Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges

Yelim Kim
University of British Columbia
Vancouver, BC, CA
yelim27@student.ubc.ca

Mohi Reza
University of British Columbia
Vancouver, BC, CA
mohireza@alumni.ubc.ca

Joanna McGrenere
University of British Columbia
Vancouver, BC, CA
joanna@cs.ubc.ca

Dongwook Yoon
University of British Columbia
Vancouver, BC, CA
yoon@cs.ubc.ca

ABSTRACT

This work investigates the practices and challenges of voice user interface (VUI) designers. Existing VUI design guidelines recommend that designers strive for *natural* human-agent conversation. However, the literature leaves a critical gap regarding how designers pursue naturalness in VUIs and what their struggles are in doing so. Bridging this gap is necessary for identifying designers' needs and supporting them. Our interviews with 20 VUI designers identified 12 ways that designers characterize and approach naturalness in VUIs. We categorized these characteristics into three groupings based on the types of conversational context that each characteristic contributes to: Social, Transactional, and Core. Our results contribute new findings on designers' challenges, such as a design dilemma in augmenting task-oriented VUIs with social conversations, difficulties in writing for spoken language, lack of proper tool support for imbuing synthesized voice with expressivity, and implications for developing design tools and guidelines.

CCS CONCEPTS

• **Human-centered computing** → **HCI theory, concepts and models**; **Natural language interfaces**; **Interaction design theory, concepts and paradigms**.

KEYWORDS

naturalness, voice user interfaces (VUI), designers, characterize, practices, challenges

ACM Reference Format:
Yelim Kim, Mohi Reza, Joanna McGrenere, and Dongwook Yoon. 2021. Designers Characterize Naturalness in Voice User Interfaces: Their Goals, Practices, and Challenges. In *CHI Conference on Human Factors in Computing Systems (CHI '21), May 8–13, 2021, Yokohama, Japan.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3411764.3445579

CHI '21, May 8–13, 2021, Yokohama, Japan

ACM ISBN 978-1-4503-8096-6/21/05…$15.00
https://doi.org/10.1145/3411764.3445579

1 INTRODUCTION

With substantial industrial interest, conversational Voice User Interfaces (VUIs) [1] are becoming integral to the plethora of digital systems, from smartphones to smart homes, that feature voice agents such as Siri, Alexa, and Google Assistant. VUI designers are those who architect the conversational experience between the user and VAs. Hence, understanding designers' practices and challenges is of fundamental importance in providing them with useful design resources. In the VUI literature, however, designers are underrepresented as subjects of studies; by contrast, there's no lack of an existing body of literature on users' experiences [22, 27, 63, 88]. Overlooking VUI designers' perspectives is a significant impediment to providing comprehensive and practical VUI design support to them: Despite attempts to establish guidelines for VUIs [105, 112], previous studies have emphasized that existing VUI design guidelines are still immature [69–71, 73] and HCI/UX curricula are lacking detailed coverage for VUI design [72]. Our study focuses on design practices and challenges of VUI designers to inform the creation of better tools and guidelines to support them.

If there is one common motif in existing, albeit immature and fragmented, guidelines and standards for VUI design, it is that at the heart of the desired properties of VUIs is *naturalness*. In the field of VUI design, it has long been believed that the design of human-agent conversation should be modeled after practices and mechanics of naturally occurring human conversation, so the VUI can incorporate the beneficial attributes of natural languages, such as flexibility [69], intuitiveness [86], and accessibility [93]. Multiple VUI design textbooks and guidelines recommend that designers make VUIs that provide natural conversational experiences to the users [6, 40, 70], sound more natural [5, 29], feature natural dialogues [10, 41, 54], or offer natural interactions [9, 53]. Particularly in industry, there is a strong drive towards promoting naturalness as the holy grail of VUIs. Google's conversation design guidelines recommend designers to "craft conversations that are natural and intuitive for users." [41] In Amazon's "re:MARS", the company's

[1] Conversational VUI systems are one of the two general types of VUI systems [102]. In a conversational VUI system, users perceive the voice agents as conversation partners and accomplish their goals by having conversations with the agents [102]. While in a command-based VUI system, which is the other general type of VUI system, users are expected to learn and use the appropriate voice commands to accomplish their goals [46]. Hereafter, we use the term 'VUI' to refer to 'conversational VUI' and we use the term 'voice agent' to refer to 'conversational VUI agent' [43].

latest public-facing tech conference, "making conversation natural" is presented as the mantra of VUI that their voice assistant products aspire and strive for.

Despite the emphasis on the importance of naturalness, to our knowledge, there has been no systematic study uncovering designers' current practices and challenges in creating natural VUIs. There does exist a small number of articles and videos in the popular press where designers report some of their difficulties in creating natural VUIs [11, 65]. For example, David Attwater, with over 17 years of experience in VUI design, said *"the current tools do not come with any underlying knowledge of language, meaning, or interrelationships between words and concepts. Each app designer currently has to develop their own complete dialog and meaning models"* and he said he expects *"enhanced support for natural language"* in near future [65]. This suggests that better tool support is one area worth investigation. We focus on naturalness as the central notion in our investigation of VUI designers' practices, and also probe on challenges, including tool support.

