Please quote as: Schmitt, Anuschka; Zierau, Naim; Janson, Andreas & Leimeister, Jan Marco: Voice as a Contemporary Frontier of Interaction Design. 2021. - European Conference on Information Systems (ECIS). - Virtual.

VOICE AS A CONTEMPORARY FRONTIER OF INTERACTION DESIGN

Research Paper

Anuschka Schmitt, University of St. Gallen, St.Gallen, Switzerland, anuschka.schmitt@unisg.ch

Naim Zierau, University of St.Gallen, St. Gallen, Switzerland, naim.zierau@unisg.ch

Andreas Janson, University of St. Gallen, St.Gallen, Switzerland, andreas.janson@unisg.ch

Jan Marco Leimeister, University of St. Gallen, St.Gallen, Switzerland, janmarco.leimeister@unisg.ch

Abstract

Voice assistants' increasingly nuanced and natural communication bears new opportunities for user experiences and task automation, while challenging existing patterns of human-computer interaction. A fragmented research field, as well as constant technological advancements, impede a common apprehension of prevalent design features of voice-based interfaces. As part of this study, 86 papers across domains are systematically identified and analysed to arrive at a common understanding of voice assistants. The review highlights perceptual differences to other human-computer interfaces and points out relevant auditory cues. Key findings regarding those cues' impact on user perception and behaviour are discussed along with the three design strategies 1) personification, 2) individualization and 3) contextualization. Avenues for future research are lastly deducted. Our results provide relevant opportunities to researchers and designers alike to advance the design and deployment of voice assistants.

Keywords: Voice Assistant, Conversational Interface, Voice Design, Human-Computer Interaction

1 Introduction

In 2013, the ever-present and empathic artificial voice personality, which accompanied the protagonist Theodore Twombly in the Sci-Fi Drama "Her", seemed shockingly imaginable, yet still far from reality. With the integration of voice assistants (VAs) in mobile devices, personal homes, and service interfaces, voice has arrived in our everyday life, just as it has in Twombly's. *Amazon's Alexa*, for instance, learns from mistakes, is taking breathing breaks, adapts her tone of voice to the context of the conversation and proactively offers her help (Low, 2020; Tarantola, 2020). In short, interactions with VAs promise to be increasingly natural and effortless. Adoption and use numbers underline the wide-spread acceptance of such agents: With over 110 million users in the US alone, the number of VAs is expected to reach 8.4 billion by 2024 (Smith, 2020). More interestingly, one is quick to imagine the potential of VAs: The hands- and eyes-free nature of VAs offers promising applications for certain user groups and contexts, especially against the backdrop of 2.2 million people with vision impairment worldwide (Abdolrahmani et al., 2018; World Health Organization, 2019).

While the fields of application seem endless and the implementation benefits intuitive, the empirical understanding of VAs is not as straightforward. Research on VAs is growing rapidly yet is highly fragmented in terms of publication outlet. As a result, two key research challenges arise: First, different conceptualizations of voice-based agents exist. Second, it becomes challenging to interpret and predict user perceptions and behaviour for different deployment contexts. The fragmented research body calls for a context-sensitive knowledge base on voice design. Existing frameworks and categorisations of design-relevant features for VAs are characterized by an either isolated perspective focusing on certain types of VAs only, or by considering VAs as one type of conversational agents and subsequently black-boxing voice-specific cues (Cambre & Kulkarni, 2019). Potential for VAs has been recognized with researchers from different domains alike calling to explore how to design a conversational interface, and how to consecutively evaluate user satisfaction (McTear, 2017). Pfeuffer, Benlian, Gimpel and Hinz (2019) raise the importance of considering individual auditory features of such assistants. This study aims to contribute to the field of IS by bringing together disparate research streams on voice-based conversational agents towards a more holistic perspective on related design cues. More precisely, auditory cues found in VAs are mapped against related user perception and interaction outcomes for a more integrated and context-sensitive analysis of VA design. Ultimately, this study poses the following research question:

What are voice-relevant cues of VAs and how can they be leveraged for interaction design?

This paper is organized as follows. Section 2 presents the conceptual background, followed by section 3 illustrating the method of the systematic literature review. Section 4 reviews the identified literature and discusses main findings along three design strategies. A discussion and research agenda are presented in section 5. Finally, conclusions and limitations are pointed out.

2 Conceptual Background

2.1 Voice as an Interaction Modality

VAs, which are also referred to as conversational agents, or dialogue systems (Balasuriya et al., 2018; O'Brien et al., 2020), are computer programs that interact with customers via automatic speech recognition and natural language processing in form of voice-mediated communication (Pfeuffer et al., 2019). The increasingly sophisticated interaction quality of VAs and omnipresence of such interfaces has been propelled by ongoing advances in Artificial Intelligence (AI). However, with a plethora of existing personal agents and conversational interfaces, what makes voice so different?

As illustrated in Table 1, the nature of voice exhibits several features which ultimately differentiate voice-based interaction from other prevalent interface modalities, such as text. First, voice-based interaction follows a sequential processing, which resembles natural human communication. Text-based conversation, on the other hand, allows to re-read and re-write messages, which, however, also introduces less natural pauses (Rubin et al., 2000). Second, voice-based interaction is marked by more colloquial language, which allows for intuitive expression and reception of information (Dennis et al., 2008). Third, voice encompasses signals beyond mere information provision and thus changes the nature of interaction between users and a conversational agent sustainably (Rosenthal & Ryan, 2000). Voice not only exhibits a greater variety and intuitively appropriated cues, yet also, and most importantly, a great extent of nonverbal cues (Redeker, 1984). In fact, voice has been defined as "the carrier of speech" (Belin, Fecteau and Bédard, 2004, p. 129) as it becomes distinguishable from other modalities by signalling social cues beyond the pure verbal information communicated (i.e. content). In that sense, voice provides cues about the personality and identity (i.e., gender, age), as well as emotions (i.e., happiness, anger), of a speaker (Sutton et al., 2019). Overall, distinguishing features of voice-based interaction offer novel opportunities for VAs as an interaction interface.

	Text-Based Interactions	**Voice-Based Interactions**
Interaction Flow	Permanent, parallel processing	Temporal, sequential processing
Language & Syntax	More complex vocabulary and sentence structure	Simpler, colloquial vocabulary Imprecise sentence structure
Naturalness	Limited extent of nonverbal cues	Great extent of nonverbal cues

Table 1. Key differentiating features of text- and voice-based interaction.

Against the backdrop of technological advancements, established theories and previously made comparisons between text and voice-based interaction might only hold to a certain extent. In that sense, both text- and voice-based interfaces enable a more dynamic, one-to-one interaction between users and machines as compared to non-conversational digital interfaces based on 1) a dialogue flow that mimics human-to-human communication through sequential turn-taking 2) and enriched language with nonverbal, social cues (Hildebrand & Bergner, 2020). State-of-the-art text-based conversational interfaces, such as chatbots, also offer nonverbal cues, including reading and typing delays, typing indicators or emojis, which ultimately contribute to the naturalness of an interaction (Schuetzler et al., 2021). However, auditory cues enhance characteristics of natural interaction, as well as arouse mental simulation of contextual aspects other modalities are incapable of (Griffin et al., 2018). In the following, these differentiating properties of nonverbal auditory cues are further delineated.

2.2 Voice Design

Voice design is studied in human-robot interaction, psychology, phonetics and linguistics and has taken a closer look at auditory cues, such as speed rate or pitch, to understand how such cues and their respective modification affect user perception and behaviour (Chidambaram et al., 2012; Elkins & Derrick, 2013). In this paper, we are concerned with such non-linguistic acoustics (how vocal sounds and speech is heard), as opposed to verbal content and style. We explicitly exclude the latter, which refers to the literal meaning of a message and how this message is expressed in terms of lexical diversity (Feine et al., 2019).

