KinVoices: Using Voices of Friends and Family in Voice Interfaces

SAM W. T. CHAN, Augmented Human Lab, University of Auckland, New Zealand
TAMIL SELVAN GUNASEKARAN, Empathic Computing Lab, University of Auckland, New Zealand
YUN SUEN PAI, Empathic Computing Lab, University of Auckland, New Zealand
HAIMO ZHANG, Augmented Human Lab, University of Auckland, New Zealand
SURANGA NANAYAKKARA, Augmented Human Lab, University of Auckland, New Zealand

With voice user interfaces (VUIs) becoming ubiquitous and speech synthesis technology maturing, it is possible to synthesise voices to resemble our friends and relatives (which we will collectively call 'kin') and use them on VUIs. However, designing such interfaces and investigating *how* the familiarity of kin voices affect user perceptions remain under-explored. Our surveys and interviews with 25 users revealed that VUIs using kin voices were perceived as more engaging, persuasive and safer yet eerier than VUIs using common virtual assistant voices. We then developed a technology probe, KinVoice, an Alexa-based VUI that was deployed in three households over two weeks. Users set reminders using KinVoice, which in turn, gave the reminders in synthesised kin voices. This was to explore users' needs, uncover challenges involved and inspire new applications. We discuss design guidelines for integrating familiar kin voices into VUIs, applications that benefit from its usage, and implications for balancing voice realism and usability with security and diversification.

CCS Concepts: • **Human-centered computing → User studies**; **Natural language interfaces**.

Additional Key Words and Phrases: Voice user interface; voice interface; speech interface; voice design; voice synthesis; conversational agent; virtual assistant; intelligent personal assistant; smart speaker; Amazon Echo; Amazon Alexa; Google Assistant; voice cloning; voice reminder; voice notification

ACM Reference Format:
Sam W. T. Chan, Tamil Selvan Gunasekaran, Yun Suen Pai, Haimo Zhang, and Suranga Nanayakkara. 2021. KinVoices: Using Voices of Friends and Family in Voice Interfaces. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 446 (October 2021), 25 pages. https://doi.org/10.1145/3479590

1 INTRODUCTION

Humans are relational and social beings. We tend to converse more with, easily relate to and trust the people we are close to, namely, our friends and family [59]. Due to this, we see the emergence of technologies that utilise *voice recordings* of caregivers and loved ones in reminder systems [2, 24, 29] and smart home systems [57]. These technologies attempt to leverage the closeness and familiarity of the voices to improve the users' responses. Beyond using *voice recordings*, current commercially available voice user interfaces (VUIs), including virtual assistants (e.g., Google Assistant [44], Siri [36] and Alexa [37]), use computer-synthesised speech made from voice samples of professional

Authors' addresses: Sam W. T. Chan, Augmented Human Lab, University of Auckland, New Zealand, sam@ahlab.org; Tamil Selvan Gunasekaran, Empathic Computing Lab, University of Auckland, New Zealand, themastergts007@gmail.com; Yun Suen Pai, Empathic Computing Lab, University of Auckland, New Zealand, ; Haimo Zhang, Augmented Human Lab, University of Auckland, New Zealand, haimo@ahlab.org; Suranga Nanayakkara, Augmented Human Lab, University of Auckland, New Zealand, suranga@ahlab.org.

2573-0142/2021/10-ART446 $15.00
https://doi.org/10.1145/3479590

voice actors. These interfaces provide users with multiple (albeit limited) voice and language options. With the advancements in speech synthesis technology and increased access to them [69], voice synthesis is becoming more personalised and easier to use. Users now have greater opportunity to generate voices just from samples of their own voices or audio they have access to [39, 64, 65, 70]. It has become possible to build VUIs that use synthesised voices that resemble our friends' and relatives' voices, which we will collectively refer to as 'kin' voices.

The way VUIs sound has been known to have visceral and unconscious effects on humans, user experience and society [67]. According to the 'Computers are Social Actors' (CASA) theory [52], human-computer interactions (HCI) are *also* fundamentally social and voices are considered social actors. These lead us to apply human social norms and personas to computers. With a growing body of work and interest in the HCI community in designing and selecting voices [11–13, 16, 67], we were inspired to explore two research questions:

RQ1: What kinds of personas and attitudes do we apply to VUIs using kin voices?
RQ2: What are the key considerations when designing VUIs that use kin voices?

By harnessing kin voices as a design material for VUIs [67] and investigating these under-explored areas, this could be a step towards having kin voices as persuasive design [12], creating VUIs which are more relatable and better-received than current ones. However, an overly realistic and familiar-sounding voice could bring misuse, harm and distrust [47, 53]. We also aim to explore how they could be designed to tackle these issues.

To answer our research questions, we conducted two user studies. The first study involved surveys and interviews with 25 participants to examine user attitudes, perceptions and experience when hearing the voices from their kin in the context of voice interfaces and virtual assistants. We examined aspects such as likeability, eeriness, perceived safety, persuasiveness, credibility, social presence and co-presence, and we put these in comparison with existing commonly-used virtual assistant voices. The interviews helped us to further understand their experience, preferences and thoughts on the potential scenarios and context of use and misuse of kin voices in VUIs. Informed by our findings, we designed our second study involving a technology probe, KinVoice, an Alexa-based VUI that synthesised reminders in kin voices. KinVoice was deployed within three households for two weeks followed by interviews and co-design activities. This was to gain insights into the social science factors of user needs and motivations in real-life settings, the engineering factors regarding technical requirements and challenges, and to elicit new applications to inform future designs. We propose design guidelines for building and using kin voices for VUIs, and discuss the need to balance voice realism and usability with goals of security and diversification. To summarise, our work contributes with the following:

- An investigation of the impact of kin voices on user attitudes and experience through our surveys and interviews. We show how interfaces with kin voices were perceived as having higher social presence and co-presence, more persuasive and safer (yet eerier) than interfaces using current virtual assistant voices.
- An analysis based on the findings from the interviews and technology probe. We form guidelines for integrating kin voices with VUIs and highlight applications that may best benefit from its usage.

2 RELATED WORK

In this section, we discuss related literature on voice realism, familiarity and design.

2.1 Effects of Voice Realism and Familiarity on User Perception

Prior studies in HCI and human-robot interaction (HRI) have examined the impact of voice realism on user perceptions by comparing human voices with machine-generated voices. Human speech was rated more favourably and perceived as more persuasive than computer-generated speech [66]. Authors found that if the voices came from computers, both human and computer voices were rated similarly. This meant that users were equally comfortable with either voice. Another study, involving physical robots, compared users' impressions of persuasiveness and charisma of two synthesised speeches which replicated the speech characteristics of Steve Jobs and Mark Zuckerberg [28]. The study showed that robots could benefit from increased speech melodies and charismatic speech. Other aspects that researchers have looked into were the effects of human voices on social presence and trust of recommendation agents [17] and virtual assistants [18] compared to using synthetic voices. Human voices induced higher social presence and trust for both studies. By contrast, Abdulrahman et al. [1] revealed that there was no difference in co-presence perception, trust or working alliance between the human voice and machine-generated voice for virtual humans. Their research also showed that machine voices were perceived as more eerie-sounding and less likeable than human voices. Commonly-used speech systems were rated more comprehensible and natural because users were familiar with them [32]. It is important to note that the effect of embodiment as seen in virtual humans and robots might have affected user perceptions in many of these studies [3] and is different from our case of investigating with VUIs without such human-like embodiment. Thus, our work is similar to works involving intelligent personal assistants (IPAs). In Doyle et al.'s work [22], repertory grids were used to identify the dimensions that influence user perceptions of humanness in IPAs. Within each dimension, construct pairs were formed (e.g., real or fake). Our work examined several of these construct pairs, especially within the dimensions of vocal qualities, interpersonal connection, partner identity and role. We also focused on quantifying the bipolar pairs as measures for user perception. These studies gave us a preview of the potential effects that kin voices in VUI could have on user perception and the evaluation methods and metrics that could be used. Unlike these prior works that compare human and generated voices, we want to study the effects on perception with the added factors of kinship and close familiarity.

The concepts of familiarity, recognition and trust could help us understand the effects that kin voices have on perception and experience. Many entities (voices, images, etc.) are perceived as familiar due to frequent exposure, and this 'mere exposure' effect has been found to alter personal preferences and improve affect-based trust [43]. In interpersonal conversations, having trust allows for more personal conversations and is essential for long-term and deep relationships in humans [19]. Trust might serve as a bridge to more interactions with IPAs [19]. A design study with elderly care home residents found that they would prefer virtual assistants which prompted them in their caregivers' voices [42]. Another study on designing smart home systems revealed that, contrary to expectations of confusion, people living with dementia seemed to respond better to voices they recognise and trust, such as someone close to them [57]. The CHIPP prompting system used caregivers' recorded voice reminders as they were thought to help with recognition [38]. These works focused on specific groups of people and looked at caregivers as the primary familiar voices. Our work encompasses friends and relatives who might be more familiar to users. A study to develop a wearable audio-based system for mediating wisdom from mentors [60] found that users wished to receive wisdom from those they consider as good mentors and those who they can relate to; particularly, personal heroes, friends and family members. The authors have yet to explore these wishes. Epp et al. [25] explored facilitating the recording of audio e-books in the voices of a family member for improved accessibility of the e-books. However, their study was focused on evaluating the system's usability and not user perceptions on hearing the e-books in the familiar voices. Lastly,

a study comparing user's own voice, familiar voices (of instructors and lab mates) and unfamiliar voices to design voice notifications showed that familiar voices attracted more attention (were more recognisable) than unfamiliar ones [7]. Users reacted faster to notifications with familiar voices than unfamiliar ones. Our work builds on the idea that familiarity brings attention and recognition. We further explore it with kin. Altogether, these works provide a basis for understanding user preferences and perceptions of various kinds of familiar voices. However, in this paper, we are more interested in the voices of people with strong kinship or friendship, which is distinct from a familiar voice such as that of a known actor's. Kin voices are possibly more familiar and could be viewed as having higher empathy and trustworthiness, yet their effects remain an under-explored research gap that we aim to address. Furthermore, many of these works either propose using familiar voices or have used voice recordings of familiar people. Our work not only investigates recordings of kin voices but also the use of synthesised voices that sound familiar and similar to our kin.

