacted the results of synthetic humanoid TTS systems in the past. This work should be replicated with modern TTS systems. Furthermore, much custom technology was used to make comparisons, especially between voice and body and voice and human. More work using established suites (e.g., the Nao system, comprised of a physical robot, virtual robot, and TTS) or creating suites for comparisons such as these is needed. Finally, there needs to be greater consideration of "real" agent voices (through TTS systems or pre-recordings, etc.) and "fake" voices produced by real humans in realtime through the popular Wizard of Oz method. Without evaluations of readily available voice technologies, it is unclear how much of the findings generated will impact real life experience with the vHAI that use them.

Greater ecological validity such as through in situ work is also needed. Many voice-based agents—such as Siri and Alexa—are already "in the wild," encouraging if not enforcing, even unwittingly, ideas and values that may or may not be ethical. Future research should explicitly consider the real world through ethnographies (e.g., Reference [70]), field studies (e.g., Reference [16]), and "in the wild" methods [152], as well as longitudinal work to better understand the reach of this technology's influence in everyday life. Mixed methods can also augment the trend of reliance on subjective measures and self-reports [113].

4.7.4 Explorations Within and Beyond the Voice–Body Continuums. The vHAI experiences in most of the work surveyed fall along expected lines: humanoid morphologies in voice and body, regardless of type of agent (see Figure 6). Despite the uptake of smart speakers, smart vehicles, and other smart devices and "things," there has been little work on the voice of these machines. We can isolate gaps in the areas of abstract and object-based morphologies for voice and body. In these cases, the body of the agent may be obscure (or not). The idea of "disembodiment" should be considered in light of the uncertain effects of different agent bodies in different contexts, as well as the apparent vocaloid shift. Smart spaces that have a voice, for instance, have not been well studied. Yet, plenty of examples from popular culture suggest that certain contexts and use cases might challenge what we know about matching the voice to the body. For example, the *Enterprise* spaceships in Star Trek have a human voice augmented with tones and beeps, which may disrupt

the artificial–intuitive continuum if such agents work as effectively in reality as in these imagined worlds.

A match between voice and body is usually preferred. Still, it is unclear whether and what kind of *mismatches* are acceptable—or can be made to *feel* acceptable, given the right circumstances. There could possibly be an "uncanny voice" effect equivalent to the Uncanny Valley effect for "not quite right" humanoid robots and 3D virtual characters. Additionally, while adding a body to the voice generally had no ill effects, a lack of a perceivable body sometimes did, at least before the vocaloid shift. An open question is how this relates to voice assistants, smart speakers, and other body-less or body-ambiguous voice agents. Indeed, it is unclear to what (or where) people ascribe the body of such technologies—the speaker? The room? Someone teleconferencing from a distance?—and what perception of differing body sources means, if anything.

Other opportunities can be imagined when using the voice–body continuums vHAI framework as a starting point. Non-human animal voices and biological voices, as well as musical voices; whispering to but also the whispers of a vocal agent [142]. We can again turn to popular culture for inspiration. In Star Trek there are tribbles, a kind of space rabbit reminiscent of companion agents (e.g., Reference [185, 194]) that represent the audio modality as a natural complement to haptic experiences, similar to a cat purring. We can explore what kind of "voice" would be appropriate in various bodies, types of agents, contexts, and so on. There is much room for studies of semantic and paralinguistic features: gibberish [193] but also in terms of vocal fillers, semantic-free utterances, and so on. Indeed, the surveyed papers suggest several pathways for exploration. For instance, the use of vocal fillers by an agent may depend on the formality of the situation (informal better, formal worse) and the role of the agent (experienced worse, inexperienced better).

Multimodality theory can also be brought in. The visual modality is an important one, but others are worth exploring. For example, audio voices with text, akin to closed captioning systems and subtitles in television, movies, and video games. More radical combinations have recently been proposed, such as voice and heat [86]. Future work can explore multimodal patterns (e.g., redundancy, complementarity, etc. [66]) between various forms of voice in one agent. For instance, given the various difficulties and feasibility issues with speech systems, text could be used to provide linguistic content while the vocalic aspects of voice could be used to provide paralinguistic cues of personality, emotion, and so on. Relatedly, the survey findings confirm that low-level paralinguistic characteristics of voice correspond with higher order social constructs, such as gender, affect, and personality. Moreover, work on gibberish and non-semantic utterances suggests the radical idea that speech (as in linguistic, semantic content) may not always be necessary. Thus, future work can explore sociolinguistic questions of agent voice: modalities, semantics, and vocalics.

4.8 Limitations

We did not use all possible databases. As such, some relevant papers may have been unintentionally excluded. We manually added papers as needed, but follow-up work can confirm the validity of our selections. Additionally, we used strict criteria in terms of what research designs could be included, potentially excluding valuable findings from qualitative, autoethnographic, and other non-experimental work. We also restricted our focus to quantitative methods and measures; future work should review the qualitative work on voice.

In our voice–body continuums, we did not include the situatedness of the context in terms of being more or less simulated or realistic (from Reference [117]). It was not clear to us how to map such a continuum onto the voice and body morphology and type of agent continuums. We could imagine that any kind of voice and body could be represented at any point along the continuum, even if certain combinations cannot be found in the current body of work. Yet, some findings

indicate that important differences between virtuality and reality exist; as such, situatedness should be considered as a possible factor in future work.

5 CONCLUSION

"The voice of the machine" is rapidly becoming a key feature of human–agent interaction. As Pieraccini [144] and others have noted, enabling agents to recognize human speech and other communication channels is integral. This survey has shown that understanding how people interpret, respond to, and are influenced by agent voice is of equal import. The voice of the agent is socially situated, perceived, contextual, and dynamic. Features of the voice (and body) can have dramatic effects on vHAI experiences. This is not limited to the ability of an agent to *recognize* speech, gesture, nonverbal behaviors, or emotion but also its ability to *express* and *influence* through voice. There is great need and opportunity to consider how the voice of the machine is engineered and embedded in people's lives—and then how it is actually perceived in relation to the intended effects and goals of the larger system, service, or experience. We are already grappling with AI bias, poor usability, privacy and safety concerns, and other social factors within a technoscape increasingly featuring Alexa and Siri and other commercial voice assistants. We must determine and evaluate the fundamental human factors of human–agent interaction, and for this we add voice as one more "call" to action.

ACKNOWLEDGMENTS

We thank the editors and reviewers. We also thank the Japanese Society for the Promotion of Science (JSPS) for supporting this work.

REFERENCES

[1] Sameera A. Abdul-Kader and J. C. Woods. 2015. Survey on chatbot design techniques in speech conversation systems. *Int. J. Adv. Comput. Sci. Appl* 6, 7 (2015), 72–80. DOI: https://doi.org/10.14569/IJACSA.2015.060712

[2] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2019. A comparison of human and machine-generated voice. In *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology (VRST'19)*. ACM, 1–2. DOI: https://doi.org/10.1145/3359996.3364754

[3] ACM Council. ACM Ethics. Retrieved September 3, 2019 from https://ethics.acm.org/.

[4] Icek Ajzen. 1991. The theory of planned behavior. *Organ. Behav. Hum. Decis. Process* 50, 2 (1991), 179–211.