Sutton et al.'s (2019) framework on socio-phonetic design cues considers how social cues can be derived from voice. In their framework, voice is viewed as a combination of social and physical properties. They argue that current voice systems are homogenised in terms of voice output (i.e., gender, accent), as well as that voices and related cues shape user perceptions and experiences. However, non-human sounding voices are not considered as part of their paper. Murad, Munteanu, Cowan and Clark (2019) offer design guidelines for speech interaction by reviewing a set of existing guidelines for user interfaces and extending these, yet only in consideration of mobile voice interfaces. Taking a contextual perspective, another framework only considers embodied voice interfaces (Cambre and Kulkarni, 2019). While a few voice-based agent classifications already exist, established categorisations in the field of IS view voice as one of several modalities without providing a more refined differentiation of individual auditory cues. As part of their taxonomy on pedagogical agents, Wellnhammer, Dolata, Steigler and Schwabe (2020) illustrate that agent voice can differ among a classic text-to-speech engine (TTS), a modern TTS engine, and a human voice. Beyond the type of voice, however, no further modifications of auditory cues are discussed. Feine, Gnewuch, Morana and Maedche's (2019) taxonomy on conversational agents provides a more nuanced understanding of different voice-relevant cue categories, such as a differentiation between voice qualities (i.e., volume, pitch) and vocalizations (i.e., laughing). Nevertheless, their taxonomy does not include a review of related effects on users, which makes it difficult to derive an understanding of related interaction design implications.

Overall, previous categorisations exhibit several gaps. As mentioned by Cambre and Kulkarni (2019), existing reviews lack a mapping of VA-relevant voice cues against user outcomes and their existing exploration in research contexts. To overcome gaps in existing literature, our study suggests the development of a categorisation of VA-relevant design cues 1) that covers auditory cues prevalent and introduced through the deployment of synthetic voices, 2) that considers different types of VAs, 3) and that derives prevalent design strategies for VAs to be considered in the future.

3 Methodology

The following section lays out the chosen methodology to review existing research on auditory cues relevant to voice assistants. First, a systematic literature review (SLR) on voice-based conversational agents was conducted, following established guidelines (Webster & Watson, 2002). An overview of the literature search strategy is given in Figure 1.

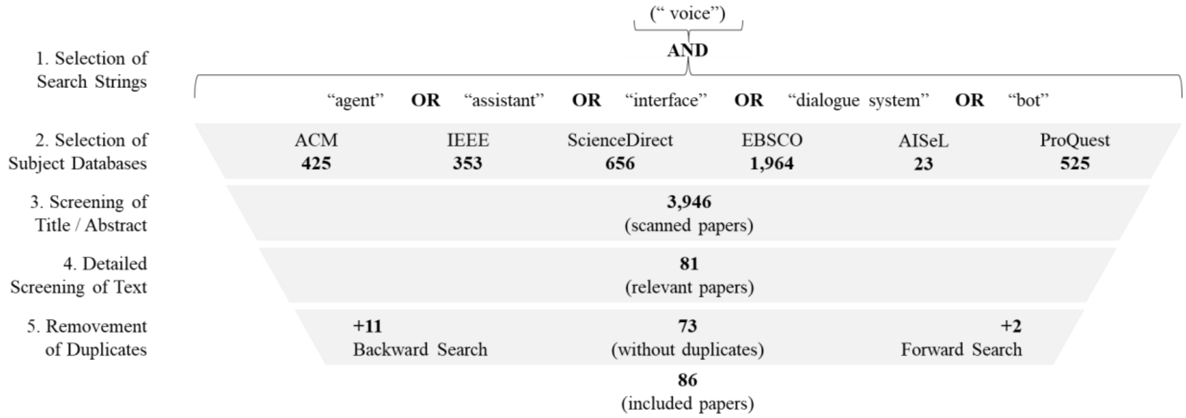
1. Selection of Search Strings
(" voice")
AND
"agent" OR "assistant" OR "interface" OR "dialogue system" OR "bot"
2. Selection of Subject Databases
ACM 425
IEEE 353
ScienceDirect 656
EBSCO 1,964
AISeL 23
ProQuest 525
3. Screening of Title / Abstract
3,946 (scanned papers)
4. Detailed Screening of Text
81 (relevant papers)
5. Removement of Duplicates
+11 Backward Search
73 (without duplicates)
+2 Forward Search
86 (included papers)

Figure 1. Systematic literature search and screen.

Six databases were selected, namely EBSCOhost, IEEE, ACM, ScienceDirect, ProQuest and AISeL. To account for synonyms and varying conceptualizations of conversational agents (e.g., personal assistant, virtual assistant, interactive artificial agent, interface, bot, assistive robot) and vocal cues (e.g., voice-activated, voice-based, speech-based, spoken language, spoken dialogue, conversational), an exploratory search in open access and specific subject databases was conducted to build a search string. An initial review of around 30 papers ensured that we identified most common definitions and synonyms for VAs. This was particularly important as the preliminary search exhibited great fragmentation across research fields that connotated VAs differently. We stopped looking at additional papers once connotations for voice assistants started to reappear with additional papers sighted. The search string (*(voice) AND (agent OR assistant OR interface OR dialogue system OR bot)*) was employed, resulting in 3,946 search results. The database searches were limited to include abstracts, titles, or keywords of relevant publications only. Publications were assessed with respect to the following inclusion criteria: All publications needed to 1) refer to a VA that could either be based on a deterministic or probabilistic model, 2) analyse vocal cues that had an effect on the user (perceptual or behavioural), 3) be peer-reviewed and written in English, 4) be published in a recognized academic journal or reputable conference and 5) be published in between 2000 and 2020. After removing all duplicates, a forward and backward search was lastly conducted to identify further significant contributions to voice-based IT systems, i.e., also outside of the chosen date range. A total of 86 papers are considered relevant to this study.

The 86 selected papers were analysed from a concept-centric perspective. A concept matrix was created based on the analysis of literature search results (Webster & Watson, 2002; Wolfswinkel et al., 2013) in order to derive distinct perspectives on voice cues in the context of conversational agents, which represent the unit of analysis of this study. For each selected paper, core concepts, research method, units of analysis including relevant independent, dependent variables and their relationships, were identified. Moreover, contextual criteria such as the application domain and task of the VA were captured. Findings were synthesised into central research perspectives that amount to vocal design features of VAs. Process steps were iteratively validated by discussion among three researchers to ensure stable, valid and reproducible results. For the systematic review, intercoder reliability was checked by letting three authors code ten selected papers based on an initial coding scheme drafted by the first author. Differences in coding allowed to refine the coding scheme and come to a common understanding of the research focus among the authors. In a similar vein, an initial framework for relevant voice cues was created by the first author. In a subsequent step, the co-authors were asked to place a selected number of papers in this framework. The framework was iteratively refined over three rounds until authors agreed on the classification of individual papers within the framework. The feedback of all authors helped to improve the systematic review since they have numerous years of research experience in the field of conversational agents. Nevertheless, a subjective bias cannot be fully excluded as the process of deriving concepts requires individual judgement.

4 Results

Building upon the systematic literature review, the following subsections address this study's research question by first understanding voice as a type of modality and by subsequently exploring individual voice cues relevant to the design of VAs. Before doing so, an overview of the selected literature is given, touching upon temporal and contextual research trends, as well as prevalent domains covered.

Over the analysed time period ranging from 2000 to 2020, the number of publications has been growing steadily. A steep increase in relevant publications can be detected from 2018 onwards, with half of all reviewed studies being published since then. These findings coincide with the recent market penetration of smart speakers such as *Amazon Alexa* and *Google Assistant* (Smith, 2020). This development supports our initial assumptions that research on VAs is an emerging and to be further developed research field. The commercial availability of VAs for personal use also becomes apparent when reviewing domains covered. More than a fourth (28%) of all included studies focus on general purpose VAs, which were mainly explored in the context of daily activities and tasks. A number of publications have studied the implications and suitability of VAs for specific target groups, partially as part of ethnographic home studies, including children (Aeschlimann et al., 2020), elderly (Straßmann et al., 2020) or disabled (Masina et al., 2020). However, a few studies did not further specify the usage context or analysed voice cues in simulated task settings and lab environments. Based on nine archetypical purposes of conversational agents by Scarpellini and Lim (2020), we further identified the respective role of each VA reviewed. While general purpose, standalone VAs used in home settings mostly assist with or perform pragmatic tasks, VAs in the Education, Healthcare, and Service domains exhibit a focus on the need to be reliable and explicative, as well as consider users' preferences.