2.2 Voice Design

Previous literature have proposed theories to explain these effects of voices in order to form frameworks for designing voices in HCI. Many frameworks stem from the CASA theory [52]. A core tenet of the CASA theory was that the notions of 'self' and 'other' are applied to voices, or in other words, as Sutton et al. [67] have put it: "different voices indicate different personas". This means that people appear to apply the same social norms and expectations to computers that we apply to humans. The theory was derived from The Media Equation [63] which claimed that "media equals real life" and that humans are not evolved enough to handle digital media. The CASA and Media Equation theories have been argued to have theoretical limitations [21, 30], and were developed when human-computer interactions were rare and less complex [30]. With the interactions becoming increasingly pervasive and advanced, researchers suggested extending CASA to include that humans may develop and apply human-media social scripts (mental models for interactions) to computers. The voices/personas we have in VUIs today might strengthen stereotypes and ideologies that affect our social and cultural futures. Thus, Sutton et al. [67] proposed to integrate sociophonetics and HCI into voice design through three strategies: 1) individualisation through enabling a wider range and control for the selection of voices, 2) context awareness through understanding in what context the voices will be used in, and 3) diversification through the intentional design of voices which challenge stereotypes and speech ideologies. Cambre et al. [12] agreed with Sutton et al. on the strategies of individualisation and context awareness, and added that the medium (device) in which the voice is embodied also plays a part in the design as devices have different appearances and functionality. They further postulate that voices can be designed to intentionally misalign the voice with expectations and stereotypes as a form of persuasive design to change user attitudes and behaviours. In this paper, we seek to analyse the effects of kin voice VUI, and design them through these lenses and strategies. When looking into design guidelines for KinVoice, we refer to three other important design considerations in which we will discuss:

Similarity-Attraction Effect: The similarity-attraction effect [51] states that people will prefer to interact with others who are similar to them in personality. This theory has been expanded upon over the years through HCI and HRI studies. Research on developing systems that mirror users' conversation style [10] and grammar construction [20] support this theory. We see evidence of this for sex and gender perceptions towards voices from robots as well [27]. A study on robots and voice accents [68] show that users prefer their native accent. However, this might not be true for all accents as a study in Singapore uncovered that a British English accent was preferred over the regional Singaporean accent [54]. Thus, it is not a straightforward approach to explain the effects of kin voices in VUIs through the similarity-attraction effect. Our relatives might have social and

cultural influences, and possibly even personalities that are close to us. However, our gender and age would likely be different. Conversely, we are likely to have a diversity of friends who may not have similar social and cultural backgrounds, but we might share interests, activities, geographical location, gender and age. A broader range of factors would need to be considered.

Human-Like or Robot-Like: A major goal in VUI design has been to improve the machine-synthesised voices' human-likeness. Techniques, such as voice puppetry, have been used to design synthesised speech that sound more natural and human-like through fine control by human voices [4]. Users will expect human-like qualities from human-sounding interfaces [13]. However, users might be disappointed if these expectations are not met. Nass et al. [52] argued that it does not take much modification for computers to feel like social actors (CASA theory) and exhibit personality [51]; nor do VUIs need to emulate the characteristics of human speech [49]. However, it remains undefined and subjective how much modification (social cues) computers require [30]. Cabral et al. noted that although human voices are more expressive, understandable and likeable than synthetic voices, reducing human-likeness may not definitively cause different perceptions in qualities such as the virtual characters' appeal and credibility [10]. Cambre et al. [12] further suggested in their framework to consider making voices in VUI that are "distinctly and deliberately non-human". Aylett et al. [3] suggested that the designed unnaturalness of the voice or how robotic the voice is should not undermine but should support the robot/interface's personality. These are key considerations to consider when designing kin voice VUIs.

Linguistic Content: Although it was shown that social cues in voices of VUI would override the social cues of linguistic content [50], content is still a factor in how users experience and perceive the voices [12]. Users would expect consistency between voice characteristics and content and how well aligned it is with the voice persona. As such, users of kin voices VUIs might consider it strange and might not appreciate it if the synthesised voices said content that their real-life counterparts would not.

3 STUDY 1: IMPACT OF KIN VOICES ON USER ATTITUDES AND EXPERIENCE

This study aims to address RQ1, have an overview of how users might perceive and experience kin voices in the context of VUIs and to compare this with an existing commonly-used VUI voice.

3.1 Preliminary Survey

The applications specifically afforded by VUIs using kin voices were initially unknown. Our first assumptions were that the usage scenarios would be the same as how we use virtual assistants, such as playing music and information search, due to their prevalence. Hence, as a starting point, we issued a preliminary online survey which had an open question for respondents to imagine being given a chance to use a virtual assistant which used a kin's voice ("If the virtual assistant has the voice of a person close to you (friend/family), what would you use it for?"). The survey was issued through the researchers' social media platforms, personal friends and relatives via word-of-mouth, email and chat networks. We received 48 responses (mean age = 39.9, SD = 17.4, age range: 19 to 77, 24 male and 24 female).

From the responses, we extracted four scenario themes which we identified as unique to VUI design space and particularly suited for kin voice use: 1) giving a reminder that is directly related to the person whose voice is playing (e.g., birthday or social meeting) (13 responses), 2) reading out text or chat messages from the friend or relative in their voice (5 responses), 3) saying out motivational phrases (3 responses) and 4) reading out stories (2 responses). We realised from the responses that our assumptions had to change. VUIs have a wide range of usage scenarios and a few of them (e.g., reading aloud on-screen text, long-form texts like reading audiobooks) are not

limited to the applications we would normally use with virtual assistants. As such, virtual assistants could be viewed as a subset of VUIs. Therefore, this preliminary survey helped us to expand our understanding of VUI design, and these extracted scenarios were kept in mind as we designed the main study.

3.2 Study Design

The main study used a within-subjects approach in which the same participant had conversations with a simulated voice interface and heard phrases from the two voices representing two conditions: 1) kin voice (pre-recorded from participants' kin) and 2) generic voice (synthesised voice used in VUIs). For each condition, participants were asked to rate their attitudes towards and impressions of the voice through a survey. After the surveys, we held interviews to discuss their experience hearing the two voices, their thoughts on the usage scenarios presented to them and future scenarios, and lastly, their thoughts on trust and how usage of the technology might affect their relationship.

3.3 Participants

Participants were recruited via invitation through university and community groups email networks, the university website and word-of-mouth. Participants were included in the study if they were aged 18 or above, were fluent or native English speakers, had access to a computer with a camera and microphone, and had access to the internet. Participants who did not meet these criteria were excluded from the study. Participants were invited to ask a friend or relative to take part in the study with them as their 'study partner'. Ethical approval from the university's ethics review board was obtained and signed consent was given by participants prior to the start of the study. A total of 25 participants took part (mean age = 42, SD = 17.4, age range = 19 to 74, 11 male and 14 female). The study partner pairings ranged from siblings (1 pair, known each other for 67 years), to child-parent (2 pairs, average relationship time: 21.5 years), grandchild-grandparent (1 pair, for 21 years), co-workers (2 pairs, average relationship time: 2.5 years), friends (4 pairs, average relationship time: 8.3 years) and spouses/couples (5 pairs, average relationship time: 21.1 years). Three participants were non-native but fluent English speakers and were accustomed to speaking to their kin in English. All of the 25 participants, except four of them, had used a VUI before. Six participants were frequent VUI users (at least once a week), five participants used VUIs about once a month and ten participants reported not using VUIs often or rarely using them. Participants indicated that they had used Google Assistant (12 mentions), Siri (10 mentions), Bixby (3 mentions) and Alexa (1 mention).

3.4 Survey Measures

Three groups of 14 measures were formed for the survey (see survey items in Supplementary Materials) which were inspired by studies comparing human versus computer-generated voices as mentioned in Related Work (Section 2.1).

Social Presence, Telepresence and Co-presence (4 measures): We utilise the measures that were previously used to assess virtual humans [1] and avatars in virtual environments [55, 56]. 'Social Presence' refers to how users perceive the interface's ability to provide salience of another person or to allow that person to be *noticed*. The four items were measured on a sliding scale (0 to 100). 'Telepresence' refers to the feeling of *immersion* or that another person is present with the user through the interface. The five items were assessed on a 7-point scale. Co-presence has been referred to as the psychological *connection* to and with another person and is usually measured in two dimensions: 'Perceived Co-Presence' (11 items) which refers to the feeling that the 'interaction partner' is able to perceive them and 'Self-Reported Co-Presence' (6 items) which refers to the

feeling of being able to perceive their 'interaction partner'. Items for these scales were measured on a 5-point metric.

Persuasive and Charismatic Speech (4 measures): The measures of persuasive speech (speaker persuasiveness, credibility, and strength) were constructed based on a study on human versus computer voices in robots [66]. Each item was measured on a 5-point semantic differential scale in which the options were adjectives of opposite meaning. The 'Persuasiveness' of speech was assessed via four items (unconvincing-convincing, negative-positive, harmful-beneficial, and ineffective-effective). The perceptions of the speaker was assessed in two dimensions of speaker 'Credibility' (honest-dishonest, uninformed-informed, untrustworthy-trustworthy, unqualified-qualified, and insincere-sincere) and speaker 'Strength' (unassertive-assertive, timid-bold, inactive-active, and meek-forceful). 'Persuasiveness' items and 'credibility and strength' items were adapted from a 9-point scale and 7-point scale, respectively, because we wanted them on the same and simpler scale, making them easier to administer with the other measures in the next subsection. Measures retained acceptable reliability: Persuasiveness (Cronbach's α = .89), Credibility (Cronbach's α = .83) and Strength (Cronbach's α = .70). We assessed how 'Charismatic' the speech was using the items from the study by Fischer et al. [28] that replicated and compared the synthesised voices of Steve Jobs and Mark Zuckerberg from a robot. We followed their measure in which users rated on a 7-point scale how much (1: not at all, 7: very much) the speech corresponded to five-word items to describe charismatic speech (enthusiastic, charming, persuasive, passionate, engaging).