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, and Kori Inkpen. 2019. Guidelines for human-AI interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, Paper No. 3. DOI: https://doi.org/10.1145/3290605.3300233

[6] Sean Andrist, Micheline Ziadee, Halim Boukaram, Bilge Mutlu, and Majd Sakr. 2015. Effects of culture on the credibility of robot speech: A comparison between English and Arabic. In *Proceedings of the 10th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'15)*. ACM, 157–164. DOI: https://doi.org/10.1145/2696454.2696464

[7] Sercan Ö Arik, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, and Jonathan Raiman. 2017. Deep voice: Real-time neural text-to-speech. In *Proceedings of the 34th International Conference on Machine Learning*. JMLR.org, 195–204. Retrieved from https://dl.acm.org/citation.cfm?id=3305402.

[8] Robert K. Atkinson, Richard E. Mayer, and Mary Margaret Merrill. 2005. Fostering social agency in multimedia learning: Examining the impact of an animated agent's voice. *Contemp. Educ. Psychol.* 30, 1 (2005), 117–139. DOI: https://doi.org/10.1016/j.cedpsych.2004.07.001

[9] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Nicholas Cummins, Simone Hantke, and Björn Schuller. 2018. The perception of vocal traits in synthesized voices: Age, gender, and human likeness. *J. Audio Eng. Soc* 66, 4 (2018), 277–285.

[10] Alice Baird, Stina Hasse Jørgensen, Emilia Parada-Cabaleiro, Simone Hantke, Nicholas Cummins, and Björn Schuller. 2017. Perception of paralinguistic traits in synthesized voices. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences (AM'17)*. ACM, 17:1–17:5. DOI: https://doi.org/10.1145/3123514.3123528

[11] Lawrence W. Barsalou, Paula M. Niedenthal, Aron K. Barbey, and Jennifer A. Ruppert. 2003. Social embodiment. In *The Psychology of Learning and Motivation: Advances in Research and Theory*, Brian H. Ross (ed.). Academic Press, San Diego, CA, 43–92.

[12] Simone de Beauvoir. 2011. *The Second Sex*. Vintage Books, New York, NY.

[13] Sofie Ingeman Behrens, Anne Katrine Kongsgaard Egsvang, Michael Hansen, and Anton Mikkonen Møllegård-Schroll. 2018. Gendered robot voices and their influence on trust. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI'18)*. ACM, 63–64. DOI: https://doi.org/10.1145/3173386.3177009

[14] S. M. Bhagya, P. Samarakoon, M. A. Viraj, J. Muthugala, A. G. Buddhika, P. Jayasekara, and M. R. Elara. 2019. An exploratory study on proxemics preferences of humans in accordance with attributes of service robots. In *Proceedings of the 28th IEEE International Conference on Robot and Human Interactive Communication*. IEEE, 1–7. DOI: https://doi.org/10.1109/RO-MAN46459.2019.8956297

[15] Cheryl Campanella Bracken, Leo W. Jeffres, and Kimberly A. Neuendorf. 2004. Criticism or praise? The impact of verbal versus text-only computer feedback on social presence, intrinsic motivation, and recall. *Cyberpsychol. Behav.* 7, 3 (2004), 349–357. DOI: https://doi.org/10.1089/1094931041291358

[16] Michael Braun, Anja Mainz, Ronee Chadowitz, Bastian Pfleging, and Florian Alt. 2019. At your service: Designing voice assistant personalities to improve automotive user interfaces. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 40:1–40:11. DOI: https://doi.org/10.1145/3290605.3300270

[17] Cynthia Breazeal, Kerstin Dautenhahn, and Takayuki Kanda. 2016. Social robotics. In *Springer Handbook of Robotics*. Springer, Cham, Switzerland, 1935–1972. https://doi.org/10.1007/978-3-319-32552-1_72

[18] Susan E. Brennan and Eric A. Hulteen. 1995. Interaction and feedback in a spoken language system: A theoretical framework. *Knowl.-Based Syst.* 8, 2–3 (1995), 143–151.

[19] Mark Burgin and Gordana Dodig-Crnkovic. 2009. A systematic approach to artificial agents. Retrieved from https://ArXiv09023513.

[20] Julia Cambre, Jessica Colnago, Jim Maddock, Janice Tsai, and Jofish Kaye. 2020. Choice of voices: A large-scale evaluation of text-to-speech voice quality for long-form content. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 1–13. DOI: https://doi.org/10.1145/3313831.3376789

[21] Julia Cambre and Chinmay Kulkarni. 2019. One voice fits all?: Social implications and research challenges of designing voices for smart devices. *Proc. ACM Hum.-Comput. Interact.* 3 (2019), 1–19. DOI: https://doi.org/10.1145/3359325

[22] Angelo Cangelosi and Tetsuya Ogata. 2016. Speech and language in humanoid robots. In *Humanoid Robotics: A Reference,* A. Goswami and P. Vadakkepat (eds.). Springer, Dordrecht, The Netherlands, 1–32. Retrieved from https://link.springer.com/referenceworkentry/10.1007%2F978-94-007-7194-9_135-1.

[23] Hoang-Long Cao, Lars Christian Jensen, Xuan Nhan Nghiem, Huong Vu, Albert De Beir, Pablo Gomez Esteban, Greet Van de Perre, Dirk Lefeber, and Bram Vanderborght. 2019. DualKeepon: A human-robot interaction testbed to study linguistic features of speech. *Intell. Serv. Robot.* 12, 1 (2019), 45–54. DOI: https://doi.org/10.1007/s11370-018-0266-9

[24] Rebecca Cherng-Shiow Chang, Hsi-Peng Lu, and Peishan Yang. 2018. Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Comput. Hum. Behav.*84, (2018), 194–210. DOI: https://doi.org/10.1016/j.chb.2018.02.025

[25] Emna Chérif and Jean-François Lemoine. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Rech. Appl. En Mark. Engl. Ed* 34, 1 (2019), 28–47. DOI: https://doi.org/10.1177/2051570719829432

[26] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. In *Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*. ACM, 293–300. DOI: https://doi.org/10.1145/2157689.2157798

[27] Erin K. Chiou, Noah L. Schroeder, and Scotty D. Craig. 2020. How we trust, perceive, and learn from virtual humans: The influence of voice quality. *Comput. Educ.* 146, (2020), 103756. DOI: https://doi.org/10.1016/j.compedu.2019.103756

[28] Meia Chita-Tegmark, Monika Lohani, and Matthias Scheutz. 2019. Gender effects in perceptions of robots and humans with varying emotional intelligence. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 230–238. DOI: https://doi.org/10.1109/HRI.2019.8673222

[29] Mark Coeckelbergh. 2011. Humans, animals, and robots: A phenomenological approach to human-robot relations. *Int. J. Soc. Robot.* 3, 2 (2011), 197–204. DOI: https://doi.org/10.1007/s12369-010-0075-6

[30] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* 20, 1 (1960), 37–46.