Domain	**Task**
Automotive (11)	Assist (4), Bond (1), Control (2), Educate (1), Entertain (1), Guide (1), Inform (1)
Collaboration (2)	Guide (2)
E-Commerce (11)	Advise (8), Inform (3)
Education (12)	Coach (1), Educate (5), Entertain (2), Inform (4)
General (25)	Advise (1), Assist (9), Control (2), Entertain (1), Guide (5), Inform (2), Unspecified (5)
Healthcare (11)	Advise (4), Assist (1), Coach (2), Control (1), Guide (2), Inform (1)
Leisure (5)	Assist (1), Bond (1), Entertain (1), Guide (1), Inform (1)

Table 2. Publications by domain and respective task.

The systematic literature review revealed two main perspectives on the current understanding and exploration of VAs and their impact on user perceptions and behavioural interaction. First, literature on interface modalities (e.g., text or visual cues) view voice on a meta-level and as a design strategy next to other types of modalities. A second stream of literature explores individual vocal cues, such as gender or volume of voice. While our analysis predominantly focuses on the current knowledge base of auditory cues relevant to VAs, we see the first literature stream as an important entry point to our analysis. More specifically, by illustrating the variability in user and interaction outcomes for different tasks and combinations of modality-mediated interaction, relevant strategies for VA design and critical aspects to consider in the subsequent analysis can be identified.

4.1 Voice in Intermodal Comparison

In what technological contexts does the inclusion or the isolated deployment of voice enhance or even outperform other forms of VA interaction? One theme that emerged from the literature was a comparison of different modalities and on how the presence (or absence) of voice affected perceptions of, responses to and interaction with a (multimodal) agent. Hess, Fuller and Campbell's (2009) study on a website's recommendation agent found that *including a voice feature* next to features of text alone or text and human animation, led to greater perceptions of social presence and hence trust. In another retailing experiment where the conversational agent supported information requests by shoppers, a touch-based interface led to higher user engagement with the platform as compared to an interface that additionally deployed voice. In contrast, the voice-plus-touch interface weakened the relationship between platform engagement and brand trust. The authors argue that the additional voice feature makes the interface unnecessarily complex for the task at hand (Pagani et al., 2019).

A number of studies have shown that *voice modalities on their own* perform worse than other modalities of text, robotic face and body movements (Schlögl et al., 2013; Shi et al., 2018). Chidambaram, Chiang and Mutlu (2012) find that, while both nonverbal bodily and vocal cues lead to high user compliance with a robot's suggestions, bodily cues of the robot are perceived as more persuasive than vocal cues. In a study on learning outcomes, a fully embodied agent including gaze and gesture outperformed a voice-only agent (Lusk and Atkinson, 2007). An experiment with the health and nutrition assistant GRETA demonstrated that GRETA interacting in the absence of text was perceived as less likeable. However, in the voice-only interaction, users could more easily memorize the information GRETA provided and trusted the robot more (Berry et al., 2005). In a similar vein, in a setting of foreign language learning, the agent's voice, operating here as an educational instructor, contributed more positively to learning outcome than agent movement, gestures and pointing (Carlotto & Jaques, 2016).

It appears that the deployment of voice-based assistants and the respective effect on user and interaction must be viewed *embedded in the use context and task* at hand, which is touched upon more elaborately later in this paper. In line with this reasoning, Cho, Molina and Wang (2019) found that in the context of utilitarian tasks, voice leads to more positive attitudes towards the agent, which is mediated by perceived human likeness of the agent. Participants of their study were asked to interact with Microsoft's voice agent Cortana involving two types of tasks, namely one utilitarian and one hedonic. As part of a mixed factorial experiment, modality (text versus voice) and device (mobile versus laptop) represent two within-subject factors of their study. Whether and how the nature of tasks interacts with the influence of voice is still relatively unexplored in the context of conversational agents. Interestingly, Qiu and Benbasat (2005) explored the effect of different modalities in a customer service context where the agent supported users with their product recommendations. Among the six conditions combining the modalities text, voice and animation, the text-to-speech voice of the agent influenced emotional trust, as well as cognitive trust in the agent's competence.

Ultimately, voice is one of several *interaction modalities*, depending on a number of different variables, including context, user and task at hand. The choice for inclusion of voice or voice-only also seems to be a question of *habitude and experience*: In an ethnographic study with elderly using personal assistants at home, an affinity towards text-based communication could be identified (Schlögl et al., 2013). In the past, voice as a design element has rather been an additional or complementary feature to existing interfaces and information systems. Nowadays, an increased deployment of voice as a stand-alone and disembodied feature can be experienced. With the prevalence of voice-only conversational agents in our personal homes, business settings, and service frontlines, this study takes a closer look at individual cues relevant to voice design.

4.2 Voice Cues and Voice Design

What concrete features must researchers and designers consider when designing voice-based interfaces? More so, how should such features be modified when designing for a specific context or interaction? In a second step of this paper's analysis, we overcome the shortcomings of previously touched upon categorisations and taxonomies by including voice-related findings on user impact. More specifically, our aim is to examine the role of individual voice cues on user perception and behavioural outcomes. We consider auditory cues of VAs without limiting the considered literature to a certain device, task or embodiment of the agent. A combination of several vocal cues by Feine et al.

(2019) is taken as a point of departure for organizing a categorisation of vocal cues and their effect on user perception and behaviour, which includes auditory voice qualities and vocalizations, as well as chronemics. While Feine et al.'s (2019) taxonomy focuses on different types of conversational agents, their consideration of a more nuanced variation in vocal cues as compared to other taxonomies in the field of IS, is deemed suitable for our study. As presented in Table 2, our categorisation of auditory features is one way to view the analysis of different types of voice-relevant design cues, yet the categories are not intended to be collectively exhaustive. Rather, the categorisation focuses only on voice cues that have been explicitly explored and manipulated in existing research on VAs.

		Voice Cue	Acoustic Operationalization	User Perception of VA	Interaction Outcomes with VA	Selected Tasks of VA	Selected Research
Auditory Cues	Voice Qualities	Loudness	Amplitude [dB]	↑Fairness, ↑Dominance ↑Credibility, ↑Trust ↑Acceptance	↑Negotiation, ↑Cooperation	Negotiate price offers, play a quiz	Mania et al. (2020), Nass & Lee (2001), Tay, Jung & Park (2014), Tielman et al. (2014)
		Rate of Speech	Number of words per minute / second	↑Credibility, ↑Trust, ↑Acceptance	↑Cooperation	Rate book reviews, control blood measurement	Nass & Lee (2001), Tay, Jung & Park (2014), Tielman et al. (2014)
		Pitch	Fundamental Frequency (F0) and Formant Frequency (Df)	↑Trust, ↓Trust, ↑Credibility, ↑Robot Appeal, ↑Acceptance, ↑Enjoyment, ↑Interaction Quality, VA Gender & Emotion	↑Cooperation	Guide security check for travel, advise money investment	Elkins & Derrick (2013), McGinn & Torre (2019), Nass & Lee (2001), Tay, Jung & Park (2014), Torre et al. (2020), Walters et al. (2008)
	Chronemis	Pausing / Latency	Spectral Centroid Response Latency Voice Breaks		↓Speech Collisions	Guide hotel room reservation	Funakoshi et al. (2010)
		Filled Pauses	Backchannel / Fillers	↑Naturalness	↑Progressivity	Assist in setting a quiz, chat about the news	Fischer et al. (2019), Marge et al. (2010)
VA-Specific Cues		Synthetic Voice	SSML or TTS (as compared to human voice)	↑ Trust, ↑Preference, ↓ Enjoyment, ↓Naturalness, ↓ Social presence	↑Learning, ↓Learning ↑Engagement, ↑Training efficiency, ↓Confirmation to suggestions, ↑Physical distance to robot	Educate & test on subject material, advise on product choice	Craig & Schroeder (2017), Gravano & Hirschberg (2011), Mayer & DaPra (2012), Marge et al. (2010), McGinn & Torre (2019), Qui & Benbasat (2005), Qui & Benbasat (2009), Tamagawa et al. (2011), Walters et al. (2008)
			Gender (matched to user)	↑Trust, ↑Comprehension, ↑Liking, ↑Perceived ease of use, ↑Perceived usefulness	Robot Appearance Selection	Answer general purpose requests, practice driving skills	Chang, Lu & Yang (2018), Cowan et al. (2016), Ernst & Herm-Stapelberg (2020), Lee, Ratan & Park (2019), Louwerse et al. (2005), McGinn & Torre (2019)
			Accent (matched to user)	↑Naturalness, ↑Preference, ↑Trust		Collaborate on picture task	Cowan et al. (2016), Tamagawa et al. (2011)

Note: User Perceptions and Interaction Outcome findings associated with presence or an increase of voice cue at hand.
E.g. Marge et al. (2010) found that the use of filled pauses increased perceived naturalness of interaction

Table 3. Relevant voice cues along user perceptions and interaction outcomes.