Perceptions of Voices (6 measures): The Godspeed indices [5] of 'Likeability' (five items), 'Humanness' (four items), perceived 'Intelligence' (five items), and perceived 'Safety' (three items) were included. The item, 'Moving rigidly-Moving elegantly', on the 'Humanness' index was removed as it applied more to animated entities and was irrelevant to VUIs. Reliability for 'Humanness' measure remained excellent (Cronbach's α = .93). We also evaluated voices through the perceived 'Eeriness' index (eight items), adapted from Abdulrahman et al. [1], that was meant to be an alternative to the Godspeed indices to include the factor of affect (emotion) [34]. All of these items were measured on a 5-point scale. Lastly, the study compared the 'Understandability' of the voices. We adapted two items from [1] to refer to the VUI voice instead of the authors' virtual human voice. The items were statements ("The voice was difficult to understand" and "the voice made it easy to listen to the phrases") and were to be rated on a 5-point scale how much users agreed on the statements (1: Strongly Disagree, 5: Strong Agree).

3.5 Voices, Conversations and Phrases Presented

Since we were interested in including the measures of presence and co-presence, it was appropriate to simulate the voice interactions using the Wizard-of-Oz method through conversations as those measures asked about the user's perception of and with an 'interaction partner'. In our case, the 'interaction partner' was a VUI and its conversations with the user were controlled by the Wizard-of-Oz (researcher). We decided to present the conversations and phrases separately. We first presented participants with the conversations (we expected their verbal responses and answers) and let them answer the survey items regarding presence, telepresence and co-presence (first group of measures in Section 3.4). Next, we presented the phrases (participants need not verbally respond) and the rest of the survey items to not confuse participants with regards to what was meant by 'interaction partner'.

We developed our set of conversations and phrases, and prepared them via Google Slides[1] (see sample slides in Supplementary Materials). We used the recorded speech of the participant's study

[1] https://www.google.com/slides/about/

partner saying out the required conversation content and phrases as the kin voices. Pre-recordings were used instead of synthesised kin voices because large voice samples (more than 1-hour worth) and good quality (clear and noise-free) samples would be needed from each participant to produce synthesised voices of a quality and naturalness comparable to the current generic VUI voices. From our pilot tests, the existing quality of synthesised kin voices that we could achieve with small samples would have confounded Study 1's results as the synthesised kin voices were much less natural-sounding than synthesised generic voices. Thus, this was the simplest way to answer RQ1 without introducing such confounds. To minimise confounding effects of gender perceptions, as highlighted in Related Work (Section 2.2), we prepared two conversation and phrase sets with male and female synthesised voices to match the gender of their kin. For example, if the participant's kin was female (kin voice condition), we would present to the participant a female-sounding synthesised speech for the generic voice condition. For the generic voice, we chose the voices available for Google Dialogflow[2] (for male: Automatic, for female: en-US-Wavenet-C) as they represented the current common and quality voices used by VUIs. Each participant would be presented four slide sets with one set for conversations and another set for phrases for the two conditions.

Conversations: We developed three conversation scenarios with a total of six lines of conversation content that were related to setting up reminders, responding to a motivational prompt to exercise, and asking the VUI to play music. Setting up reminders and the motivation prompt were from the usage scenarios extracted from our preliminary survey. Playing music was one of the most common uses for VUIs [6, 23]. When presented to the participants, they saw a quoted text in blue indicating the responses that they could say to the 'interaction partner' but were not restricted to say them in verbatim. The slide included an animated microphone icon to indicate to the user that the 'interaction partner' was listening and that the user could go ahead and speak. When they did, the Wizard-of-Oz would click to the next slide to play the attached audio clip giving the illusion of interactivity.

Phrases: We developed 11 phrases that incorporated the common applications for VUIs (alarms, weather, search and recommendations, smart home control, navigation, and normal meeting reminders) [6, 23, 71] and the four scenarios extracted from the preliminary survey in Section 3.1 (personalised reminders, personal messages, motivational phrases, and stories/poems). The last scenario was hearing quotes/wisdom from famous figures [60].

3.6 Procedure

Interested participants were first sent the participant information sheet and consent form to read and sign. Researchers also confirmed if participants had a study partner. Upon receiving the consent forms, researchers would request each person in the study partner pair record the conversation content (six lines) and phrases (11 lines) in a total of 17 separate audio recordings. Participants were not told the purpose of the recordings, they were asked not to discuss the recording to their partner and to record privately. Once the recordings were received, we prepared the slide sets and arranged to meet each participant individually (remotely or physically) for a main experimental session.

Participants were briefed that the session would be about one hour in total as shown in Figure 1. In the first 40 minutes, they were instructed that they would converse with the 'interaction partners' and hear phrases from the two voices in separate 20-minute blocks (Voice A and Voice B). They were not told that they would hear the kin or generic voices. For each block, the conversation set was presented first and participants were asked to fill in the first page of an online questionnaire

[2] https://dialogflow.cloud.google.com/

that asked questions about co-presence, presence and telepresence. The phrases set was presented next and participants were asked to continue to the second page of the questionnaire with the rest of the questions. The order of the blocks was counterbalanced with either the kin voice condition or generic voice condition first. For the remaining 20 minutes, they were interviewed by the researchers to describe their experience and discuss usage scenarios, trust and effect on relationship (see Interview Questions in Supplementary Materials).

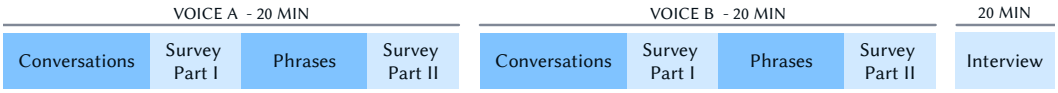

Fig. 1. Procedure for Study 1 Experimental Session. The survey, conversations and phrases for each condition (kin voice and generic voice) were presented in two counterbalanced 20-minute blocks (Voice A and B).

3.7 Apparatus

The majority of the sessions were conducted remotely via the Zoom video-conferencing software.[3] In the remote set-up, the simulated interface was presented via the screen sharing function on Zoom and the interview was recorded via Zoom's built-in recording function. Four participants wished to have it conducted physically where we set up a laptop with an external screen for researchers to present the sets. An external keyboard was used for researchers to silently change the slides.

3.8 Survey Results

The survey results are presented according to the three groups of measures as stated in Section 3.4. The scores for each measure was calculated by taking the average of the ratings for the items in the measure. The normality assumption was met according to the Shapiro-Wilk normality test ($p>.05$) for the paired differences in all of the measures except for perceived 'Safety'. Thus, the Wilcoxon signed rank test was conducted to compare the differences in scores for perceived 'Safety' and paired t-tests was conducted for the rest of the measures. Interval plots are shown in Figure 2. Statistical values for the tests for each measure are summarised in Table 1.

3.8.1 Social Presence, Telepresence and Co-presence: Participants rated experiencing a significantly higher 'Perceived Co-presence' when conversing with VUIs using the kin voices compared to the ones using generic voices ($p<.001$). It was not the same for 'Self-Reported Co-presence' where there was no significant difference in ratings ($p=.1$). This meant that they felt that the VUIs with kin voices were able to perceive them more than VUIs with generic voices but felt equally neutral (scores were near 3) about being able to perceive the VUIs, regardless of voice. They also rated the VUI with kin voices as having significantly more 'Social Presence' ($p<.001$) and higher 'Telepresence' ($p<.001$) than the VUI with the generic voices.

3.8.2 Persuasive and Charismatic Speech: Kin voices in VUIs were shown to be perceived as significantly more persuasive ($p<.001$), credible ($p<.001$) and charismatic ($p<.001$) than generic voices. There were no significant differences in their perception of speaker 'Strength' for kin voices and generic voices ($p=.1$).

3.8.3 Perceptions of Voice: Kin voices are rated as significantly more likeable ($p<.001$), more intelligent ($p<.001$) and safer ($p<.001$) than generic voices. As expected, participants rated kin voices as significantly more human-like than generic ones ($p<.001$). By contrast, we found that

[3] https://zoom.us/

kin voices were perceived as significantly more eerie-sounding (p<.05) than generic ones. Kin and generic voices were equally intelligible (understandable) (p=.1).

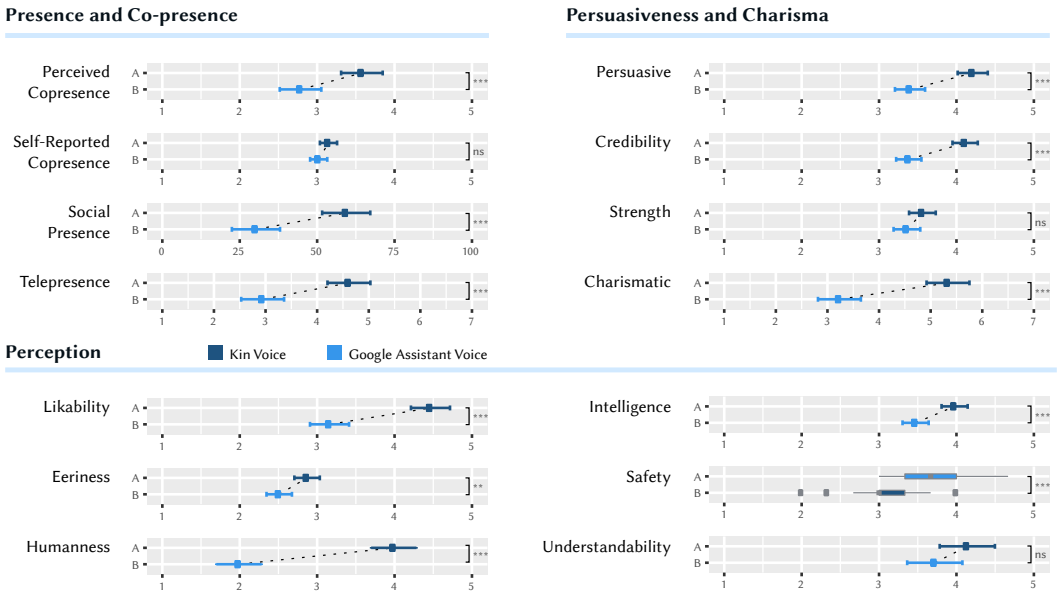

Fig. 2. Interval plots for survey measures showing the mean scores (square dots) and confidence intervals (error bars). Top and dark blue item (A) for each plot represents values for kin voices while the bottom and light blue item (B) represents the values for generic voices. Dotted lines between the means of item pairs represent the difference in means with longer lines representing higher differences. Significance levels are presented on the right of each plot, p<.001:***, p<.05:*, no significant differences: ns.