[31] Sarah Cosentino, Salvatore Sessa, and Atsuo Takanishi. 2016. Quantitative laughter detection, measurement, and classification—A critical survey. *IEEE Rev. Biomed. Eng.* 9, (2016), 148–162. DOI: https://doi.org/10.1109/RBME.2016.2527638

[32] Scotty D. Craig and Noah L. Schroeder. 2017. Reconsidering the voice effect when learning from a virtual human. *Comput. Educ.* 114, (2017), 193–205. DOI: https://doi.org/10.1016/j.compedu.2017.07.003

[33] Chris Creed and Russell Beale. 2008. Psychological responses to simulated displays of mismatched emotional expressions. *Interact. Comput.* 20, 2 (2008), 225–39. DOI: https://doi.org/10.1016/j.intcom.2007.11.004

[34] John W. Creswell and Cheryl N. Poth. 2016. *Qualitative Inquiry and Research Design: Choosing Among Five Approaches* (4th ed.). Sage, Thousand Oaks, CA.

[35] Charles R. Crowell, Michael Villanoy, Matthias Scheutzz, and Paul Schermerhornz. 2009. Gendered voice and robot entities: Perceptions and reactions of male and female subjects. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3735–3741. DOI: https://doi.org/10.1109/IROS.2009.5354204

[36] Joe Crumpton and Cindy L. Bethel. 2016. A survey of using vocal prosody to convey emotion in robot speech. *Int. J. Soc. Robot.* 8, 2 (2016), 271–285. DOI: https://doi.org/10.1007/s12369-015-0329-4

[37] Nils Dahlbäck, Arne Jönsson, and Lars Ahrenberg. 1993. Wizard of Oz studies—Why and how. *Knowl.-Based Syst.* 6, 4 (1993), 258–266. DOI: https://doi.org/10.1016/0950-7051(93)90017-N

[38] Nils Dahlbäck, Seema Swamy, Clifford Nass, Fredrik Arvidsson, and Jörgen Skågeby. 2001. Spoken interaction with computers in a native or non-native language—Same or different? In *Proceedings of the International Conference on Human-Computer Interaction (INTERACT'01)*. 294–301.

[39] Nils Dahlbäck, QianYing Wang, Clifford Nass, and Jenny Alwin. 2007. Similarity is more important than expertise: Accent effects in speech interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'07)*. ACM, 1553–1556. DOI: https://doi.org/10.1145/1240624.1240859

[40] Data for Democracy. 2019. *Global Data Ethics Pledge (GDEP)*. Data for Democracy. Retrieved September 3, 2019 from https://github.com/Data4Democracy/ethics-resources.

[41] F. D. Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q* 13, 3 (1989), 319–339. DOI: https://doi.org/10.2307/249008

[42] Robert O. Davis, Joseph Vincent, and Taejung Park. 2019. Reconsidering the voice principle with non-native language speakers. *Comput. Educ.* 140, (2019), 103605. DOI: https://doi.org/10.1016/j.compedu.2019.103605

[43] Thomas J. Donahue and Matthias Scheutz. 2015. Investigating the effects of robot affect and embodiment on attention and natural language of human teammates. In *Proceedings of the 6th IEEE International Conference on Cognitive Infocommunications*. IEEE, 397–402. DOI: https://doi.org/10.1109/CogInfoCom.2015.7390626

[44] Xiao Dou, Chih-Fu Wu, Kai-Chieh Lin, Senzhong Gan, and Tzu-Min Tseng. 2020. Effects of different types of social robot voices on affective evaluations in different application fields. *Int. J. Soc. Robot.* (2020). DOI: https://doi.org/10.1007/s12369-020-00654-9

[45] Kathryn D. Drager, Joe Reichle, and Carrie Pinkoski. 2010. Synthesized speech output and children: A scoping review. *Am. J. Speech Lang. Pathol.* 19, 3 (2010), 259–273. DOI: https://doi.org/10.1044/1058-0360(2010/09-0024)

[46] Aaron C. Elkins and Douglas C. Derrick. 2013. The sound of trust: Voice as a measurement of trust during interactions with embodied conversational agents. *Group Decis. Negot.* 22, 5 (2013), 897–913. DOI: https://doi.org/10.1007/s10726-012-9339-x

[47] Steven L. Elmlinger, Jennifer A. Schwade, and Michael H. Goldstein. 2019. The ecology of prelinguistic vocal learning: Parents simplify the structure of their speech in response to babbling. *J. Child Lang.* 46, 5 (2019), 998–1011. DOI: https://doi.org/10.1017/S0305000919000291

[48] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychol. Rev.* 114, 4 (2007), 864. DOI: https://doi.org/10.1037/0033-295X.114.4.864

[49] Rochelle E. Evans and Philip Kortum. 2010. The impact of voice characteristics on user response in an interactive voice response system. *Interact. Comput.* 22, 6 (2010), 606–614. DOI: https://doi.org/10.1016/j.intcom.2010.07.001

[50] Friederike Eyssel, Dieta Kuchenbrandt, Frank Hegel, and Laura de Ruiter. 2012. Activating elicited agent knowledge: How robot and user features shape the perception of social robots. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 851–857. DOI: https://doi.org/10.1109/ROMAN.2012.6343858

[51] Friederike Eyssel, Laura de Ruiter, Dieta Kuchenbrandt, Simon Bobinger, and Frank Hegel. 2012. "If you sound like me, you must be more human": On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction*. 125–126. DOI: https://doi.org/10.1145/2157689.2157717

[52] Jasper Feine, Ulrich Gnewuch, Stefan Morana, and Alexander Maedche. 2019. A taxonomy of social cues for conversational agents. *Int. J. Hum.-Comput. Stud.* 132, (2019), 138–161. DOI: https://doi.org/10.1016/j.ijhcs.2019.07.009

[53] Leah Fessler. 2017. We tested bots like Siri and Alexa to see who would stand up to sexual harassment. *Quartz*. Retrieved September 18, 2019 from https://qz.com/911681/.

[54] Chris Fields, Donald D. Hoffman, Chetan Prakash, and Manish Singh. 2018. Conscious agent networks: Formal analysis and application to cognition. *Cogn. Syst. Res.* 47, (2018), 186–213. DOI: https://doi.org/10.1016/j.cogsys.2017.10.003