Auditory cues refer to features of voice and utterance which may not be encoded by phonetic segments, grammar or choice of vocabulary. They include the loudness or volume, the speech rate, and pitch of a voice. Each of these cues can be acoustically operationalized through different measures. Loudness, for instance, can be measured through the amplitude as measured in decibels, while speech rate can be adapted by modifying the number of words spoken per minute. Tweaking prosodic voice cues is important as they can influence users' perceptions of a VA with regard to its credibility, appeal or acceptance (Nass & Lee, 2001; Sutton, 2020; Tay et al., 2014). Second, such modification can evoke associations of personality (i.e., gender) or the emotional state of a voice. Torre, Goslin and White (2020) identified a higher formant frequency (measured as the number of cycles of a soundwave per second of vibrating air particles) and higher spectral centroid (which refers to how energy is distributed across different frequency bands) as automated measures of pitch and acoustic correlates of a smiling voice. Deploying this voice-based robot in an investment task found an increase in trust resulting from a higher pitch. Interestingly, an increased pitch has been associated with a decrease in trust in an interview setting with an embodied conversational agent (Elkins & Derrick, 2013).

As part of auditory cues, *chronemics* are timing-related cues in communication and hence related to turn-taking (Feine et al., 2019). Turn-taking is crucial to the design of voice-based interactions since it defines the flow and progressivity of an interaction (Fischer et al., 2019). Gravano and Hirschberg (2011) identified seven turn-yielding cues from human-to-human interaction which are also automatically computable. Backchannels, for instance, are short expressions such as "uh-hu" or "mm-hm" which convey that the listener is paying attention and encourage the speaker to continue talking. Response latency is another turn-taking metric, often measured as reply speed or the time in between two conversation turns. Funakoshi et al. (2010) found that non-humanlike turn-taking (measured as slower than average reply speed in milliseconds) resulted in less speech collisions and hence proves as suitable for an interaction with a spoken dialogue system. Contrary to Funakoshi et al.'s (2010) study in the context of a hotel room reservation task, Marge, Miranda, Black and Rudnicky (2010) explore the effect of voice breaks and filled pauses on perceived naturalness in a social conversation with a spoken dialogue system. They find that the inclusion of such pauses increases the perceived naturalness of the interaction and can compensate for a lack of logical flow.

A second dimension titled *VA-specific cues* relates to cues that have been identified as either newly arising features in the design of VAs or predominant features which cannot be further broken down into any of the previously mentioned cues and which can be directly manipulated with a synthesized or human recorded voice. Synthesized voice, as opposed to recorded human voice, enables VAs to be matched to certain tasks and contexts. Platforms offer several synthesized text-to-speech (TTS) voice libraries and Speech Synthesis Markup Language (SSML) for further personalization (Branham & Roy, 2019). Information and content can be fed through written dialogue or spoken corpora (Wellnhammer et al., 2020). While the deployment of a synthetic voice as compared to recorded human voice has been associated with less naturalness and social presence, numerous studies have shown a significant positive effect on trust and preference. More so, in learning contexts, a synthetic voice has proven to positively affect learning transfer, training efficiency and engagement (Craig and Schroeder, 2017; Komiak and Benbasat, 2003; Qiu and Benbasat, 2005; Tamagawa et al., 2011).

Based on the analysis and mapping of voice cues and related user outcomes, we identify three dominant design themes the reviewed literature. These findings are in line with the two voice design strategies mentioned by Sutton, Foulkes, Kirk and Lawson (2019), referring to the individualisation of voice output, and context awareness when designing VAs. While the two design goals have been introduced focusing on sociophonetic aspects of VAs only, we extend these respective design strategies to all auditory cues identified in this review. More so, we expand this previous work in suggesting a third relevant design strategy, namely the impersonation, or personification of VAs. It is important to note here that these output strategies should not be viewed as mutually exclusive, collectively exhaustive nor ultimately desirable for VA design. Rather, we see these three design strategies as a way to structure the prevalent themes found in the reviewed literature and as topics around which previous, related design choices should be challenged.

5 Strategies for Voice Design

5.1 Personification of Voice Assistants

In line with Media Equation Theory, humans do not only assign personalities to other humans yet also to machines (Reeves & Nass, 1996). Our categorisation of voice cues and related mapping of user perceptions is congruent with this theory, demonstrating that both individual voice cues, as well as the combination of cues translate into certain social cues and associations with the VA to create a "voice personality". Aiming to establish more natural interactions, existing studies predominantly assess the human likeness of VAs. Last, the deployment of a "default" voice has also been found as a prevalent voice design strategy and is criticized by numerous papers, pointing towards a lack of conscious and embedded voice design.

In fact, Völkel et al. (2020) developed a *personality model* for speech-based conversational agents. Based on an aggregated set of 349 adjectives, they developed a set of synonym clusters to describe the personality of a VA. Their analysis was based on an examination of commercially available VAs, including Google Assistant, Cortana and Alexa. More so, their study demonstrates how VAs are ascribed certain personality traits and that perceptions of agent personality can be shaped deliberately. Interestingly, they propose an initial set of VA personality dimensions, arguing that existing models to describe human personality, such as the Big Five, are inadequate for VAs. The research interest in VA anthropomorphism becomes apparent by considering the sheer number of studies exploring the humanness and naturalness of individual voice cues. Existing research has demonstrated that the imitation of voice cues found in human-to-human interaction, such as response latency, can increase the perceived naturalness of VAs (Marge et al., 2010). However, the investigation of individual acoustic measures and their tweaking towards greater human likeness should be further researched.

Chang, Lu and Yang (2018), who compared eight synthetic voices differing in gender, pitch, speech rate and intonation, not only found that the adoption of voice characteristics to user preferences might perform better than a "one size fits all" approach, but that the association and combination of various cues created certain *personality trait associations* with the social robots. In a similar vein, Mania et al. (2020) modified pitch, amplitude and intonation to create impressions of vocal dominance of the VA. While the modulation of vocal cues was sufficient in conveying different degrees of dominance, a related difference in negotiation action was only present for participants who perceive themselves as competitive negotiators. In an e-commerce setting, Nass and Lee (2001) manipulated voice with regard to four parameters (volume, fundamental frequency, frequency range, and speed rate) to create extroverted and introverted voice personalities. The personality conveyed by the voice had a main effect on user perceptions: Participants who listened to the same book review rated those differently on liking and credibility based on hearing these reviews in different voices. To maximize liking and trust, the authors argue, voice cues should be set in such a way which creates a personality that is consistent with the user and the content being presented, thus shaping intimate service interactions. Modifying a VA's extraversion by twitching multiple voice cues has been a common method found in several studies, such as Tay et al.'s (2014) experiment in healthcare, as well as Hess, Fuller and Campbell (2009) study on multimedia vividness. The latter modified speed rate, frequency range, pitch and volume to evoke an extroverted or introverted agent personality, showing that such personality difference influences social presence and ultimately trust in the VA. In a more recent study, increased formant frequency and spectral centroid were used as acoustic correlates of a smiling robot voice (Torre, Goslin, et al., 2020). Not only did participants rate "smiling voices" as happier but trusted those more.

A number of researchers have criticized a default set of voice parameters, as commonly chosen by commercial interfaces (Cambre & Kulkarni, 2019; Torre et al., 2018). By relying on a predefined set of design parameters, a certain voice personality (mostly female) is deployed by default without considering the effects and suitability of this voice character for the use case at hand. Hwang, Oh, Lee and Lee (2019) have reviewed 1,602 responses from five South Korean VAs, finding that female voices exist as the only default voice setting, ultimately enhancing stereotypical femininity. In fact, the deployment of a female voice as the default gender has been associated with increased likeability and preference (Chang et al., 2018; Ernst & Herm-Stapelberg, 2020). Technologoical companies are slowly starting to offer several voice options instead of a default female voice, such as *Apple's Siri* (Khaled, 2021). Beyond a default voice, commercial developers are currently exploring the vast opportunities of voice modification to deliver a natural experience and facilitated interaction flow, including a "One-Breath Test" to guide speech rate of synthetic voice or the design of pauses in-between a certain number of words (Branham and Roy, 2019).