Summary of Values				
Measures	'Kin' Voice ($M \pm S.D.$)	Generic Voice ($M \pm S.D.$)	t_{24} Value/ Z-Value	p-Value
Perceived Co-presence	3.58 ± .65	2.79 ± .65	4.31	< .001
Self-Reported Co-presence	3.15 ± .27	3.02 ± .27	1.67	= .1
Social Presence	59.39 ± 18.8	30.23 ± 18.8	5.48	< .001
Telepresence	4.62 ± 1.0	2.94 ± 1.0	5.87	< .001
Persuasive	4.21 ± .47	3.40 ± .47	6.08	< .001
Credibility	4.11 ± .40	3.38 ± .40	6.51	< .001
Strength	3.56 ± .42	3.36 ± .42	1.69	= .1
Charismatic	5.34 ± 1.0	3.23 ± 1.0	7.39	< .001
Likability	4.46 ± .61	3.16 ± .61	7.54	< .001
Eeriness	2.87 ± .40	2.51 ± .40	3.20	< .05
Humanness	3.99 ± .71	1.99 ± .71	9.97	< .001
Intelligence	3.98 ± .41	3.47 ± .41	4.36	< .001
Understandability	4.14 ± .86	3.72 ± .86	1.72	= .1
Safety	3.67 ± .51	3.0 ± .46	216	< .001

Table 1. Summary of means/median (M), standard deviations (S.D.), t-values/Z-value and p-values for the survey measures.

3.9 Interview Results

The interviews were transcribed. The transcripts were coded independently by two researchers and analysed using the thematic analysis method [8] to generate initial themes. The researchers then reviewed the coded data and themes to come up with our final themes and analysis as presented below.

3.9.1 Kin voices as being familiar thus fostering connection and sense of presence: As expected, the presented 'kin' voices were described as being familiar. This sense of familiarity may have created the feeling of closeness and personal connection with the VUIs with 'kin' voices. P4 mentioned how she *"felt connected"* with the kin voice. The idea was also evident from descriptions of generic-voiced VUIs, *"I felt there was sense of distance in the voice, it was like no connection and [it was] strange to hear"* (P11). Additionally, the familiarity of the kin voices may have fostered the feeling that their study partner was there with them. This was felt to varying extents: from feeling that the study partner was talking through the VUI, *"I felt like she was talking over the phone"* (P2), to feeling that the study partner was present in the same space, *"Sounded like she was in the room questioning me and reminding me"* (P24).

3.9.2 Kin voices as being engaging: The increased sense of familiarity, connection and presence may have made the VUIs with kin voices more engaging to listen to. We see this idea through descriptions of kin-voiced VUIs that they were interesting to hear (P2, P14, P16), made them listen more (P19, P20, P24) and could be summed up with P24's statement: *"Maybe you're more attentive to it because it is a familiar voice that you would listen to it more."*

The difference in voice realism and tone of the kin voices and generic voices may be another reason for why kin voices were more engaging than generic voices. The generic voices sounded better (more natural) than many machine-generated ones they have heard before. However, they were indeed described as not as natural nor expressive as kin voices. An interviewee used the words *"very bland"* (P16) to describe the generic voice. This could mean that the voice was not expressive or it could mean that the voice was uninteresting to hear. Perhaps it could mean both. This idea could be further unpacked from the two quotes: *"[It was] very machine-like. [It was] hard to listen to. [It was] very cold. Didn't want to listen to it."* (P12), and *"[The generic voice was] not quite the same. Wasn't as bad for a machine-generated voice. They've got a coldness to them. It's quite tone-y and tangy. Better than average. But still obviously a machine. It's the tone and intonation."* (P23). The use of the word "cold" relates to the feeling of disconnection. Thus, the lack of realism and expressivity in the voice tone may have led to feeling no personal connection, making listeners feel less engaged and interested in the generic voices than kin voices.

3.9.3 Kin voices as being comforting: The descriptions of kin voices being "comforting" (P3) and "calming" (P16) may have stemmed from the participants' familiarity with the voices as well as how they perceived their study partners. We see other descriptors, such as "heart-warming" and "reassuring", as synonyms that reinforce the idea that the kin voices were "comforting". It is possible that the sense of familiarity made hearing from the kin-voiced VUI more enjoyable, as seen from P24: *"Definitely when it's a familiar voice, it's definitely nicer to hear."* This, in turn, may have led to them associating kin voices to the feeling of comfort and calmness.

3.9.4 Kin voices as being similar but still different to original voices: Participants felt that the kin voices were very similar to the voices of their actual study partner's voice. This made sense as their real voice recordings were used. However, a few noted that the VUIs with kin voices were saying things that their study partner normally would not (P6, P18). Thus, the VUIs felt different and as if something was off.

3.9.5 Kin voices encourage interaction and trust but in limited forms: If kin voices are engaging to hear, they may also encourage acceptance, leading to increased interaction and disclosure. This idea is captured by P24's response: *"I think I would be more accepting of it (VUIs with kin voices), if it was a resembled voice instead of being just a total machine... You would interact more with it and you would be more responsive to it rather than it being a machine generated voice."* Furthermore, trust in others or interfaces is often related to how much people are willing to disclose their thoughts, feelings and information [59]. Several interviewees mentioned that if the VUIs did use voices of their kin, they would be willing to disclose more to the interfaces: *"Oh definitely, [it] would come with associated trust and relationship. Having a voice that I know would definitely [make me] open up more"* (P23). However, several also felt that they would be willing to engage in small talk and casual conversations; however, as they would still be aware that they are disclosing to a voice interface, they would not share as much with the interface as they would with their study partner. This was captured in P19's quote *"I probably wouldn't share really intimate stuff. But if it just [said], "Hey, how's your day going." I'll probably keep it light-hearted."* Therefore, users might disclose more to VUIs with kin voices but might not disclose personal or more intimate information, showing that there might be improved but limited interaction and trust.

3.9.6 Interaction with VUIs with kin voices might affect relationship: There were mixed opinions regarding whether the use of kin voices in VUIs might affect their relationship with their real-life study partner. Several foresaw that there would not be any change. Several believed that they might get bored or annoyed with the VUI and possibly with their study partners if they heard the kin voices too often. Others felt it might make the relationship stronger, as P1 put it *"It won't affect the relationship. On the contrary, it would make the relationship with [study partner] more intimate."*

3.9.7 Usage scenarios are useful if they are personal: From the conversations and phrases presented to them in the kin voice condition, participants picked out the most useful and intriguing usage scenarios for them. There were: personalised reminders, greetings/alarms, reading messages and navigation in the kin's voice. P19 explained *"...because I felt more connection"* and P16 mentioned that it was because they had a *"more personalised feeling"*. The scenarios of hearing weather information, famous quotes and smart home control were brought up as not being useful to use with kin voices. P18 felt that these tasks were *"neutral"*. P17 associated them with being *"impersonal"*. She emphasised that these scenarios were mainly informational and she imagined these scenarios could be said with any voice. From our comparison between their reasons and motivations for the scenarios they deemed useful or not useful, we infer that participants found the scenarios useful to use with kin voices if they were either directly related to their study partner (reminders and messages) or there were strong associations and matching between the scenarios and what their study partner would normally do for them (wake-up greetings and assisting in in-car navigation). In the cases of reading stories and motivational messages, the mixed comments were because they either imagined that their study partner could have read the stories and motivated them, or they could not imagine they would.

3.9.8 Kin voice VUIs as a chance to connect with familiar and meaningful personas: The topic of being able to use various voices in future scenarios was frequently brought up. Several wished to continue using their study partner's voice or another relative's voice including parents, children and even loved ones who passed away. Several wanted to use other friends' voices including mentors. Several also wanted to use the voices of famous people and celebrities, mainly because of their impressions of the voices. P5 elaborated, *"Someone like Alfred Jackson would be great...[he] has a calming and relaxing voice"*. A few talked about the chance to use voices of fictional cartoon, or anime characters. Two participants thought about using their own voice. Overall, interviewees saw

the chance of using the concept of synthesising a voice in VUIs as a way to connect with people who they are familiar with or personas who are meaningful to them.mentia. P9 elaborated that her father had dementia and she believes that it would calm her down if it is a familiar voice as it is more smoothing.

3.10 Discussion on User Attitudes and Personas Assigned to VUIs using Kin Voices

Social Presence, Telepresence and Co-presence: VUIs with kin voices were rated to have higher perceived co-presence, social presence, and telepresence than those with generic voices. These aligned with interview descriptions that kin voices made them feel connected (high co-presence), like the study partner was there (high telepresence), and the VUIs with kin voices were engaging and noticeable (high social presence). From previous work, there was no difference in the perception of perceived nor self-reported co-presence between human and machine voices [1]. Thus, we postulate that the familiarity of the kin voices improves the perceived co-presence — the feeling that VUIs with kin voices are able to perceive and connect with them.

Persuasiveness, Charisma and Trust: Kin voices were also rated as more persuasive, credible and charismatic than generic voices but not necessarily more bold and active (higher speaker strength). The interview themes of kin voices making participants feel attentive and wanting to listen in more ties together with the idea of persuasiveness and charisma. Since human voices have been shown that it could sound equally credible to generic voices [10], we posit that having the familiarity of kin voices in VUIs could make VUIs sound more credible than those using generic voices.