[55] William A. Firestone. 1993. Alternative arguments for generalizing from data as applied to qualitative research. *Educ. Res.* 22, 4 (1993), 16–23. DOI: https://doi.org/10.3102/0013189X022004016
[56] W. Tecumseh Fitch. 2017. Empirical approaches to the study of language evolution. *Psychon. Bull. Rev.* 24, 1 (2017), 3–33. DOI: https://doi.org/10.3758/s13423-017-1236-5
[57] Sarah Fox, Mariam Asad, Katherine Lo, Jill P. Dimond, Lynn S. Dombrowski, and Shaowen Bardzell. 2016. Exploring social justice, design, and HCI. In *Proceedings of the CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM, 3293–3300. DOI: https://doi.org/0.1145/2851581.2856465
[58] Maxine Gallander Wintre, Christopher North, and Lorne A. Sugar. 2001. Psychologists' response to criticisms about research based on undergraduate participants: A developmental perspective. *Can. Psychol. Can.* 42, 3 (2001), 216. DOI: https://doi.org/10.1037/h0086893
[59] P. Gangamohan, Sudarsana Reddy Kadiri, and B. Yegnanarayana. 2016. Analysis of emotional speech—a review. In *Toward Robotic Socially Believable Behaving Systems*. Springer, 205–238. Retrieved September 12, 2019 from https://doi.org/10.1007/978-3-319-31056-5_11.
[60] Aimi Shazwani Ghazali, Jaap Ham, Panos Markopoulos, and Emilia Barakova. 2019. Investigating the effect of social cues on social agency judgement. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 586–587. DOI: https://doi.org/10.1109/HRI.2019.8673266
[61] Henry Goble and Chad Edwards. 2018. A robot that communicates with vocal fillers has…uhhh…greater social presence. *Commun. Res. Rep.* 35, 3 (2018), 256–260. DOI: https://doi.org/10.1080/08824096.2018.1447454
[62] Li Gong and Jennifer Lai. 2003. To mix or not to mix synthetic speech and human speech? Contrasting impact on judge-rated task performance versus self-rated performance and attitudinal responses. *Int. J. Speech Technol.* 6, 2 (2003), 123–131. DOI: https://doi.org/10.1023/A:1022382413579
[63] Li Gong and Clifford Nass. 2007. When a talking-face computer agent is half-human and half-humanoid: Human identity and consistency preference. *Hum. Commun. Res.* 33, 2 (2007), 163–193. DOI: https://doi.org/10.1111/j.1468-2958.2007.00295.x
[64] D. Govind and S. R. Mahadeva Prasanna. 2013. Expressive speech synthesis: A review. *Int. J. Speech Technol.* 16, 2 (2013), 237–260. DOI: https://doi.org/10.1007/s10772-012-9180-2
[65] Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Comput. Speech Lang.* 25, 3 (2011), 601–634. DOI: https://doi.org/10.1016/j.csl.2010.10.003
[66] Patrizia Grifoni. 2009. *Multimodal Human Computer Interaction and Pervasive Services*. IGI Global, Hershey, PA.
[67] Jonathan Grudin. 2009. AI and HCI: Two fields divided by a common focus. *AI Mag.* 30, 4 (2009), 48–48. DOI: https://doi.org/10.1609/aimag.v30i4.2271
[68] David J. Gunkel. 2012. Communication and artificial intelligence: Opportunities and challenges for the 21st century. *Commun.* 11, 1 (2012), 1–25. DOI: https://doi.org/10.7275/R5QJ7F7R
[69] Andrea L. Guzman. 2016. Making AI safe for humans: A conversation with Siri. In *Socialbots and Their Friends: Digital Media and the Automation of Sociality*, R. W. Gehl and M. Bakardjieva (eds.). Routledge, New York, NY, 69–85.
[70] Andrea L. Guzman. 2019. Voices in and of the machine: Source orientation toward mobile virtual assistants. *Comput. Hum. Behav.* 90, (2019), 343–50. DOI: https://doi.org/10.1016/j.chb.2018.08.009
[71] Andrea L. Guzman and Seth C. Lewis. 2019. Artificial intelligence and communication: A human–machine communication research agenda. *New Media Soc.* 22, 1 (2019), 70–86. DOI: https://doi.org/10.1177/1461444819858691
[72] Edward Twitchell Hall. 1974. *Handbook for Proxemic Research*. Society for the Anthropology of Visual Communication, Washington, DC.
[73] Brian Hare. 2017. Survival of the friendliest: Homo sapiens evolved via selection for prosociality. *Annu. Rev. Psychol.* 68, (2017), 155–186. DOI: https://doi.org/10.1146/annurev-psych-010416-044201
[74] Shannon Hennig and Ryad Chellali. 2012. Expressive synthetic voices: Considerations for human robot interaction. In *Proceedings of the 21st IEEE International Symposium on Robot and Human Interactive Communication*. 589–595. DOI: https://doi.org/10.1109/ROMAN.2012.6343815
[75] Mieke Heyvaert, Bea Maes, and Patrick Onghena. 2013. Mixed methods research synthesis: Definition, framework, and potential. *Qual. Quant.* 47, 2 (2013), 659–676. DOI: https://doi.org/10.1007/s11135-011-9538-6
[76] Charles F. Hockett and Charles D. Hockett. 1960. The origin of speech. *Sci. Am.* 203, 3 (1960), 88–97.
[77] Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents (IVA'19)*. ACM, 111–118. DOI: https://doi.org/10.1145/3308532.3329473
[78] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Sci. Eng. Ethics* 24, 5 (2018), 1521–1536. DOI: https://doi.org/10.1007/s11948-017-9975-2
[79] Janet Shibley Hyde, Rebecca S. Bigler, Daphna Joel, Charlotte Chucky Tate, and Sari M. van Anders. 2019. The future of sex and gender in psychology: Five challenges to the gender binary. *Am. Psychol.* 74, 2 (2019), 171–193. DOI: https://doi.org/10.1037/amp0000307

[80] IEEE Robotics and Automation Society. 2019. Robot ethics. *IEEE Robot. Autom. Soc.* Retrieved September 12, 2019 from https://www.ieee-ras.org/robot-ethics.

[81] Sally Jacoby and Elinor Ochs. 1995. Co-construction: An introduction. *Res. Lang. Soc. Interact.* 28, 3 (1995), 171–183. DOI: https://doi.org/10.1207/s15327973rlsi2803_1

[82] Jesin James, Catherine Inez Watson, and Bruce MacDonald. 2018. Artificial empathy in social robots: An analysis of emotions in speech. In *Proceedings of the 27th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 632–637. DOI: https://doi.org/10.1109/ROMAN.2018.8525652

[83] Louise Jensen and Marion Allen. 1996. Meta-synthesis of qualitative findings. *Qual. Health Res.* 6, 4 (1996), 553–560. DOI: https://doi.org/10.1177/104973239600600407

[84] Yuin Jeong, Juho Lee, and Younah Kang. 2019. Exploring effects of conversational fillers on user perception of conversational agents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA'19)*, ACM, New York, NY, 1–6. DOI: https://doi.org/10.1145/3290607.3312913

[85] Swati Johar. 2016. *Emotion, Affect and Personality in Speech: The Bias of Language and Paralanguage*. Springer International Publishing. Retrieved September 13, 2019 from https://www.springer.com/us/book/9783319280455.

[86] Seyeong Kim, Yea-kyung Row, and Tek-Jin Nam. 2018. Thermal interaction with a voice-based intelligent agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI EA'18)*. ACM, 1–6. DOI: https://doi.org/10.1145/3170427.3188656

[87] Songhyun Kim, Junseok Goh, and Soojin Jun. 2018. The use of voice input to induce human communication with banking chatbots. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI'18)*. ACM, 151–152. DOI: https://doi.org/10.1145/3173386.3176970

[88] Barbara Kitchenham. 2004. Procedures for performing systematic reviews. Technical report, Keele University, Keele, UK.