So far, little research has taken into consideration adverse and unintended effects of personification, as well as the context-sensitivity regarding the effectiveness of personification. Raising the human-likeness of VAs may be beneficial in relational interaction contexts, yet could impede application contexts that require anonymity, i.e., where sensitive data is shared. As proposed by the Uncanny Valley, the notion of an anthropomorphized, yet not fully human VA might evoke feelings of eeriness as a high human-like appearance creates expectations which are not matched by the agent's capabilities (Mori, MacDorman and Kageki, 2012). Anthropomorphizing VAs through voice cues can lead to frustration, especially if technical functionalities are not ensured or personification is unexpected. In a similar vein, researchers have criticized the impersonation of VAs to an extent where users no longer can distinguish the machine voice from a human voice, thus deceiving the real identity of a VA (O'Leary, 2019).

5.2 Individualization of VA Interaction

The previously raised inappropriateness of a default design approach for VAs points towards another key finding of this study's review: The importance of tailoring a VA towards users and their individual personality, preferences and expectations. In a real-world driving study, VA characters were trusted more and rated as significantly more likeable as compared to a default voice character that matched the driver's personality (Braun et al., 2019). Traits of the characterized VAs were expressed through both choice of words and intonation. Their study even shows that a mismatched assistant personality is perceived as much less useful and satisfying than the default VA. In a similar vein, matching system accent to users' nationality has been studied in several papers, exhibiting increased levels of naturalness, preference and trust (Cowan et al., 2016; Tamagawa et al., 2011). Zepf, Gupta, Krämer and Minker (2020) explore acoustical mimicry as a strategy to improve empathy within automotive user interfaces and demonstrate a positive effect on perceived empathy, personalization and naturalness without affecting the efficiency of the driving task. In a study by Nguyen, Ta and Prybutok (2019), female and male participants exhibit different factors that drive attitudes toward VA use and adoption, illustrating how gender presents an important user characteristic to consider when designing for system adoption and use. Adapting towards users or even mirroring their demographics and behaviour has been explored as a potential design strategy for voice-based interactions. Eyssel et al. (2012) underline this hypothesis with their study, finding that same-gender robots are perceived more positively, are associated with more psychological closeness and are more strongly anthropomorphized. Significant differences in perceived pleasantness and VA gender preference between male and female participants have also been found in Obinali's (2019) study on information acceptance of voice output by *Apple's Siri.* Even in the context of a TTS voice, matching the agent voice's gender towards the user at hand reduces cognitive demands while driving and makes the interface more effective to use (Truschin et al., 2014).

Turning towards the intelligible nature of state-of-the-art VAs, Lubold, Walker and Pon-Barry (2020) find that VAs adopting their pitch towards the user, in combination with social dialogue, can lead to higher measures of rapport. In commercial cases, so-called *tapering* is used as a strategy to adopt the amount of detail provided in each interaction and reducing it in line with the number of times a service has been used (Branham & Roy, 2019). In addition, speech rate and timeout periods have been considered as computable voice features to be tailored towards certain target populations. As visually impaired people can usually process voice much faster, design papers have raised the need for speech rate to be adaptable. In a similar vein, elderly or disabled humans may take longer to formulate a command. Hence, such personal requirements translate into a need for longer timeout periods of VAs (Branham & Roy, 2019). However, the current research field is lacking an empirical exploration of the effects of individual design cues of such commercially available VAs (Knote et al., 2019). Beyond qualitative document reviews of commercial design guidelines, little research can be found on VA-specific cues.

An individualized VA interaction enables novel opportunities for minorities, i.e., by adapting to the needs of elderlies or tailoring an interaction towards a specific accent and thus reaching previously neglected user groups (Sutton et al., 2019). However, this room for adaptation also leaves room for the manifestation of stereotypes, especially if the choice of design cues is based on stereotypical functionalities and human bias. For instance, the design of a VA's gender has shown that a male voice is deployed for security tasks (Trovato et al., 2017), whereas standalone VAs deployed for home use are female (Carpenter et al., 2009). Beyond designing for increased likability and adoption of VAs, designers should consider how certain design choices manifest or even enhance stereotypes.

5.3 Context Awareness in VA Interaction

The importance of designing voice "in relation to the contexts in which it is used" (Sutton et al., 2019, p. 7) has been raised by a number of researchers. In 2006 already, Mutlu et al. argued that voice design should be viewed as an interplay among user characteristics, device and context. Accordingly, the previous two design output strategies cannot be viewed isolated from this third one, context awareness. Relevant aspects that have already been touched upon with regard to contextualization include the consideration of geographical accent (Cowan et al., 2016; Tamagawa et al., 2011), yet extend to social qualities a voice is expected to embody in a certain use case. The previous discussion around adaption towards socio-demographical functionalities can already be viewed as a form of contextualization. When tailoring voice towards an activity or task, these expected qualities differ from interaction to interaction. In the following, we therefore refer to a specific task or domain when talking about context.

In a study by Torre, Latupeirissa and McGinn (2020), participants were asked to match different voices to images of robots. A predefined number of contexts consisted of a home, hospital, restaurant and school setting each, and were added as background noise to the robot voice. Interestingly, different robot images were chosen for different contexts despite other vocal features (naturalness, gender, accent) remaining the same. The findings illustrate that the effect of individual voice cues interact with the context of an agent's deployment. Supporting this line of reasoning, a study by Tay et al. (2014) explored the impact of a VA's gender and personality in two different occupational settings, namely a healthcare and a security scenario. The study demonstrated that VA gender and personality do not monotonically influence user acceptance, yet rather interact with corresponding role stereotypes of certain occupations in their effect on user acceptance (Tay et al., 2014). Sensitivity in voice design also becomes relevant when a VA is deployed in contexts of high cognitive load, such as a driving scenario. When safety is of utmost importance, creating high perceptions of VA trustworthiness or naturalness might not be the desired design goal. More specifically, Wong, Brumby, Babu and Kobayashi (2019) found that a more assertive VA alerting the driver to driving events resulted in faster reaction times and was perceived as more urgent.

Turning towards commercial design elements, Tabassum et al. (2019) found that users would be comfortable with an always listening function when sharing non-sensitive or non-personal information. It can be assumed that preferences and perceptions about voice design and individual design elements vary per use case. While the personification of VAs through voice design cues and the tailoring towards users has already been empirically investigated by a number of studies, research considering contextual granularities, such as the differentiation between relational and transactional tasks of VA deployment, is quite scarce. More so, the previously mentioned studies show that voice cues alone do not affect of how a VA should be designed but rather interact with the context of the VA's deployment.

6 A Research Agenda for Voice-Based Interaction

As part of this section, we discuss the contributions of this study's review and provide a research agenda. We see the importance of progressing the understanding of auditory cues for VAs we arrived at as part of this study by further exploring seemingly contradictory findings as well as extending our categorisation towards a more nuanced and interlinked refinement of relevant VA design cues.

Personification of VAs: Despite a great number of studies exploring the personification, or anthropomorphising, of VAs, mixed results have been found regarding individual dependent variables, such as perceptions of trust and naturalness. These contradicting findings might be due to differing expectations towards a VAs across contexts and tasks. Design strategies in human-human interaction or previous human-computer interaction do not necessarily hold for interactions with VAs, especially in varying contexts and, hence, task requirements. For instance, non-humanlike spoken dialogue design has proven successful for reducing speech collisions when interacting with a VA (Funakoshi et al., 2010). These results should be corroborated by further studies and replicated in different contexts or with different target groups to understand in what interactions personification is, for one, socially and ethically desirable, and second, effective regarding the interaction goals from both user and provider perspective. Building upon such empirical testing, future research should provide a more nuanced understanding of when to follow principles for human-like spoken dialogue.