Participants identified in the interviews that the kin voices were similar but the content of what they said was not what the study partner would normally say. This presented a mismatch in the content and perceived persona of the VUI, and might cause a negative experience as people are more likely to trust and like a speech interface when the voice and content match in personality [50]. It could be why the interviewees were more willing to engage in small talk than to disclose more intimate information. To some extent, this might indicate that there is still a lack of trust towards VUIs with kin voices.

General Attitudes: Our findings that kin voices were more likeable than generic voices were consistent with findings related to human versus machine voices [1, 10]. People perceived VUIs with kin voices as more intelligent and they felt safer with them compared to VUIs with generic voices. This is supported by their interview descriptions that kin voices were comforting and calming. Contrary to previous findings that human voices were more understandable than machine voices [1], it seems that in our study, kin voices and generic voices in VUI were equally understandable. This could be because we used the Google Dialogflow TTS voices which might be more natural-sounding than the Microsoft Hazel voice that they used. Prior work also found that machine voices were perceived as more eerie-sounding than human voices [1]. However, in our study, kin voices were rated as eerier than generic voices, even though participants did not consciously use the word 'eerie' or other synonyms to describe kin voices in their interviews. This perception of eeriness could be due to the mismatch [12] between the medium (VUI) and the voices (kin); people do not expect kin voices to come from VUIs. It is plausible that this effect of eeriness and mismatch helps to explain why kin voices and voice notifications in the study by Bhatia et al. [7] attracted more attention. It might then seem contradictory that people find it 'comforting' yet 'eerie' but we think it could just mean that the perception of other factors (likeability, intelligence and safety) had a higher impact on people's attitudes than the perception of eeriness.

Finally, our participants had varied backgrounds and personalities, were from different countries and were related to their 'study partners' in different ways. This could mean that kinship and

familiarity have higher effects on user attitudes than the other factors such as social, cultural, linguistic, voice quality and similarity-attraction.

4 STUDY 2: TECHNOLOGY PROBE

Study 2 was aimed at addressing RQ2 on the key considerations for designing VUIs using kin voices through results from a field test with one of the usage scenarios which was found from Study 1. We employed a technology probe approach [35] to achieve this aim, with the social science goal of understanding usage behaviours and motivations, the engineering goal of testing in-the-wild and design goal of eliciting responses and new applications.

4.1 Design of Probe

Out of the usage scenarios discussed in Study 1, we selected the use case of reminder services for our probe as it was identified as one of the most useful applications and it made use of the ability of text-to-speech (TTS) engines to generate dynamic speech content. We used the Real-Time Voice Cloning (RTVC) tool developed by Jemine et al. [39, 40], based on the research by Jia et al. [41], as it was open-source and was able to synthesise realistic voices with small amounts of voice samples (a minimum of five seconds).

The probe, KinVoice, was implemented on Amazon Echo Dot devices. We developed a custom 'skill', a third-party feature that can be installed and accessed via Echo devices, which enables users to set reminders, issues reminders to users and plays the reminder audio in the voice of the user's kin (Figure 3). The user first invokes the KinVoice 'skill' by calling its name and asking it to set a reminder. It retrieves information on the reminder message, day, and time from the user. Then, it updates the Alexa Developer Console server which helps to keep track of and issue the reminder. The reminder data is also posted to a custom-made Django[4] framework API that is hosted on a Google Cloud[5] server. The API generates the reminder message as an audio file in a kin voice based on a speech recording sample of the user's kin using the RTVC tool [39, 40], typically within 10 seconds and uploads the file to an Amazon S3 bucket database.[6] Due to a security safeguard for Alexa development, the Echo does not allow the synthesised audio file to be played when the reminder is issued. Thus, when the reminder is issued, the Echo Dot announces there is a reminder for them and asks the user to play the reminder message. Finally, KinVoice plays the reminder audio file from the S3 bucket when the user asks and could be played at any time after the reminder is issued, until the next reminder is issued.

[4] https://www.djangoproject.com/
[5] https://cloud.google.com/
[6] https://aws.amazon.com/s3/

Fig. 3. KinVoice interaction: User sets the reminder and KinVoice confirms (left image). KinVoice issues the reminder (middle image) and plays the reminder when the user asks (right image).

4.2 Participants

We invited participants from Study 1 to take part. Two pairs of participants in three separate households were selected. One of the pairs involved two sisters, one of them (who we will refer to as T1, older sister, age 70) lived alone and the other (who we will refer to as T2, younger sister, age 67) lived with her husband. The other pair involved a couple living together (who we will refer to as T3 and T4, age 61 and 64, respectively). We note that they were community-dwelling older adults (aged 60 and above). We chose the pairs because they were able to commit at least two weeks of their time for the probe, lived in the same city as the researchers' lab so that we could assist in setting up and troubleshooting the probe if required and they represented two different kin pairings (siblings and couple). They were allowed to contact the researchers at any time for inquiries and were allowed to stop at any time.

4.3 Procedure

The probe was deployed for two weeks with each participant receiving an Echo Dot. In the week prior to deployment, the researchers went to the participants' households to setup the Echo devices, to get them habituated to using the Echo and to get an idea of what their daily routines were. We asked them to select amongst themselves whose voice they wanted to use and for that person to narrate an article/book for 40 to 60 seconds as we recorded the audio samples. At the start of the first week of the probe, we met the participants remotely via Zoom for about an hour to introduce the probe and its functions. We gave them a demonstration of how it worked and confirmed that they are able to use it without trouble. An instruction card of the phrases that they could say was emailed to them, which they printed out and placed beside the device. At the start of the second week, researchers met with the participants remotely (T1 and T2 individually, T3 and T4 together) for 15 to 30 minutes to check if they had any issues with the probe and to elicit comments. At the end of the two weeks, the researchers met with each pair together at one of the pair's home for a 1.5-hour co-design activity in which we presented a summary of their usage data. We encouraged them to discuss their motivations for their usage, their thoughts on the synthesised kin voices and the probe. We guided them to map out daily activities which were meaningful and important to them. Lastly, they brainstormed opportunities and potential ways to use VUIs with kin voices to address any frustrations associated with their activities.

4.4 Findings

The check-in interviews were transcribed and coded independently by two researchers. We referred to this coded data and the participants' usage data as we prepared our discussion and co-design activity materials at the end of the probe. For each pair, we report our findings on 1) their needs, motivations and usage behaviour, 2) their opinions on the generated kin voices and the probe, and 3) their responses and the most relevant ideas from their co-design activity.

4.4.1 Needs, Motivations and Usage Behaviour: T1 lives alone with two cats, spends most of her time reading and visits the gym with her sister (T2) every Friday. She works occasionally as a committee member on a statutory board and frequently does volunteer work. The probe was placed on her dining table near the kitchen. She chose to use T2's voice, set nine different reminders on KinVoice and played each reminder audio. One example of a reminder she set was to check emails for replies regarding an event she organised for her volunteer work.

T2 lives with her husband and works eight hours per week as a General Practitioner. She spends the rest of her time going to dance classes, reading and watching television (TV). The probe was placed along their hallway table as it was a high traffic area for them and it was close to the internet router. T2 decided to use her own voice for the probe as she felt that it would be awkward to have

T1's voice played in a common area where her husband could also hear it. She set ten reminders through KinVoice and played six of the reminders. She commented that this was because it was a 'two-step process' to play the reminder and it was more cumbersome to do so. One of the reminders was a daily (recurring) reminder to watch her favourite TV program. T2 seems to take a central role in managing the home and reminding others, setting reminders for her husband to take pills, cook dinner and feed her pet dog.

T3 works full-time at a university and as a life coach on an ad-hoc basis. Her probe was placed on the living room TV table that was near to where she would usually work on her laptop and close to the kitchen. She used T4's voice, set eight reminders and played six of the audio files. Her reminders mainly consisted of doing outdoor tasks such as to *"go pick up the parcel from the warehouse"* and *"to go and buy some lettuce for dinner"*.

T4 works at a bank as a data analyst and has knowledge and interest in Artificial Intelligence. He used T3's voice and placed the probe on the desk in his study room that was one level below their living room. He set eleven reminders and asked to play the audio thirteen separate times. When asked about this, he replayed one or two of the reminders just to hear how T3's voice was synthesised.

4.4.2 Opinions on Generated Voices and KinVoice: When discussing with T1 about T2's synthesised voice, she felt that it had elements of T2's voice: *"There's elements of similarity"* but it sounded *"Americanised"*. She added *"I think it sinks in unconsciously, instead of it being too jarring, it's easier to hear...You're used to hearing a familiar voice. When it's familiar, you tend to pay a bit more attention, 'cause you're used to it."* We found that T2 would usually call her on Wednesday evenings; T2 would remind her of events and initiates exercise at the gym. T1 remarked *"...in fact, today she called me to go for a walk with her."* Therefore, synthesised kin voice made her more attentive to the reminder and the role of KinVoice seems to match the reminding role that T2 takes.

T2 commented on the synthesised voice: *"Do I really sound like that?...our voices as we hear them always sound different than when we hear it played out."*. She added: *"At least it doesn't sound too American"* and that the quality of the voice was a little slurred. When we discussed how the generated voices could be made less slurred and more realistic through having more voice samples, she remarked that having it more realistic "would be very weird". She emphasised that potentially having the voice too realistic might be dangerous as well and she preferred to still be able to distinguish the difference in the voices.

T3 commented that she could recognise that the synthesised voice was T4's voice but it was different in terms of accent. Upon further thought, she said *"I think it's really clever what you've done. Because you've taken [T4's] voice from whatever he recorded and somehow used it in words that you probably didn't"*. T4 continued with her point, saying *"Instead of having us say a whole dictionary of words, you've managed to sample us and put a bit of effect over the generated voice"* .