[89] Robin Knote, Andreas Janson, Matthias Söllner, and Jan Marco Leimeister. 2019. Classifying smart personal assistants: An empirical cluster analysis. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*. 2024–2033. DOI: https://doi.org/10.24251/HICSS.2019.245

[90] Takanori Komatsu and Seiji Yamada. 2011. How does the agents' appearance affect users' interpretation of the agents' attitudes: Experimental investigation on expressing the same artificial sounds from agents with different appearances. *Int. J. Hum.–Comput. Interact.* 27, 3 (2011), 260–279. DOI: https://doi.org/10.1080/10447318.2011.537209

[91] Jody Kreiman, Gerratt R. Bruce, Gail B. Kempster, Andrew Erman, and Gerald S. Berke. 1993. Perceptual evaluation of voice quality. *J. Speech Lang. Hear. Res.* 36, 1 (1993), 21–40. DOI: https://doi.org/10.1044/jshr.3601.21

[92] Jody Kreiman, Diana Vanlancker-Sidtis, and Bruce R. Gerratt. 2003. Defining and measuring voice quality. In *Proceedings of the Conference on Voice Quality: Functions, Analysis, and Synthesis*. ISCA, 115–120. Retrieved December 17, 2019 from https://www.isca-speech.org/archive_open/voqual03/voq3_115.html.

[93] Brigitte Krenn, Stephanie Schreitter, and Friedrich Neubarth. 2017. Speak to me and I tell you who you are! A language-attitude study in a cultural-heritage application. *AI Soc.* 32, 1 (2017), 65–77. DOI: https://doi.org/10.1007/s00146-014-0569-0

[94] Jon A. Krosnick and Stanley Presser. 2010. Question and questionnaire design. In *Handbook of Survey Research* (2nd ed.). Emerald Group Publishing, Bingley, UK, 263–314.

[95] Neha Kumar and Naveena Karusala. 2019. Intersectional computing. *Interactions* 26, 2 (2019), 50–54. DOI: https://doi.org/10.1145/3305360

[96] Marianne Latinus and Pascal Belin. 2011. Human voice perception. *Curr. Biol.* 21, 4 (2011), R143–R145. DOI: https://doi.org/10.1016/j.cub.2010.12.033

[97] John Laver. 1980. *The Phonetic Description of Voice Quality*. Cambridge University Press, Cambridge, UK.

[98] Nicole Lazzeri, Daniele Mazzei, Maher Ben Moussa, Nadia Magnenat-Thalmann, and Danilo De Rossi. 2018. The influence of dynamics and speech on understanding humanoid facial expressions. *Int. J. Adv. Robot. Syst.* 15, (2018), 1729881418783158. DOI: https://doi.org/10.1177/1729881418783158

[99] Eun Ju Lee, Clifford Nass, and Scott Brave. 2000. Can computer-generated speech have gender?: An experimental test of gender stereotype. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI'00)*. ACM, 289–290. DOI: https://doi.org/10.1145/633292.633461

[100] Kwan Min Lee and Clifford Nass. 2003. Designing social presence of social actors in human computer interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'03)*. ACM, 289–296. DOI: https://doi.org/10.1145/642611.642662

[101] Kwan Min Lee and Clifford Nass. 2004. The multiple source effect and synthesized speech: Doubly disembodied language as a conceptual framework. *Hum. Commun. Res.* 30, 2 (2004), 182–207. DOI: https://doi.org/10.1111/j.1468-2958.2004.tb00730.x

[102] Sanguk Lee, Rabindra Ratan, and Taiwoo Park. 2019. The voice makes the car: Enhancing autonomous vehicle perceptions and adoption intention through voice agent gender and style. *Multimodal Technol. Interact.* 3, 1 (2019), 20. DOI: https://doi.org/10.3390/mti3010020

[103] Seul Chan Lee, Harsh Sanghavi, Sangjin Ko, and Myounghoon Jeon. 2019. Autonomous driving with an agent: speech style and embodiment. In *Proceedings of the 11th International Conference on Automotive User Interfaces and Interactive Vehicular Applications: Adjunct Proceedings (AutomotiveUI'19)*. Association for Computing Machinery, 209–214. DOI: https://doi.org/10.1145/3349263.3351515

[104] Michael Lewis. 1998. Designing for human-agent interaction. *AI Mag.* 19, 2 (1998), 67–67. DOI: https://doi.org/10.1609/aimag.v19i2.1369

[105] Judith Lorber. 1994. Night to his day: The social construction of gender. In *Paradoxes of Gender*. Yale University Press, 13–15, 32–36.

[106] Nichola Lubold, Erin Walker, and Heather Pon-Barry. 2016. Effects of voice-adaptation and social dialogue on perceptions of a robotic learning companion. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 255–262. DOI: https://doi.org/10.1109/HRI.2016.7451760

[107] Martina Mara, Simon Schreibelmayr, and Franz Berger. 2020. Hearing a nose? User expectations of robot appearance induced by different robot voices. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI'20)*. ACM, 355–356. DOI: https://doi.org/10.1145/3371382.3378285

[108] George M. Marakas, Richard D. Johnson, and Jonathan W. Palmer. 2000. A theoretical model of differential social attributions toward computing technology: When the metaphor becomes the model. *Int. J. Hum.-Comput. Stud.* 52, 4 (2000), 719–750. DOI: https://doi.org/10.1006/ijhc.1999.0348

[109] Richard E. Mayer, Kristina Sobko, and Patricia D. Mautone. 2003. Social cues in multimedia learning: Role of speaker's voice. *J. Educ. Psychol.* 95, 2 (2003), 419–425. DOI: https://doi.org/10.1037/0022-0663.95.2.419

[110] Roger C. Mayer, James H. Davis, and F. David Schoorman. 1995. An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 3 (1995), 709–734.

[111] Derek McColl, Alexander Hong, Naoaki Hatakeyama, Goldie Nejat, and Beno Benhabib. 2016. A survey of autonomous human affect detection methods for social robots engaged in natural HRI. *J. Intell. Robot. Syst.* 82, 1 (2016), 101–133. DOI: https://doi.org/10.1007/s10846-015-0259-2

[112] Robert R. McCrae and Paul T. Costa, Jr. 2008. A five-factor theory of personality. In *Handbook of Personality: Theory and Research* (3rd ed.), Oliver P. John, Richard W. Robins and Lawrence A. Pervin (eds.). The Guildford Press, New York/London, 159–181.

[113] Jennifer Dodorico McDonald. 2008. Measuring personality constructs: The advantages and disadvantages of self-reports, informant reports and behavioural assessments. *Enquire* 1, 1 (2008), 1–19. DOI: https://doi.org/10.1.1.523.2142

[114] Conor McGinn and Ilaria Torre. 2019. Can you tell the robot by the voice? An exploratory study on the role of voice in the perception of robots. In *Proceedings of the 14th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 211–221. DOI: https://doi.org/10.1109/HRI.2019.8673305

[115] Michael F. McTear, Zoraida Callejas, and David Griol. 2016. *The Conversational Interface: Talking to Smart Devices*. Springer International Publishing, Switzerland.

[116] Joseph Mendelson and Matthew Aylett. 2017. Beyond the listening test: An interactive approach to tts evaluation. In *Proceedings of the 18th Annual Conference of the International Speech Communication Association*. 249–253. DOI: https://doi.org/10.21437/Interspeech.2017-1438

[117] Paul Milgram and Fumio Kishino. 1994. A taxonomy of mixed reality visual displays. *IEICE Trans. Inf. Syst.* E77-D, 12 (1994), 1321–1329.