Individualization of VA Interaction: Reviewed literature has demonstrated that a VA design matching and considering individual user personality yield desirable outcomes with regard to user perception and interaction outcomes as compared to a default, “one size fits all” approach and allow to better cater the needs of previously neglected target groups, such as elderly (Torre et al., 2018; Cambre and Kulkarni, 2019). Besides privacy issues (Dickhaut et al., 2021), the matching of user personality, expectations and preferences might bring along (potentially) unforeseen implications, which have been neglected in research so far. For instance, the matching of gendered voice to corresponding gendered features and occupations presents a double-edged sword as higher interaction success is accomplished while reinforcing existing stereotypes (Tay et al., 2014). Hence, we propose that researchers should explore design strategies for VAs that combine potentially conflicting goals (Dickhaut, E.; Janson, A. & Leimeister, 2020). In other words, future studies should investigate the effect of designing for the elimination or reduction of gender, occupational or geographical stereotypes and related impact on success and perceptions of user interaction (Tolmeijer et al., 2021). The suggested research efforts for both personification and individualization could be tailored towards the specific requirements of encounters by comparing needs and effectiveness of personification and individualization between transactional and relational tasks, for instance (Huang & Rust, 2020).

Context Awareness in VA Interaction: The conducted analysis and research agenda for the previous two design strategies have emphasized that the impact of voice is very much context-dependent. The contributions regarding contextualization of user interaction with a VA are quite scarce according to our understanding derived from this study’s review. Two different studies have shown that depending on task and context, human-like turn taking, for instance, might be more or less suitable. Furthermore, a number of studies did not further specify the usage context and analysed voice cues in simulated task settings and isolated lab environments, which may imply a low external validity of the current body of knowledge that is also reflected in the low number of conducted field studies. A threat to internal validity is posed by the study of multimodal agents which challenges to isolate the effect of voice-based interaction in comparison to other modalities. The prevalence of studies in task-separated, simulated lab environments leaves questions around the suitability and effect of VAs deployed in encounters beyond the individual home unanswered. We propose conducting field studies and collecting “real world” voice data when analysing the implications of voice design for interaction contexts.

Moreover, we have identified some *overarching research opportunities* based on the reviewed literature. We differentiated between auditory cues originating from the linguistics literature and VA-specific auditory cues that arise with the emergence of this novel system class. Thus, we see the future development of two important streams elaborating on this differentiation: First, a holistic understanding of existing vocal cues in human-human interaction, relying upon the understanding and conceptualisations of other research domains such as linguistics, should be developed that maps cues relevant to human-VA interaction and hence demonstrates the key differences between those two types of interactions. Hall, Coats and LeBeau's (2005) categorisation of nonverbal cues offers a starting point to do so. As part of this research effort, our cue categorisation should be extended to linguistic cue s to identify interrelations with speech-related cues or even verbal content other than auditory cues. Second, vocal cues that arise through the deployment of synthetic voice should be further investigated. With SSML offering a novel conversational speaking style for VAs (i.e., whispering), as well as a more granular level of personalization, commercial products and actual deployment might differ starkly from our categorisation of voice-relevant cues. While some preliminary research on commercially available design guidelines exist (Branham & Roy, 2019), we see the development of a comprehensive overview of existing and malleable auditory cues in those commercially available synthetic voices as another important stream for future research. As a first step, a review of such commercially available synthetic voices and related libraries such as Google Wavenet and Amazon Polly and also the ecosystem affordances should be conducted (Knote et al., 2021). Consecutively, identified design cues in commercially available VAs could be mapped against the current overview of features identified as part of this study.

7 Limitations and Conclusion

We consider several limitations of this study. First, the scope of this literature review cannot be considered fully exhaustive. However, to reach a high coverage, a database-driven search strategy was chosen over an outlet-based one to include peer-reviewed conference proceedings and thus most nascent research on VAs. To minimize the risk of subjective bias, concise selection criteria were defined and followed up upon. Coverage may be reduced due to the selection of the initial keyword search that was focused on different terms used for VAs. However, we thoroughly analysed the sizeable amount of identified publications based on full-text analysis using a concept matrix. As we focused on vocal cues which had already been empirically explored in the context of (voice-based) conversational agents, we did not endeavour to provide a categorisation of auditory design cues that is complete in terms of an exhaustive overview of potentially relevant design cues. In that sense, our categorisation is lacking a consideration of more diverse literature from related domains, i.e., linguistics or psychology for a richer understanding of what constitutes voice. As mentioned earlier, the inclusion of more diverse literature could help to further contrast and progress the understanding of how voice in human-VA interaction differs from human-human interaction. Future work may extend and substantiate our analysis by including verbal content and style (e.g., lexical emotional content). The inter-relatedness of the three proposed design strategies should also be further addressed in subsequent research. While it is difficult to discuss any of our findings detached from the contexts of the underlying studies, all analyses and discussions should already be embedded within the notion of contextualization.

Concluding, the holistic evaluation of the progress of the specific research stream on VAs and related design cues is crucial to this study’s consecutive identification of potential research fields and gaps. As a result, recommendations for future research on the design and deployment of VAs is provided. For this purpose, we conducted a systematic literature review to identify, code and analyse empirical findings for VAs and related design features. By discussing the key insights of our review along three design strategies, namely personification, individualization and contextualization, a set of research questions for future research in the context of VAs was developed.

Acknowledgements

We thank the Swiss National Science Foundation for funding parts of this research (100013_192718). The third author also acknowledges funding from the Basic Research Fund (GFF) of the University of St. Gallen.

References

Abdolrahmani, A., Kuber, R., & Branham, S. M. (2018). Siri talks at you: An empirical investigation of voice-activated personal assistant (VAPA) usage by individuals who are blind. *ASSETS 2018 - Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility*, 249–258. https://doi.org/10.1145/3234695.3236344

Aeschlimann, S., Bleiker, M., Wechner, M., & Gampe, A. (2020). Communicative and social consequences of interactions with voice assistants. *Computers in Human Behavior*, *112*, 106466. https://doi.org/10.1016/j.chb.2020.106466

Balasuriya, S. S., Sitbon, L., Bayor, A. A., Hoogstrate, M., & Brereton, M. (2018). Use of voice activated interfaces by people with intellectual disability. *ACM International Conference Proceeding Series*, 102–112. https://doi.org/10.1145/3292147.3292161

Belin, P., Fecteau, S., & Bédard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in Cognitive Sciences*, *8*(3), 129–135. https://doi.org/10.1016/j.tics.2004.01.008

Berry, D. C., Butler, L. T., & De Rosis, F. (2005). Evaluating a realistic agent in an advice-giving task. *International Journal of Human Computer Studies*, *63*(3), 304–327. https://doi.org/10.1016/j.ijhcs.2005.03.006

Branham, S. M., & Roy, A. R. M. (2019). Reading between the guidelines: How commercial voice assistant guidelines hinder accessibility for blind users. *ASSETS 2019 - 21st International ACM SIGACCESS Conference on Computers and Accessibility*, 446–458. https://doi.org/10.1145/3308561.3353797

Braun, M., Mainz, A., Chadowitz, R., Pfleging, B., & Alt, F. (2019). At your service: Designing voice assistant personalities to improve automotive user interfaces a real world driving study. *Conference on Human Factors in Computing Systems - Proceedings*, 1–11. https://doi.org/10.1145/3290605.3300270

Cambre, J., & Kulkarni, C. (2019). One voice fits all? Social implications and research challenges of designing voices for smart devices. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW). https://doi.org/10.1145/3359325

Carlotto, T., & Jaques, P. A. (2016). The effects of animated pedagogical agents in an English-as-a-foreign-language learning environment. *International Journal of Human Computer Studies*, *95*, 15–26. https://doi.org/10.1016/j.ijhcs.2016.06.001

Carpenter, J., Davis, J. M., Erwin-Stewart, N., Lee, T. R., Bransford, J. D., & Vye, N. (2009). Gender representation and humanoid robots designed for domestic use. *International Journal of Social Robotics*, *1*(3), 261–265. https://doi.org/10.1007/s12369-009-0016-4

Chang, R. C. S., Lu, H. P., & Yang, P. (2018). Stereotypes or golden rules? Exploring likable voice traits of social robots as active aging companions for tech-savvy baby boomers in Taiwan. *Computers in Human Behavior*, *84*, 194–210. https://doi.org/10.1016/j.chb.2018.02.025

Chidambaram, V., Chiang, Y. H., & Mutlu, B. (2012). Designing persuasive robots: How robots might persuade people using vocal and nonverbal cues. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 293–300. https://doi.org/10.1145/2157689.2157798

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

Cowan, B. R., Gannon, D., Walsh, J., Kinneen, J., O'keefe, E., & Xie, L. (2016). Towards understanding how speech output affects navigation system credibility. *Conference on Human Factors in Computing Systems - Proceedings*, *07-12-May-*, 2805–2812. https://doi.org/10.1145/2851581.2892469

Craig, S. D., & Schroeder, N. L. (2017). Reconsidering the voice effect when learning from a virtual human. *Computers and Education*, *114*, 193–205. https://doi.org/10.1016/j.compedu.2017.07.003

Dennis, A. R., Fuller, R. M., & Valacich, J. S. (2008). Media, Tasks, and Communication Processes: A Theory of Media Synchronicity. *MIS Quarterly*, *32*(3), 575–600.