T4 said (while smiling) that the synthesised voice of T3 *"was quite gruff and almost sounds annoyed"*, T3 interjected (while laughing), *"But that's how I always sound"*. He did note that he could pick up T3's intonations and it was more similar to T3 in the first half of the message with the second half turning into a different accent.

Overall, our interactions with T1 and T2 sparked implications in terms of attentiveness, role matching and dangers of voice realism. T3 and T4 appreciated the ability to generate dynamic content using each other's synthesised voices using only a short voice sample.

4.4.3 Co-Design Activity: In the co-design activity, T1 and T2 thought about how they would at times be unsure about what to cook for meals. They imagined using each other's voices to make suggestions for meals and having the voice interface to guide them through as they cook. T1 recounted how one of her cats reacted to T2's synthesised kin voice when the reminder audio

was played. This sparked a conversation on how they could use their own voices to remotely communicate with their pets. T1 mentioned that she could tell her cats that she would be coming back home soon to feed them. T2 remarked that she could have her synthesised voice assist in giving instructions to her dog when she is away from home.

In the co-design activity, T3 and T4 talked about how they had different routines but try to communicate with each other their plan for the day so that they meet for activities together like going for walks, exercising and having dinner. They thought if having a communication device like an inter-comm that could automatically pass messages and daily plans to each other. T4 had a thought that instead the voice interface only saying *'Go to the gym"* when reminding in T3's voice, it could be rephrased into *"Let's go to the gym together"* to give it a feeling of a joint appointment. Another area of interest was regarding how they often missed their daughter and granddaughter, wishing to hear their voices more often and because they believed the synthesised kin voices could *"make you feel like they are still around"* (T3). T3 was interested in the occasional use of her granddaughter's voice to greet her or tell her a joke. They extended this idea by identifying that their generated voices could be periodically played on the digital noticeboard in their daughter's home to encourage their daughter and granddaughter to call them or inform that they were thinking of them. Like T1 and T2, T3 and T4 were adamant about user burden in having to trigger the intent to play the reminder audio, believing that it should play out immediately when reminding: *"It's just the double step that puts me off doing it."* (T3). This point on usability was a key challenge that needs to be overcome.

To summarise, T1 and T2 ideated two new use areas for supporting meal preparation and interactions with their pets. T3 and T4 formed ideas to support communication between themselves, and promote communication between them and their child/grandchild.

5 DISCUSSION

From our findings and previous research, we discuss implications for voice design and kin voice in VUI design which we detail in this section.

5.1 Rethinking Voice Realism

Content and accent mismatches could undermine the sense of realism for generated kin voices. The mismatch in content between what the kin would normally say and the content generated in the VUI could lead to distrust (Study 1). Probe users (Study 2) also felt a mismatch in the perceived accent [12] of synthesised kin voice and their pair's (T1, T3, T4) or their own accent (T2). This is likely due to the familiarity between the user pairs, where they are familiar with each other's voices. It is also likely due to the kin voices using read speech which tends to be pronounced more formally than casual conversational speech [46]. Although it would be challenging to achieve accent matching practically, it should be considered as it affects the sense of familiarity and hence, the connection users might have with the VUIs.

With the advancements of voice synthesis and language models, content and accent mismatches could soon be fully addressed. To match accents, the accent and voice qualities could first be identified. Then, developers could select a more suitable voice data set to train the voice synthesis engine. To match content, language models like GPT-3 [9] are beginning to be able to generate speech content that is more human-like and believable. However, in a practical sense, training new language models from new data sets would require high computational resources and might not be scalable.

We argue that the strive for increased realism and humanness of the synthesised voices might not be necessary in the case of kin voice VUIs. Voice impersonation attacks [47, 53] could be a major source of misuse of kin voice VUIs where the attacker could use the synthesised voice of the

victim's kin to ask for important personal information, such as bank details and passwords, or trick the victim into transferring funds to the attacker's account. Since kin voices are more engaging and persuasive (Study 1), people are more likely to comply when requested through such VUIs. Attackers might even use kin voices to pass voice identification systems and change the victim's personal details. As also discussed with T2 (Study 2), over-realism would be dangerous if misused. Therefore, the content and accent mismatch might not be bad because they make the generated voice distinct and enable users to tell that the voice is not from the real person.

Our argument builds upon the suggestions from Cambre et al. [12] and Sutton et al. [67] to consider reducing human-likeness of the generated voices. We add that we should consider finding the balance between how human-like and how robotic the generated voices should sound. The kin voice should not be too real so as to prevent potential impersonation attacks but it should also not be too unrecognisable or unfamiliar that the user would not apply the positive attitudes (found from Study 1 and 2) to the VUI. Furthermore, as inspired by the diversification design strategy by Sutton et al. [67], we could alternatively and deliberately use a diverse range of data sets (covering many accents) to train a language and voice model, and build a TTS engine with a new accent. As such, the synthesised voices from it would be 'accent-neutral' and might prevent reinforcement of negative stereotypes. This strategy could be embedded in the kin voice design itself, serving as a deterrent to the misuse and impersonation attacks while supporting diversification.

5.2 Use Kin Voices for Personalised Tasks

According to previous work, users expect human-like qualities from voice interfaces if they sound close to human [13] and it has been argued that computers do not need to exhibit high realism to be treated as social beings [52]. Our probe showed evidence to support these views (Study 2). T1 was the most positive of the users with the system, likely because T2 usually reminds her of things in the first place. Therefore, the functions and qualities of KinVoice, in some way, matched her expectations as it 'embodied' the persona and social role of T2. This meant that even with a less realistic-sounding synthesised voice, T1 might have somehow perceived KinVoice as how she would perceive T2. To an extent, this aligns with the CASA theory. In another perspective, perhaps the familiarity between kin primes us to form human-machine social scripts [30] that are close to human-kin scripts. The idea of supporting meal selection and preparation generated from T1 and T2 shows their need for such role matching. It also aligns with Study 1 interview results that the scenarios were thought to be useful to use with 'kin' voices if there was strong role matching (Section 3.9.7).

Therefore, our combined findings showed that scenarios, which were personal in nature and enabled role matching, were suitable and useful for VUIs using kin voices. Usage scenarios should directly relate to your kin (e.g., reading messages from your kin in their voice) or match what your kin usually does for you (e.g., your kin usually navigates you while driving). Other usage scenarios and tasks which were factual and impersonal (e.g., reading weather information) would not benefit much from using kin voices.

From T4's realisation in the co-design activity (Study 2) that the kin voice from the VUI should say "Let's go together..." as going to the gym was usually a joint activity between T3 and T4, we add to our findings on content matching that there should also be a match in the context of the activity and the content. Inspired by the individualisation design strategy by Sutton et al. [67] and to support role and content matching, VUIs using kin voices could offer customisation for users to select and match different voices to different usage scenarios or based on the context of use [12, 67] (location, activity, etc.). We imagine interfaces where users would invite their friends to share their voice. Each user would have multiple voice profiles of friends and relatives. We acknowledge that

this opens the potential for people to use others' voices without consent and hence, a form of authentication and consent process is needed.

5.3 Mix the Use of Synthetic and Real Voices If Needed

Unlike the KinVoice technology probe, the kin voices used could be real voices instead of synthesised ones. The system could be implemented with recordings of reminders which are crowd-sourced from friends or relatives. The use of real, recorded kin voices would be more personal and meaningful. However, a drawback might be that your kin themselves might forget to record the reminder. As noted by Gong et al. [31], synthetic speech is more suited for dynamic content like messages and emails, while human voices are better suited for fixed content as it requires minimal input to set up. They also found that both synthetic and human voices have their own set of advantages. Users interacting with interfaces that used synthetic speech performed tasks significantly better, while users who interacted with a mixed-speech interface felt that they performed better and responded more positively. Pearson et al. [61] found that for question-and-answer VUIs, human answers were more relevant than machine answers, and machine answers were instant while human answers were delayed. They proposed a hybrid interface with machine-generated and human-curated answers. As the quality and naturalness of synthesised kin voices improve over time, synthesised and real kin voice may become indistinguishable [69]. Synthesised kin voices may have the same effect on users as real kin voices. However, the voice and speech content from kin may continue to be personal and meaningful as the kin has put in time, effort and thought into recording the speech. Thus, we agree with and advocate for the combination of using synthetic and human kin voices [31, 33] in a hybrid voice interface.

5.4 Build on Attributes of Kin Voices

Therapists, Companions and Confidants: The feelings of connection, safety and comfort that participants had while interacting with the VUIs that kin voices could make them useful as virtual therapists and companions. VUIs could indeed act as therapists which promote self-compassion [48]. Emotional disclosure through speech could be ideal for supporting self-compassion as it resulted in higher improvements in cognitive change, self-esteem and adaptive coping strategies compared to disclosure through writing [26, 48]. Additionally, people have been shown to be more willing to disclose their emotions with virtual interfaces due to the lack of judgement by them [45, 48]. People might share or disclose with VUIs with kin voices in a way that we normally would not with the actual person. Thus, such VUIs could serve as a confidant or a way to practice saying things before telling them to the real person. From Study 1, it might be awkward to use the VUI when the other person is there and boring and annoying if heard too frequently. An opposing view is that it might be fun and could boost the relationship. Kin voice VUIs could be used in private settings in which only the individual user could interact with it and synthetic voices should be generated in moderation. It might also work well for those who are seeking more intimacy, or to maintain their relationship.

Notifications: Kin voices from VUIs seem to draw attention, be noticeable and be charismatic, making them suitable for use when immediate attention is required. T1 mentioned how the synthesised kin voice made her more attentive to the reminders and that it was easier to hear (Study 2). This aligns with our results from Study 1 that kin voices were noticeable, persuasive and charismatic. Our discussion with her on how it might be an unconscious process seems to be in line with a proven and previously discussed phenomenon [67]. Tying back with the study by Bhatia et al. [7] on voice notifications in instructors' and lab mates' familiar voices, we add that kin voices could be useful for notifications too. Since kin voices in VUIs were also seen as persuasive and credible, notifications

could be issued for motivating users to form a habit or to do an activity such as exercising (Study 2).