[118] Blanca Miller and David Feil-Seifer. 2019. Embodiment, situatedness, and morphology for humanoid robots interacting with people. *Humanoid Robot. Ref.* (2019), 2313–2335. DOI: https://doi.org/10.1007/978-94-007-7194-9_130-1

[119] Tom Mitchell, William Cohen, Estevam Hruschka, Partha Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, B. Dalvi, Matt Gardner, and Bryan Kisiel. 2018. Never-ending learning. *Commun. ACM* 61, 5 (2018), 103–115. DOI: https://doi.org/10.1145/3191513

[120] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Med.* 6, 7 (2009), e1000097. DOI: https://doi.org/10.1371/journal.pmed.1000097

[121] Csaba Molnár, Frédéric Kaplan, Pierre Roy, François Pachet, Péter Pongrácz, Antal Dóka, and Ádám Miklósi. 2008. Classification of dog barks: A machine learning approach. *Anim. Cogn.* 11, 3 (2008), 389–400. DOI: https://doi.org/10.1007/s10071-007-0129-9

[122] Roxana Moreno, Richard E. Mayer, Hiller A. Spires, and James C. Lester. 2001. The case for social agency in computer-based teaching: Do students learn more deeply when they interact with animated pedagogical agents? *Cogn. Instr.* 19, 2 (2001), 177–213. DOI: https://doi.org/10.1207/S1532690XCI1902_02

[123] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The uncanny valley [from the Field]. *IEEE Robot. Autom. Mag.* 19, 2 (2012), 98–100. DOI: https://doi.org/10.1109/MRA.2012.2192811

[124] David A. Moses, Matthew K. Leonard, Joseph G. Makin, and Edward F. Chang. 2019. Real-time decoding of question-and-answer speech dialogue using human cortical activity. *Nat. Commun.* 10, 1 (2019), 1–14. DOI: https://doi.org/10.1038/s41467-019-10994-4

[125] John W. Mullennix, Keith A. Johnson, Meral Topcu-Durgun, and Lynn M. Farnsworth. 1995. The perceptual representation of voice gender. *J. Acoust. Soc. Am.* 98, 6 (1995), 3080–3095. DOI: https://doi.org/10.1121/1.413832

[126] John W. Mullennix, Steven E. Stern, Stephen J. Wilson, and Corrie-lynn Dyson. 2003. Social perception of male and female computer synthesized speech. *Comput. Hum. Behav.* 19, 4 (2003), 407–424. DOI: https://doi.org/10.1016/S0747-5632(02)00081-X

[127] Christine Murad and Cosmin Munteanu. 2020. Designing voice interfaces: Back to the (curriculum) basics. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–12. DOI: https://doi.org/10.1145/3313831.3376522

[128] Christine Murad and Cosmin Munteanu. 2020. Alexa, how do I build a VUI curriculum? In *Proceedings of the 2nd Conference on Conversational User Interfaces (CUI'20)*. ACM, 1–3. DOI: https://doi.org/10.1145/3405755.3406137

[129] Kevin L. Nadal. 2017. The SAGE encyclopedia of psychology and gender. In *Neurosexism*. SAGE Publications, Thousand Oaks, CA, 1243–1246.

[130] Clifford Nass, Ulla Foehr, Scott Brave, and Michael Somoza. 2001. The effects of emotion of voice in synthesized and recorded speech. In *Proceedings of the AAAI Symposium Emotional and Intelligent II*. AAAI.

[131] Clifford Ivar Nass and Scott Brave. 2005. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*. MIT Press, Cambridge, MA.

[132] Clifford Nass, Ing-Marie Jonsson, Helen Harris, Ben Reaves, Jack Endo, Scott Brave, and Leila Takayama. 2005. Improving automotive safety by pairing driver emotion and car voice emotion. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI EA'05)*. ACM, 1973–1976. DOI: https://doi.org/10.1145/1056808.1057070

[133] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *J. Exp. Psychol. Appl.* 7, 3 (2001), 171. DOI: https://doi.org/10.1037//1076-898X.7.3.171

[134] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are machines gender neutral? Gender-stereotypic responses to computers with voices. *J. Appl. Soc. Psychol.* 27, 10 (1997), 864–876. DOI: https://doi.org/10.1111/j.1559-1816.1997.tb00275.x

[135] Clifford Nass, Erica Robles, Charles Heenan, Hilary Bienstock, and Marissa Treinen. 2003. Speech-based disclosure systems: Effects of modality, gender of prompt, and gender of user. *Int. J. Speech Technol.* 6, 2 (2003), 113–121. DOI: https://doi.org/10.1023/A:1022378312670

[136] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 72–78. DOI: https://doi.org/10.1145/191666.191703

[137] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, Haizhou Li, and Swee Lan See. 2013. Making social robots more attractive: The effects of voice pitch, humor and empathy. *Int. J. Soc. Robot.* 5, 2 (2013), 171–191. DOI: https://doi.org/10.1007/s12369-012-0171-x

[138] Andreea Niculescu, Betsy van Dijk, Anton Nijholt, and Swee Lan See. 2011. The influence of voice pitch on the evaluation of a social robot receptionist. In *Proceedings of the International Conference on User Science and Engineering*. 18–23. DOI: https://doi.org/10.1109/iUSEr.2011.6150529

[139] Nils J. Nilsson. 2009. *The Quest for Artificial Intelligence*. Cambridge University Press, Cambridge, UK.

[140] Naoki Ohshima, Keita Kimijima, Junji Yamato, and Naoki Mukawa. 2015. A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings. In *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN'15)*. IEEE, 325–30. DOI: https://doi.org/10.1109/ROMAN.2015.7333677

[141] Kengo Ohta, Norihide Kitaoka, and Seiichi Nakagawa. 2014. Modeling filled pauses and silences for responses of a spoken dialogue system. *Int. J. Comput.* 8, (2014), 136–142.

[142] Emmi Parviainen and Marie Louise Juul Søndergaard. 2020. Experiential qualities of whispering with voice assistants. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'20)*. ACM, 1–13. DOI: https://doi.org/10.1145/3313831.3376187

[143] Rolf Pfeifer and Christian Scheier. 2001. *Understanding Intelligence*. MIT Press, Cambridge, MA.

[144] Roberto Pieraccini. 2012. *The Voice in the Machine: Building Computers That Understand Speech*. MIT Press, Cambridge, MA.

[145] Jeff Pittam. 1994. *Voice in Social Interaction: An Interdisciplinary Approach*. Sage, Thousand Oaks, CA.