Dickhaut, E.; Janson, A. & Leimeister, J. M. (2020). Codifying Interdisciplinary Design Knowledge through Patterns – The Case of Smart Personal Assistants. In *Design Science Research in Information Systems and Technology (DESRIST)* (pp. 114–125). Springer International Publishing.

Dickhaut, E., Li, M., Janson, A., & Leimeister, J. M. (2021). Developing Lawful Technologies – A Revelatory Case Study on Design Patterns. *Proceedings of the 54th Hawaii International Conference on System Sciences*, *54*. https://doi.org/10.24251/hicss.2021.533

Elkins, A. C., & Derrick, D. C. (2013). The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents. *Group Decision and Negotiation*, *22*(5), 897–913. https://doi.org/10.1007/s10726-012-9339-x

Eyssel, F., Kuchenbrandt, D., Bobinger, S., De Ruiter, L., & Hegel, F. (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. *HRI'12 - Proceedings of the 7th Annual ACM/IEEE International Conference on Human-Robot Interaction*, 125–126. https://doi.org/10.1145/2157689.2157717

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

Fischer Stuart Reeves Martin Porcheron, J. E. (2019). *Progressivity for Voice Interface Design Rein Ove Sikveland*. *ii*. https://doi.org/10.1145/3342775.

Funakoshi, K., Nakano, M., Kobayashi, K., Komatsu, T., & Yamada, S. (2010). Non-humanlike spoken dialogue: A design perspective. *Proceedings of the SIGDIAL 2010 Conference: 11th Annual Meeting of the Special Interest Group OnDiscourse and Dialogue*, 176–184.

Gravano, A., & Hirschberg, J. (2011). Turn-taking cues in task-oriented dialogue. *Computer Speech and Language*, *25*(3), 601–634. https://doi.org/10.1016/j.csl.2010.10.003

Griffin, E., Ledbetter, A., & Sparks, G. G. (2018). *A First Look at Communication Theory* (10th ed.). McGraw-Hill Education.

Hall, J. A., Coats, E. J., & LeBeau, L. S. (2005). Nonverbal behavior and the vertical dimension of social relations: A meta-analysis. *Psychological Bulletin*, *131*(6), 898–924. https://doi.org/10.1037/0033-2909.131.6.898

Herm-stapelberg, N., Mainz, J. G., Herm-stapelberg, N., & Mainz, J. G. (2020). *AIS Electronic Library (AISeL) The Impact of Gender Stereotyping on the Perceived Likability of Virtual Assistants The Impact of Gender Stereotyping on the Perceived Likability of Virtual Assistants*. 0–7.

Hess, T., Fuller, M., & Campbell, D. (2009). Journal of the Association for Information Systems Designing Interfaces with Social Presence : Using Vividness and Extraversion to Create Social Recommendation Agents * Designing Interfaces with Social Presence : Using Vividness and Extraversion to Create. *Journal of the Association for Information Systems*, *10*(12), 889–919.

Hildebrand, C., & Bergner, A. (2020). Conversational Robo Advisors as Surrogates of Trust: Onboarding Experience, Firm Perception, and Consumer Financial Decision Making. *Journal of the Academy of Marketing Science, Forthcoming*.

Huang, M. H., & Rust, R. T. (2020). Engaged to a Robot? The Role of AI in Service. *Journal of Service Research*. https://doi.org/10.1177/1094670520902266

Hwang, G., Oh, C. Y., Lee, J., & Lee, J. (2019). It sounds like a woman: Exploring gender stereotypes in South Korean voice assistants. *Conference on Human Factors in Computing Systems - Proceedings*, 1–6. https://doi.org/10.1145/3290607.3312915

Khaled, F. (2021, April 1). Apple removes Siri's female voice as its default and adds two new voices. *Business Insider*. https://www.businessinsider.com/apple-removes-siri-female-voice-as-its-default-2021-4?r=US&IR=T#:~:text=Siri currently defaults to a,African%2C Irish%2C and

Australian.

Knote, R., Janson, A., Söllner, M., & Leimeister, J. M. (2019). Classifying Smart Personal Assistants: An Empirical Cluster Analysis. *Proceedings of the 52nd Hawaii International Conference on System Sciences*, *6*, 2024–2033. https://doi.org/10.24251/hicss.2019.245

Knote, R., Janson, A., Söllner, M., & Leimeister, J. M. (2021). Value co-creation in smart services: A functional affordances perspective on smart personal assistants. *Journal of the Association for Information Systems*, *22*(2), 418–458. https://doi.org/10.17705/1jais.00667

Komiak, S. X., & Benbasat, I. (2003). Understanding Customer Trust in Agent-Mediated Electronic Commerce, Web-Mediated Electronic Commerce, and Traditional Commerce. *Information Technology and Management*, *5*(1/2), 181–207. https://doi.org/10.1023/b:item.0000008081.55563.d4

Low, C. (2020, September). Alexa will seem more human with breathing pauses and learning skills. *Endgadget*. https://www.engadget.com/amazon-2020-alexa-breathing-teach-voice-profiles-for-kids-172918631.html

Lubold, N., Walker, E., & Pon-Barry, H. (2020). Effects of adapting to user pitch on rapport perception, behavior, and state with a social robotic learning companion. In *User Modeling and User-Adapted Interaction* (Issue 0123456789). Springer Netherlands. https://doi.org/10.1007/s11257-020-09267-3

Lusk, M. M., & Atkinson, R. K. (2007). Animated Pedagogical Agents: Does Their Degree of Embodiement Impact Learning from Static or Animated Worked Examples? *Applied Cognitive Psychology*, *21*, 747–764.

Mania, J., Miedema, F., Browne, R., Broekens, J., & Oertel, C. (2020). *Towards Understanding the Effect of Voice on Human-Agent Negotiation*. 1–8. https://doi.org/10.1145/3383652.3423896

Marge, M., Miranda, J., Black, A. W., & Rudnicky, A. I. (2010). Towards improving the naturalness of social conversations with dialogue systems. *Proceedings of the SIGDIAL 2010 Conference: 11th Annual Meeting of the Special Interest Group OnDiscourse and Dialogue*, 91–94.

Masina, F., Orso, V., Pluchino, P., Dainese, G., Volpato, S., Nelini, C., Mapelli, D., Spagnolli, A., & Gamberini, L. (2020). Investigating the accessibility of voice assistants with impaired users: Mixed methods study. *Journal of Medical Internet Research*, *22*(9). https://doi.org/10.2196/18431

McTear, M. F. (2017). The Rise of the Conversational Interface: A New Kid on the Block? In *Future and Emerging Trends in Language Technology* (pp. 38–49). Springer International Publishing.