Shared Use: In Study 2, T3 and T4 discussed how they could use their daughter's and granddaughter's voices to play a greeting, say a joke, or say a reminder. T3 said that it would make her feel like they were around. They identified that KinVoice systems could use one of its key attributes of telepresence (which we also found from Study 1). The ideas from T1 and T2 regarding remote commands and communication to their pets seem to utilise this attribute as well. Therefore, the attributes of telepresence, co-presence and social presence of kin voices could be used to support and promote intra-family and intra-household communications and activities. VUIs playing kin voices could help users feel like their kin are there with them and could encourage actual social interactions between the users. Since VUIs in home settings are often embedded into the family conversations and activities [62], KinVoice could support multi-user interactions where different voices could be used depending on the identified user(s) or activity. KinVoice could allow users to set a reminder on one device and ask it to issue the reminder on the other device or on both devices (Study 2). These applications would be particularly helpful when the users are physically distant or when they miss each other, giving them comfort and reducing loneliness. As social distancing has become the new norm, KinVoice holds the potential to help bridge this physical distance. The idea of supporting telepresence might be applied to using voices of kin who have passed away or for individuals with degenerate speech disorders [70]. This also helps to realise the idea of using parents' voices to read bed time stories to their children [58]. On a wider scope, we could consider shared community use of the VUI such as for a public speech-based question and answer system as seen in the work of Pearson et al. [61]. The instant machine-powered answers could be from synthesised voices of familiar local community members. This could help improve the users' perception of machine answers and encourage more advice or personal queries from their proposed hybrid machine-human respondent interface.

6 LIMITATIONS AND FUTURE WORK

We acknowledge that our findings may not be representative of a wider population. Many of our participants might be early adopters to new technology. Future studies could investigate with people who are less familiar with technology. Moreover, our current probe explored older adults due to the relevance of and need for reminder systems as well as to ground our work with previous studies looking at voice interfaces [38, 42] and memory support for elderly populations [14, 15]. It would be good to explore other pairings and age groups. Since the usage time with the VUIs is a key factor in the development of human-computer social scripts (extended CASA) [30], longer-term studies could be conducted.

The prompts and command-based interactions in our studies might still be non-natural and still adhere to predetermined structures as seen in prior work [19, 22, 62]. Future studies could employ a wider selection of prompts and commands with better-designed requests and responses [62] to promote more natural interactions. There was a possibility that participants found out that their study partners had also recorded their voices. They might have expected to hear their partners' voices or suspected that one of the voices they heard was a recording of their partner. This might have affected Study 1's results in which partners might more positively rate the kin voices. This risk was minimised by referring to the voices as A or B in the study session.

Accent: The synthesised kin voices kept being pointed out as sounding "Americanised" (T1 and T2) or different from their own accent (T3 and T4). To our knowledge, the pre-trained models for the TTS engine [40] for KinVoice were likely trained on various online-available samples [39] which might not all contain American English samples. In future studies with more controlled conditions,

we plan to use voice models that were trained on regional voice samples to minimise the risk of confounding results.

Security: There was a clear user burden of having to ask KinVoice play the reminder audio instead of the generated audio playing immediately at the reminder time, adding an extra step to the interaction process. With an Amazon Echo device, there is no current way to avoid this step as it is a security and privacy safeguard to prevent the automatic playing of audio that could be used for malicious intent. It is the same case for other present-day commercial smart speaker devices like Google Home. Developers and designers should be aware of this safeguard and potential barrier, not only for designing for VUIs with kin voices but also for designing with VUIs on commercial smart speakers. This issue could be avoided by using a custom-made VUI, but it may not have the same functionality and performance of quality speech recognition and services provided by commercial speakers. Thus, there is a trade-off between usability/function and security.

Technical: As the quality of voice synthesis technology that rely on small training sets improves over time [65, 69], we believe that our design guidelines will still hold. In fact, it will become more critical that we think about security and prevention of misuse of over-realistic synthetic voices. Without facing the technical limitation of lowered naturalness in KinVoice in Study 2, we would not have seen that a balance between naturalness and non-naturalness is needed. The open questions worth investigating are 'How natural should the voice be?' and 'At which point does the voice sound familiar?' Human-based verification combined with algorithm-based verification [47] can be utilised to give naturalness and familiarity scores to the generated voices. Then, any correlation between naturalness and familiarity may be modelled. The improvement of the quality, access, and ease of use of the tools will make it even easier to implement the proposed applications of kin voice VUIs. Future work could examine user perceptions on quality synthetic kin voices versus synthetic generic voices.

Future Directions: Our research explored how voice familiarity with added factors of kinship and friendship would be useful in VUIs. Future research could be directed toward other forms of familiar voices with added factors such as admiration; from voices of inspirational figures, mentors and community leaders. We advocate potential positive applications, for example, in social robots, virtual avatars, memory assistants, intelligent tutors and pedagogical agents. And advise caution on controversial uses like advertisements.

7 CONCLUSION

Our work brings a new understanding of how users perceive kin voices (voices of friends and family) in voice interfaces. We found that kin voices are engaging and comforting, promoting connection and the feeling of safety. Further analysis revealed that kin voices are persuasive, credible and charismatic, and any perceived eeriness, in fact, draws attention to the interface. Our findings suggest that synthesised kin voices should not be overly realistic but they should be recognisable and familiar. VUIs with kin voices are beneficial for personalised tasks which match the kin's role and content of interactions with the kin than for general tasks. We recommend leveraging the attributes resulting from the close familiarity of kin voices in virtual therapists and companions, notifications and motivation, and shared and social settings. Kin voices are exciting alternatives to common virtual assistant voices. Our studies and discussions on such VUIs contribute insights and guide future research on voice interfaces and voice design.

ACKNOWLEDGMENTS

We dedicate this work to Annette Epps. We want to thank Vipula Dissanayake for his assistance in the deployment of KinVoice, Sankha Cooray, Siddhesh Suthar and Jiashuo 'Evan' Cao for their support during the development of KinVoice. We thank Yvonne Chua for designing the figures. We thank our participants for their support and feedback. We would also like to thank the reviewers for their valuable suggestions that strengthened this paper. This work was supported by *Assistive Augmentation* research grant under the Entrepreneurial Universities (EU) initiative of New Zealand.

REFERENCES

[1] Amal Abdulrahman, Deborah Richards, and Ayse Aysin Bilgin. 2019. A Comparison of Human and Machine-Generated Voice. In *25th ACM Symposium on Virtual Reality Software and Technology* (Parramatta, NSW, Australia) *(VRST '19)*. Association for Computing Machinery, New York, NY, USA, Article 41, 2 pages. https://doi.org/10.1145/3359996.3364754

[2] David Airehrour, Samaneh Madanian, and Alwin Mathew Abraham. 2018. Designing a memory-aid and reminder system for dementia patients and older adults. *Proceedings of the 17th International Conference on INFORMATICS in ECONOMY* (2018), 75–81.

[3] Matthew P. Aylett, Selina Jeanne Sutton, and Yolanda Vazquez-Alvarez. 2019. The Right Kind of Unnatural: Designing a Robot Voice. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 25, 2 pages. https://doi.org/10.1145/3342775.3342806

[4] Matthew P. Aylett and Yolanda Vazquez-Alvarez. 2020. Voice Puppetry: Speech Synthesis Adventures in Human Centred AI. In *Proceedings of the 25th International Conference on Intelligent User Interfaces Companion* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 108–109. https://doi.org/10.1145/3379336.3381478

[5] Christoph Bartneck, Dana Kulić, Elizabeth Croft, and Susana Zoghbi. 2009. Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International journal of social robotics* 1, 1 (2009), 71–81. https://doi.org/10.1007/s12369-008-0001-3

[6] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 3, Article 91 (Sept. 2018), 24 pages. https://doi.org/10.1145/3264901

[7] Saurabh Bhatia and Scott McCrickard. 2006. *Listening to Your Inner Voices: Investigating Means for Voice Notifications.* Association for Computing Machinery, New York, NY, USA, 1173–1176. https://doi.org/10.1145/1124772.1124947

[8] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2018. *Thematic Analysis.* Springer Singapore, Singapore, 1–18. https://doi.org/10.1007/978-981-10-2779-6_103-1

[9] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[10] Joao Paulo Cabral, Benjamin R. Cowan, Katja Zibrek, and Rachel McDonnell. 2017. The Influence of Synthetic Voice on the Evaluation of a Virtual Character.. In *INTERSPEECH*. 229–233. https://doi.org/10.21437/Interspeech.2017-325

[11] 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 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376789

[12] 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, CSCW, Article 223 (Nov. 2019), 19 pages. https://doi.org/10.1145/3359325

[13] Julia Cambre, Samantha Reig, Queenie Kravitz, and Chinmay Kulkarni. 2020. "All Rise for the AI Director": Eliciting Possible Futures of Voice Technology through Story Completion. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference* (Eindhoven, Netherlands) *(DIS '20)*. Association for Computing Machinery, New York, NY, USA, 2051–2064. https://doi.org/10.1145/3357236.3395479

[14] Samantha W. T. Chan, Shardul Sapkota, Rebecca Mathews, Haimo Zhang, and Suranga Nanayakkara. 2020. Prompto: Investigating Receptivity to Prompts Based on Cognitive Load from Memory Training Conversational Agent. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 4, 4, Article 121 (Dec. 2020), 23 pages. https://doi.org/10.1145/3432190

[15] Samantha W. T. Chan, Haimo Zhang, and Suranga Nanayakkara. 2019. Prospero: A Personal Wearable Memory Coach. In *Proceedings of the 10th Augmented Human International Conference 2019* (Reims, France) *(AH2019)*. Association for Computing Machinery, New York, NY, USA, Article 26, 5 pages. https://doi.org/10.1145/3311823.3311870

[16] Fermin Chavez-Sanchez, Gloria Adriana Mendoza Franco, Gloria Angelica Martínez de la Peña, and Erick Iroel Heredia Carrillo. 2020. Beyond What is Said: Looking for Foundational Principles in VUI Design. In *Proceedings of the 2nd*

Conference on Conversational User Interfaces (Bilbao, Spain) *(CUI '20)*. Association for Computing Machinery, New York, NY, USA, Article 28, 3 pages. https://doi.org/10.1145/3405755.3406145

[17] Emna Chérif and Jean-François Lemoine. 2017. Human vs. synthetic recommendation agents' voice: The effects on consumer reactions. In *Marketing at the Confluence between Entertainment and Analytics*. Springer, 301–310.