[146] Denise F. Polit and Cheryl Tatano Beck. 2010. Generalization in quantitative and qualitative research: Myths and strategies. *Int. J. Nurs. Stud.* 47, 11 (2010), 1451–1458. DOI: https://doi.org/10.1016/j.ijnurstu.2010.06.004

[147] Fernando Poyatos. 1993. *Paralanguage: A Linguistic and Interdisciplinary Approach to Interactive Speech and Sounds*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

[148] Lingyun Qiu and Izak Benbasat. 2009. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *J. Manag. Inf. Syst.* 25, 4 (2009), 145–182. DOI: https://doi.org/10.2753/MIS0742-1222250405

[149] Robin Read and Tony Belpaeme. 2012. How to use non-linguistic utterances to convey emotion in child-robot interaction. In *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI'12)*. IEEE, 219–220. DOI: https://doi.org/10.1145/2157689.2157764

[150] Robin Read and Tony Belpaeme. 2013. People interpret robotic non-linguistic utterances categorically. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction*. IEEE, 209–210. DOI: https://doi.org/10.1109/HRI.2013.6483575

[151] Jennifer A. Rode. 2011. Reflexivity in digital anthropology. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 123–132. DOI: https://doi.org/10.1145/1978942.1978961

[152] Yvonne Rogers and Paul Marshall. 2017. *Research in the Wild*. Morgan & Claypool.

[153] Rinat B. Rosenberg-Kima, Amy L. Baylor, E. Ashby Plant, and Celeste E. Doerr. 2007. The importance of interface agent visual presence: Voice alone is less effective in impacting young women's attitudes toward engineering. In *Proceedings of the International Conference on Persuasive Technology*. Springer, 214–222. DOI: https://doi.org/10.1007/978-3-540-77006-0_27

[154] Astrid M. Rosenthal-von der Pütten, Carolin Strasmann, and Nicole C. Kramer. 2016. Robots or agents - neither helps you more or less during second language acquisition. In *Proceedings of the 16th International Conference on Intelligent Virtual Agents (IVA'16)*. 256–68. DOI: https://doi.org/10.1007/978-3-319-47665-0_23

[155] James A. Russell. 1980. A circumplex model of affect. *J. Pers. Soc. Psychol.* 39, 6 (1980), 1161–1178. DOI: https://doi.org/10.1037/h0077714

[156] Anara Sandygulova and Gregory M. P. O'Hare. 2015. Children's perception of synthesized voice: Robot's gender, age and accent. In *Proceedings of the 7th International Conference on Social Robotics*. 594–602. DOI: https://doi.org/10.1007/978-3-319-25554-5_59

[157] Anara Sandygulova and Gregory M. P. O'Hare. 2018. Age- and gender-based differences in children's interactions with a gender-matching robot. *Int. J. Soc. Robot.* 10, 5 (2018), 687–700. DOI: https://doi.org/10.1007/s12369-018-0472-9

[158] Busra Sarigul, Imge Saltik, Batuhan Hokelek, and Burcu A. Urgen. 2020. Does the appearance of an agent affect how we perceive his/her voice? Audio-visual predictive processes in human-robot interaction. In *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRICompanion'20)*. ACM, 430–432. DOI: https://doi.org/10.1145/3371382.3378302

[159] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging identity through gender, race, and class. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 5412–5427. DOI: https://doi.org/10.1145/3025453.3025766

[160] Emre Sezgin, Lisa Militello, Yungui Huang, and Simon Lin. 2019. A scoping review of patient-facing, behavioral health Interventions with voice assistant technology targeting self-management and healthy lifestyle behaviors. *Translation. Behav. Med.* 10, 3 (2019). DOI: https://doi.org/10.2139/ssrn.3381183

[161] Ameneh Shamekhi, Q. Vera Liao, Dakuo Wang, Rachel K. E. Bellamy, and Thomas Erickson. 2018. Face value? Exploring the effects of embodiment for a group facilitation agent. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'18)*. ACM, 391:1–391:13. DOI: https://doi.org/10.1145/3173574.3173965

[162] Yang Shi, Xin Yan, Xiaojuan Ma, Yongqi Lou, and Nan Cao. 2018. Designing emotional expressions of conversational states for voice assistants: Modality and engagement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–6. DOI: https://doi.org/10.1145/3170427.3188560

[163] Ryoko Shibata, Chie Fukada, Takatsugu Kojima, Kaori Sato, Yuki Hashikura, Motoyuki Ozeki, and Natsuki Oka. 2012. Does talking to a robot in a high-pitched voice create a good impression of the robot? In *Proceedings of the 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD'12)*. IEEE, 19–24. DOI: https://doi.org/10.1109/SNPD.2012.72

[164] Toshiyuki Shiwa, Takayuki Kanda, Michita Imai, Hiroshi Ishiguro, and Norihiro Hagita. 2008. How quickly should communication robots respond? In *Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction (HRI'08)*. IEEE, 153–160. DOI: https://doi.org/10.1145/1349822.1349843

[165] Mikey Siegel, Cynthia Breazeal, and Michael I. Norton. 2009. Persuasive robotics: The influence of robot gender on human behavior. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2563–2568. DOI: https://doi.org/10.1109/IROS.2009.5354116

[166] Valerie K. Sims, Matthew G. Chin, Heather C. Lum, Linda Upham-Ellis, Tatiana Ballion, and Nicholas C. Lagattuta. 2009. Robots' auditory cues are subject to anthropomorphism. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* 53, 18 (2009), 1418–1421. DOI: https://doi.org/10.1177/154193120905301853

[167] Roger Andre Søraa. 2017. Mechanical genders: How do humans gender robots? *Gend. Technol. Dev.* 21, 1–2 (2017), 99–115. DOI: https://doi.org/10.1080/09718524.2017.1385320

[168] Claude M. Steele and Joshua Aronson. 1995. Stereotype threat and the intellectual test performance of African Americans. *J. Pers. Soc. Psychol.* 69, 5 (1995), 797–811. DOI: https://doi.org/10.1037/0022-3514.69.5.797

[169] Steven E. Stern, John W. Mullennix, Corrie-lynn Dyson, and Stephen J. Wilson. 1999. The persuasiveness of synthetic speech versus human speech. *Hum. Factors* 41, 4 (1999), 588–595. DOI: https://doi.org/10.1518/001872099779656680

[170] Steven E. Stern, John W. Mullennix, and Ilya Yaroslavsky. 2006. Persuasion and social perception of human vs. synthetic voice across person as source and computer as source conditions. *Int. J. Hum.-Comput. Stud.* 64, 1 (2006), 43–52. DOI: https://doi.org/10.1016/j.ijhcs.2005.07.002

[171] Selina Jeanne Sutton, Paul Foulkes, David Kirk, and Shaun Lawson. 2019. Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI'19)*. ACM, 1–14. DOI: https://doi.org/10.1145/3290605.3300833

[172] John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cogn. Sci.* 12, (1988), 257–285.

[173] Henri Tajfel (Ed.). 2010. *Social Identity and Intergroup Relations*. Cambridge University Press, Cambridge, UK.

[174] Rie Tamagawa, Catherine I. Watson, I. Han Kuo, Bruce A. MacDonald, and Elizabeth Broadbent. 2011. The effects of synthesized voice accents on user perceptions of robots. *Int. J. Soc. Robot.* 3, 3 (2011), 253–262. DOI: https://doi.org/10.1007/s12369-011-0100-4

[175] Deborah Tannen. 2005. *Conversational Style: Analyzing Talk Among Friends*. Oxford University Press, Oxford, UK.

[176] Benedict Tay, Younbo Jung, and Taezoon Park. 2014. When stereotypes meet robots: The double-edge sword of robot gender and personality in human–robot interaction. *Comput. Hum. Behav.* 38, (2014), 75–84. DOI: https://doi.org/10.1016/j.chb.2014.05.014

[177] The Partnership on AI. 2019. Retrieved September 3, 2019 from https://www.partnershiponai.org/.