Murad, C., Munteanu, C., Cowan, B. R., & Clark, L. (2019). Revolution or Evolution? Speech Interaction and HCI Design Guidelines. *IEEE Pervasive Computing*, *18*(2), 33–45. https://doi.org/10.1109/MPRV.2019.2906991

Nass, C., & Lee, K. M. (2001). Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied*, *7*(3), 171–181. https://doi.org/10.1037/1076-898X.7.3.171

Nguyen, Q. N., Ta, A., & Prybutok, V. (2019). An Integrated Model of Voice-User Interface Continuance Intention: The Gender Effect. *International Journal of Human-Computer Interaction*, *35*(15), 1362–1377. https://doi.org/10.1080/10447318.2018.1525023

O’Brien, K., Liggett, A., Ramirez-Zohfeld, V., Sunkara, P., & Lindquist, L. A. (2020). Voice-Controlled Intelligent Personal Assistants to Support Aging in Place. *Journal of the American Geriatrics Society*, *68*(1), 176–179. https://doi.org/10.1111/jgs.16217

Obinali, C. (2019). The Perception of Gender in Voice Assistants. *Proceedings of the 22nd Southern Association for Information Systems Conference (SAIS)*, 1–6. https://aisel.aisnet.org/sais2019/39

Pagani, M., Racat, M., & Hofacker, C. F. (2019). Adding Voice to the Omnichannel and How that Affects Brand Trust. *Journal of Interactive Marketing*, *48*, 89–105. https://doi.org/10.1016/j.intmar.2019.05.002

Pfeuffer, N., Benlian, A., Gimpel, H., & Hinz, O. (2019). Anthropomorphic Information Systems. *Business & Information Systems Engineering*, *61*(4), 523–533. https://doi.org/10.1007/s12599-019-00599-y

Qiu, L., & Benbasat, I. (2005). Online consumer trust and live help interfaces: The effects of text-to-speech voice and three-dimensional avatars. *International Journal of Human-Computer Interaction*, *19*(1), 75–94. https://doi.org/10.1207/s15327590ijhc1901_6

Redeker, G. (1984). On Differences Between Spoken and Written Language. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers; television; and new media like real people and places*. Center for the Study of Language and Information; Cambridge University Press.

Rosenthal, G. G., & Ryan, M. J. (2000). Visual and Acosutic Communication in Non-Human Animals: A Comparison. *Indian Academy of Sciences*, *25*(3), 285–290.

Rubin, D. L., Hafer, T., & Arata, K. (2000). Reading and listening to oral-based versus literate-based discourse. *Communication Education*, *49*(2), 121–133. https://doi.org/10.1080/03634520009379200

Schlögl, S., Chollet, G., Garschall, M., Tscheligi, M., & Legouverneur, G. (2013). Exploring voice user interfaces for seniors. *ACM International Conference Proceeding Series*, 2–3. https://doi.org/10.1145/2504335.2504391

Schuetzler, R. M., Grimes, G. M., & Rosser, H. K. (2021). *Deciding Whether and How to Deploy Chatbots Deciding Whether and How to Deploy Chatbots*. *20*(1). https://doi.org/10.17705/2msqe.00039

Shi, Y., Lou, Y., Yan, X., Cao, N., & Ma, X. (2018). Designing emotional expressions of conversational states for voice assistants: Modality and engagement. *Conference on Human Factors in Computing Systems - Proceedings*, *2018-April*, 1–6. https://doi.org/10.1145/3170427.3188560

Smith, S. (2020, April). Juniper Research: Number of Voice Assistant Devices in Use to Overtake World Population by 2024, Reaching 8.4bn, Led by Smartphones. *Businesswire*. https://www.businesswire.com/news/home/20200427005609/en/Juniper-Research-Number-Voice-Assistant-Devices-Overtake

Straßmann, C., Krämer, N. C., Buschmeier, H., & Kopp, S. (2020). Age-Related Differences in the Evaluation of a Virtual Health Agent's Appearance and Embodiment in a Health-Related Interaction: Experimental Lab Study. *Journal of Medical Internet Research*, *22*(4), 1–15. https://doi.org/10.2196/13726

Sutton, S. J. (2020). Gender Ambiguous, not Genderless: Designing Gender in Voice User Interfaces (VUIs) with Sensitivity. *ACM International Conference Proceeding Series*, *July 2017*. https://doi.org/10.1145/3405755.3406123

Sutton, S. J., Foulkes, P., Kirk, D., & Lawson, S. (2019). Voice as a design material: Sociophonetic inspired design strategies in human-computer interaction. *Conference on Human Factors in Computing Systems - Proceedings*, 1–14. https://doi.org/10.1145/3290605.3300833

Tabassum, M., Kosinski, T., Frik, A., Malkin, N., Wijesekera, P., Egelman, S., & Lipford, H. R. (2019). Investigating users' preferences and expectations for always-listening voice assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *3*(4). https://doi.org/10.1145/3369807

Tamagawa, R., Watson, C. I., Kuo, I. H., Macdonald, B. A., & Broadbent, E. (2011). The effects of synthesized voice accents on user perceptions of robots. *International Journal of Social Robotics*, *3*(3), 253–262. https://doi.org/10.1007/s12369-011-0100-4

Tarantola, A. (2020, November). Alexa is getting better at guessing your intentions. *Endgadget*. https://www.engadget.com/alexa-is-getting-better-at-guessing-your-intentions-202813759.html

Tay, B., Jung, Y., & Park, T. (2014). When stereotypes meet robots: The double-edge sword of robot gender and personality in human-robot interaction. *Computers in Human Behavior*, *38*, 75–84. https://doi.org/10.1016/j.chb.2014.05.014

Tolmeijer, S.; Zierau, N.; Janson, A.; Wahdatehagh, J. S.; Leimeister; J. M. & Bernstein, A. (2021). Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. *ACM CHI 2021 Late Breaking Work*.

Torre, I., Goslin, J., & White, L. (2020). If your device could smile: People trust happy-sounding artificial agents more. *Computers in Human Behavior*, *105*(December 2019), 106215.

https://doi.org/10.1016/j.chb.2019.106215

Torre, I., Goslin, J., White, L., & Zanatto, D. (2018). Trust in artificial voices: A "congruency effect" of first impressions and behavioural experience. *ACM International Conference Proceeding Series*, *April*. https://doi.org/10.1145/3183654.3183691

Torre, I., Latupeirissa, A. B., & McGinn, C. (2020). *How context shapes the appropriateness of a robot's voice*. 215–222. https://doi.org/10.1109/ro-man47096.2020.9223449

Trovato, G., Lopez, A., Paredes, R., & Cuellar, F. (2017). Security and guidance: Two roles for a humanoid robot in an interaction experiment. *RO-MAN 2017 - 26th IEEE International Symposium on Robot and Human Interactive Communication*, *2017-Janua*, 230–235. https://doi.org/10.1109/ROMAN.2017.8172307

Truschin, S., Schermann, M., Goswami, S., & Krcmar, H. (2014). Designing interfaces for multiple-goal environments: Experimental insights from in-vehicle speech interfaces. *ACM Transactions on Computer-Human Interaction*, *21*(1). https://doi.org/10.1145/2544066

Völkel, S. T., Schödel, R., Buschek, D., Stachl, C., Winterhalter, V., Bühner, M., & Hussmann, H. (2020). Developing a Personality Model for Speech-based Conversational Agents Using the Psycholexical Approach. *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3313831.3376210

Webster, J., & Watson, R. T. (2002). Analyzing the Past to Prepare for the Future: Writing a Literature Review. *MIS Quarterly*, *26*(2), xiii–xxiii. https://doi.org/10.1.1.104.6570

Wellnhammer, N., Dolata, M., Steigler, S., & Schwabe, G. (2020). Studying with the Help of Digital Tutors: Design Aspects of Conversational Agents that Influence the Learning Process. *Proceedings of the 53rd Hawaii International Conference on System Sciences*, *3*, 146–155. https://doi.org/10.24251/hicss.2020.019

Wolfswinkel, J. F., Furtmueller, E., & Wilderom, C. P. M. (2013). Using grounded theory as a method for rigorously reviewing literature. *European Journal of Information Systems*, *22*(1), 45–55. https://doi.org/10.1057/ejis.2011.51

Wong, P. N. Y., Brumby, D. P., Babu, H. V. R., & Kobayashi, K. (2019). "Watch out!" Semi-autonomous vehicles using assertive voices to grab distracted drivers' attention. *Conference on Human Factors in Computing Systems - Proceedings*, 5–10. https://doi.org/10.1145/3290607.3312838

World Health Organization. (2019). World report on vision Executive Summary. *World Health Organization*, *214*(14), 1–12. https://www.who.int/news-room/detail/08-10-2019-who-launches-first-world-report-on-vision

Zepf, S., Gupta, A., Krämer, J. P., & Minker, W. (2020). EmpathicSDS: Investigating Lexical and Acoustic Mimicry to Improve Perceived Empathy in Speech Dialogue Systems. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3405755.3406125