[18] Emna Chérif and Jean-François Lemoine. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing (English Edition)* 34, 1 (2019), 28–47.

[19] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300705

[20] Benjamin R. Cowan, Holly P Branigan, Mateo Obregón, Enas Bugis, and Russell Beale. 2015. Voice anthropomorphism, interlocutor modelling and alignment effects on syntactic choices in human- computer dialogue. *International Journal of Human-Computer Studies* 83 (2015), 27–42.

[21] Paul Dourish. 1996. Book Review - The Media Equation: How People Treat Computers, Television and New Media Like Real People and Places. Retrieved 2021-04-15 from https://www.dourish.com/publications/media-review.html

[22] Philip R. Doyle, Justin Edwards, Odile Dumbleton, Leigh Clark, and Benjamin R. Cowan. 2019. Mapping Perceptions of Humanness in Intelligent Personal Assistant Interaction. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services* (Taipei, Taiwan) *(MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[23] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. *arXiv preprint arXiv:1807.04606* (2018).

[24] e-pill LLC. 2020. 25 Alarm Clock Reminder Rosie Reminder. Retrieved 2020-09-01 from https://www.epill.com/italkclock.html

[25] Carrie Demmans Epp, Cosmin Munteanu, Benett Axtell, Keerthika Ravinthiran, Yomna Aly, and Elman Mansimov. 2017. Finger Tracking: Facilitating Non-Commercial Content Production for Mobile e-Reading Applications. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Vienna, Austria) *(MobileHCI '17)*. Association for Computing Machinery, New York, NY, USA, Article 34, 15 pages. https://doi.org/10.1145/3098279.3098556

[26] Brian A. Esterling, Michael H. Antoni, Mary Ann Fletcher, Scott Margulies, and Neil Schneiderman. 1994. Emotional disclosure through writing or speaking modulates latent Epstein-Barr virus antibody titers. *Journal of consulting and clinical psychology* 62, 1 (1994), 130.

[27] 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 *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 125–126.

[28] Kerstin Fischer, Oliver Niebuhr, Lars C. Jensen, and Leon Bodenhagen. 2019. Speech Melody Matters—How Robots Profit from Using Charismatic Speech. *ACM Transactions on Human-Robot Interaction (THRI)* 9, 1 (2019), 1–21.

[29] Loretta M. Flaherty. 2004. Personal sensory reminder with customizable voice message. US Patent 6,707,383.

[30] Andrew Gambino, Jesse Fox, and Rabindra A. Ratan. 2020. Building a stronger CASA: extending the computers are social actors paradigm. *Human-Machine Communication* 1, 1 (2020), 5.

[31] 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. *International Journal of Speech Technology* 6, 2 (2003), 123–131.

[32] Jennica Grimshaw, Tiago Bione, and Walcir Cardoso. 2018. Who's got talent? Comparing TTS systems for comprehensibility, naturalness, and intelligibility. *Future-proof CALL: language learning as exploration and encounters–short papers from EUROCALL* (2018), 83–88.

[33] Randy Allen Harris. 2004. *Voice interaction design: crafting the new conversational speech systems*. Elsevier.

[34] Chin-Chang Ho and Karl F. MacDorman. 2010. Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior* 26, 6 (2010), 1508–1518. https://doi.org/10.1016/j.chb.2010.05.015 Online Interactivity: Role of Technology in Behavior Change.

[35] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.

[36] Apple Inc. 2011. Siri. Retrieved 2020-09-01 from https://www.apple.com/siri/

[37] Amazon.com Inc. 2014. Alexa. Retrieved 2020-09-01 from https://www.amazon.com/b?node=17934671011

[38] S. Jawaid and Rachel Mccrindle. 2016. Computerised help information and interaction project for people with memory loss and mild dementia. *Journal of Pain Manage* (2016), 269–272.

[39] Corentin Jemine. 2019. Master thesis: Automatic Multispeaker Voice Cloning. (2019). https://matheo.uliege.be/handle/2268.2/6801
[40] Corentin Jemine. 2019. Real-Time-Voice-Cloning. https://github.com/CorentinJ/Real-Time-Voice-Cloning.
[41] Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. 2018. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In *Advances in neural information processing systems*. 4480–4490.
[42] Alexandra König, Aarti Malhotra, Jesse Hoey, and Linda E. Francis. 2016. Designing personalized prompts for a virtual assistant to support elderly care home residents.. In *PervasiveHealth*. 278–282.
[43] Letty Y. Y. Kwan, Suhui Yap, and Chi-yue Chiu. 2015. Mere exposure affects perceived descriptive norms: Implications for personal preferences and trust. *Organizational Behavior and Human Decision Processes* 129 (2015), 48–58.
[44] Google LLC. 2016. Google Assistant. Retrieved 2020-09-01 from https://assistant.google.com/
[45] Gale M. Lucas, Jonathan Gratch, Aisha King, and Louis-Philippe Morency. 2014. It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior* 37 (2014), 94–100.
[46] Miriam Meyerhoff. 2006. *Introducing Sociolinguistics*. Routledge.
[47] Dibya Mukhopadhyay, Maliheh Shirvanian, and Nitesh Saxena. 2015. All your voices are belong to us: Stealing voices to fool humans and machines. In *European Symposium on Research in Computer Security*. Springer, 599–621.
[48] P. Muppirishetty and Minha Lee. 2020. Voice User Interfaces for mental healthcare: Leveraging technology to help our inner voice. 3rd ACM Conference on Computer-Supported Cooperative Work and Social Computing, CSCW 2020 ; Conference date: 17-10-2020 Through 21-10-2020.
[49] Clifford Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA.
[50] Clifford Nass and Kwan Min Lee. 2001. Does computer-synthesized speech manifest personality? Experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of experimental psychology: applied* 7, 3 (2001), 171.
[51] Clifford Nass, Youngme Moon, Brian J. Fogg, Byron Reeves, and Chris Dryer. 1995. Can computer personalities be human personalities?. In *Conference companion on Human factors in computing systems*. 228–229.
[52] 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* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. https://doi.org/10.1145/191666.191703
[53] Ajaya Neupane, Nitesh Saxena, Leanne M. Hirshfield, and Sarah E. Bratt. 2019. The Crux of Voice (In) Security: A Brain Study of Speaker Legitimacy Detection.. In *NDSS*.
[54] Andreea Niculescu, George M. White, See Swee Lan, Ratna Utari Waloejo, and Yoko Kawaguchi. 2008. Impact of English regional accents on user acceptance of voice user interfaces. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*. 523–526.
[55] Kristine Nowak. 2001. Defining and differentiating copresence, social presence and presence as transportation. In *Presence 2001 Conference, Philadelphia, PA*. Citeseer, 1–23.
[56] Kristine Nowak and Frank Biocca. 2003. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments* 12, 5 (2003), 481–494.
[57] R. Orpwood, C. Gibbs, T. Adlam, R. Faulkner, and D. Meegahawatte. 2005. The Design of Smart Homes for People with Dementia—User-Interface Aspects. *Univers. Access Inf. Soc.* 4, 2 (Dec. 2005), 156–164. https://doi.org/10.1007/s10209-005-0120-7
[58] Sunjeong Park and Youn-kyung Lim. 2020. Investigating User Expectations on the Roles of Family-shared AI Speakers. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[59] Malcolm R. Parks and Kory Floyd. 1996. Meanings for closeness and intimacy in friendship. *Journal of Social and Personal Relationships* 13, 1 (1996), 85–107.
[60] Pat Pataranutaporn, Tomás Vega Gálvez, Lisa Yoo, Abishkar Chhetri, and Pattie Maes. 2020. Wearable Wisdom: An Intelligent Audio-Based System for Mediating Wisdom and Advice. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–8. https://doi.org/10.1145/3334480.3383092
[61] Jennifer Pearson, Simon Robinson, Thomas Reitmaier, Matt Jones, Shashank Ahire, Anirudha Joshi, Deepak Sahoo, Nimish Maravi, and Bhakti Bhikne. 2019. StreetWise: Smart Speakers vs Human Help in Public Slum Settings. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300326
[62] Martin Porcheron, Joel E. Fischer, Stuart Reeves, and Sarah Sharples. 2018. Voice interfaces in everyday life. In *proceedings of the 2018 CHI conference on human factors in computing systems*. 1–12. https://doi.org/10.1145/3173574.3174214

[63] Byron Reeves and Clifford Nass. 1996. *The media equation: How people treat computers, television, and new media like real people*. Cambridge university press Cambridge, UK.
[64] Baidu Research. 2020. A Look Back on Baidu's AI Innovations in 2019. Retrieved 2020-09-01 from http://research.baidu.com/Blog/index-view?id=130
[65] Resemble. 2019. Resemble AI. Retrieved 2021-04-15 from https://www.resemble.ai/
[66] 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. *International Journal of Human-Computer Studies* 64, 1 (2006), 43–52.
[67] 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 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3290605.3300833
[68] 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. *International Journal of Social Robotics* 3, 3 (2011), 253–262.
[69] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. 2021. A Survey on Neural Speech Synthesis. *arXiv preprint arXiv:2106.15561* (2021).
[70] Christophe Veaux, Junichi Yamagishi, and Simon King. 2013. Towards personalised synthesised voices for individuals with vocal disabilities: Voice banking and reconstruction. In *Proceedings of the Fourth Workshop on Speech and Language Processing for Assistive Technologies*. 107–111.
[71] Xi Yang, Marco Aurisicchio, and Weston Baxter. 2019. *Understanding Affective Experiences with Conversational Agents*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300772

Received January 2021; revised April 2021; accepted July 2021