[178] Nenad Tomašev, Julien Cornebise, Frank Hutter, Shakir Mohamed, Angela Picciariello, Bec Connelly, Danielle C. M. Belgrave, Daphne Ezer, Fanny Cachat van der Haert, Frank Mugisha, Gerald Abila, Hiromi Arai, Hisham Almiraat, Julia Proskurnia, Kyle Snyder, Mihoko Otake-Matsuura, Mustafa Othman, Tobias Glasmachers, Wilfried de Wever, Yee Whye Teh, Mohammad Emtiyaz Khan, Ruben De Winne, Tom Schaul, and Claudia Clopath. 2020. AI for social good: Unlocking the opportunity for positive impact. *Nat. Commun* 11, 1 (2020), 2468. DOI: https://doi.org/10.1038/s41467-020-15871-z

[179] Ilaria Torre, Jeremy Goslin, Laurence White, and Debora Zanatto. 2018. Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. In *Proceedings of the Technology, Mind, and Society Conference (TechMindSociety'18)*. ACM, Article No. 40. DOI: https://doi.org/10.1145/3183654.3183691

[180] Cristen Torrey, Susan R. Fussell, and Sara Kiesler. 2013. How a robot should give advice. In *Proceedings of the 8th ACM/IEEE International Conference on Human-Robot Interaction (HRI'13)*. 275–282. DOI: https://doi.org/10.1109/HRI.2013.6483599

[181] Robert Trivers. 2002. *Natural Selection and Social Theory: Selected Papers of Robert Trivers*. Oxford University Press, Oxford, UK.

[182] Gabriele Trovato, Josue G. Ramos, Helio Azevedo, Artemis Moroni, Silvia Magossi, Reid Simmons, Hiroyuki Ishii, and Atsuo Takanishi. 2017. A receptionist robot for Brazilian people: Study on interaction involving illiterates. *Paladyn J. Behav. Robot.* 8, 1 (2017), 1–17. DOI: https://doi.org/10.1515/pjbr-2017-0001

[183] Christiana Tsiourti, Astrid Weiss, Katarzyna Wac, and Markus Vincze. 2019. Multimodal integration of emotional signals from voice, body, and context: Effects of (in)congruence on emotion recognition and attitudes towards robots. *Int. J. Soc. Robot.* 11, 4 (2019), 555–573. DOI: https://doi.org/10.1007/s12369-019-00524-z

[184] F. Vannucci, G. Di Cesare, F. Rea, G. Sandini, and A. Sciutti. 2018. A robot with style: Can robotic attitudes influence human actions? In *Proceedings of the 18th International Conference on Humanoid Robots*. IEEE, 1–6. DOI: https://doi.org/10.1109/HUMANOIDS.2018.8625004

[185] Kazuyoshi Wada, Takanori Shibata, Tomoko Saito, and Kazuo Tanie. 2004. Effects of robot-assisted activity for elderly people and nurses at a day service center. *Proc. IEEE* 92, 11 (2004), 1780–1788. DOI: https://doi.org/10.1109/JPROC.2004.835378

[186] Michael Leonard Walters, Dag Sverre Syrdal, Kheng Lee Koay, Kerstin Dautenhahn, and René te Boekhorst. 2008. Human approach distances to a mechanical-looking robot with different robot voice styles. In *Proceedings of the 17th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 707–712. DOI: https://doi.org/10.1109/ROMAN.2008.4600750

[187] Jess Whittlestone, Rune Nyrup, Anna Alexandrova, Kanta Dihal, and Stephen Cave. 2019. Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. *Lond. Nuffield Found.* Retrieved from https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf.

[188] Sietse Wieringa, Eivind Engebretsen, Kristin Heggen, and Trish Greenhalgh. 2018. Rethinking bias and truth in evidence-based health care. *J. Eval. Clin. Pract.* 24, 5 (2018), 930–938. DOI: https://doi.org/10.1111/jep.13010

[189] Noel Wigdor, Joachim de Greeff, Rosemarijn Looije, and Mark A. Neerincx. 2016. How to improve human-robot interaction with Conversational Fillers. In *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 219–224. DOI: https://doi.org/10.1109/ROMAN.2016.7745134

[190] Kun Xu. 2019. First encounter with robot Alpha: How individual differences interact with vocal and kinetic cues in users' social responses. *New Media Soc.* (2019), 1461444819851479. DOI: https://doi.org/10.1177/1461444819851479

[191] Svetlana Yarosh, Stryker Thompson, Kathleen Watson, Alice Chase, Ashwin Senthilkumar, Ye Yuan, and A. J. Bernheim Brush. 2018. Children asking questions: Speech interface reformulations and personification preferences. In *Proceedings of the 17th ACM Conference on Interaction Design and Children (IDC'18)*. ACM, 300–312. DOI: https://doi.org/10.1145/3202185.3202207

[192] Selma Yilmazyildiz, Robin Read, Tony Belpeame, and Werner Verhelst. 2016. Review of semantic-free utterances in social human-robot interaction. *Int. J. Hum.-Comput. Interact.* 32, 1 (2016), 63–85. DOI: https://doi.org/10.1080/10447318.2015.1093856

[193] Selma Yilmazyildiz, Werner Verhelst, and Hichem Sahli. 2015. Gibberish speech as a tool for the study of affective expressiveness for robotic agents. *Multimed. Tools Appl.* 74, 22 (2015), 9959–9982. DOI: https://doi.org/10.1007/s11042-014-2165-1

[194] Steve Yohanan, Mavis Chan, Jeremy Hopkins, Haibo Sun, and Karon MacLean. 2005. Hapticat: Exploration of affective touch. In *Proceedings of the 7th International Conference on Multimodal Interfaces*. ACM, 222–229. DOI: https://doi.org/10.1145/1088463.1088502

[195] Qian Yu, Tonya Nguyen, Soravis Prakkamakul, and Niloufar Salehi. 2019. "I almost fell in love with a machine": Speaking with computers affects self-disclosure. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI EA'19)*. ACM, 1–6. DOI: https://doi.org/10.1145/3290607.3312918

[196] Cristina Zaga, Roelof A. De Vries, Sem J. Spenkelink, Khiet P. Truong, and Vanessa Evers. 2016. Help-giving robot behaviors in child-robot games: Exploring Semantic Free Utterances. In *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI'16)*. 541–542. DOI: https://doi.org/10.1109/HRI.2016.7451846

[197] Catherine Zanbaka, Paula Goolkasian, and Larry Hodges. 2006. Can a virtual cat persuade you?: The role of gender and realism in speaker persuasiveness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1153–1162. DOI: https://doi.org/10.1145/1124772.1124945

[198] Lal Zimman. 2018. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Lang. Linguist. Compass.* 12, 8 (2018), e12284. DOI: https://doi.org/10.1111/lnc3.12284

[199] AI for the Social Good. 2019. The Dagsthul Declaration. Retrieved July 25, 2019 from https://aiforthesocialgood.com/.

Received February 2020; revised January 2021; accepted January 2021