We interviewed 20 VUI designers to answer questions on: *(1) how VUI designers approach naturalness in their design practices, (2) how they align such practices to their varying design goals, and (3) what challenges they face in striving to offer natural conversational experiences.* The basic conceptual, and partly methodological, premise in pursuing these inquiries is that the term naturalness should be handled as a *construct* that is subject to different opinions of individuals. In the literature of VUIs and conversational agents, individual scholars assigned multifaceted and fragmented meanings to the term, including the resemblance of linguistic features to interpersonal speaking style [22], spontaneous and open-ended nature of "naturally occurring conversation, as opposed to language restricted to a fixed set of commands and phrases." [68], and sounding more natural [5, 29]. Within the broader discourse on Natural User Interface (NUI) design, the preliminary conceptions of the term have remained as an abstract and generic property that refers to how the users "interact with and feel about a product" [114].

In this study, we interpreted the notion of naturalness in terms of the specific *design context as described by individual designers*, because the meaning of naturalness as a construct may vary according to their design context. Especially, given that modern voice assistants tend to be situated in complex and dynamic social settings [88], it is possible that conversational characteristics of a VUI considered natural in one setting are not perceived the same in another setting (e.g. an extremely human-like voice agent can be considered deceptively anthropomorphic and uncanny [60]). The designers in our interview study often regarded naturalness as VUI traits that make human-agent conversation natural. Using their descriptions as the basis, we crystallized such traits into sets of *naturalness characteristics* according to the types of human-VA dialogues that each characteristic contributes to.

The result of our study revealed 12 ways designers characterize and approach naturalness in VUIs. These characteristics are categorized depending on the conversational context: namely, Social for providing social dialogues, Transactional for supporting users' tasks, and Core for both. Most characteristics mirror those found in the human-to-human conversation literature, but some are "beyond-human", which reflect the machine-specific characteristics that outperform people, such as superior memory capacity and processing power. Our VUI designers also described significant challenges in achieving natural interaction related to a design dilemma in augmenting task-oriented VUIs with social conversations, difficulties for writing in spoken languages, and a lack of adequate design tools and guidelines (e.g., the poor design of existing SSML authoring interfaces).

Our work makes the following contributions:

(1) We characterize 12 ways in which designers pursue naturalness in their design practices and categorize them based on three different types of conversational context.
(2) We identify 7 challenges that hinder designers from creating natural VUIs.
(3) Our results generate implications for tools and design guidelines that support designers in creating natural VUIs.

2 RELATED WORK

Our research combines themes from VUI design, naturalness of interfaces, human-likeness of embodied agents, and existing supports for VUI design.

2.1 VUI Literature in HCI

Researchers have been investigating ways to support speech interactions since the 50s. With rapid advancements in Natural Language Understanding (NLU), VUI systems have transitioned from rudimentary speech-recognition based systems such as *Audrey* [32] and *Harpy* [62] in the 50s and 70s, to task-oriented systems like *SpeechActs* [117] in the 90s, and sophisticated conversational agents that are now pervasive.

More recently, several studies from the HCI community have been investigating how voice assistants impact users [12, 27, 88, 89]. In these studies, various issues have been explored, including how VUIs fit into everyday settings [88], how users perceive social and functional roles in conversation [27], and the disparity between high user expectations and low system capability [63].

A common theme with many of these studies is that they take into account the perspective of the users. As Wigdor [114] put it, naturalness is a powerful word because it elicits a range of ideas in *those who hear it*—in this study, we take the path less trodden, and see what *designers* think.

2.2 Discourses around Naturalness in HCI

There are several ways in which HCI research uses the term naturalness.

As a descriptor for human-likeness: Naturalness is often seen as a "mimicry of the real world" [114]. In the context of speech, humans are the natural entity of concern, and hence, *behavioral realism*, i.e. creating VUIs that behave and sound like real humans, has become a focus. We can trace the attribution of anthropomorphic traits onto computers in a seminal paper by Turing [107] on whether machines can think. In that paper, he assumes the "best strategy" to answer this question is to seek responses from machines that would be "naturally given by man". The pervasive influence of such thought can be seen in existing definitions of naturalness in VUI literature [22, 37, 76]— all treat naturalness in this light, as a pursuit of human-likeness.

As a distinguishing term between the novel and traditional modes of input: The term is also used to contrast interfaces that leverage newer input modalities such as speech and gestures, against more classical modes on input, namely, graphical and command-line interfaces [66]. In this definition, the term is in essence an umbrella descriptor of countless systems involving multi-touch [61, 103], hand-gestures [39, 58], speech [2, 15], and beyond [101, 120].

As interfaces that are unnoticeable to the user: Another usage draws from Mark Weiser's notion of transparency introduced in his seminal article on ubiquitous computing [113]. In this formulation, naturalness is a descriptor for technologies that "vanish into the background" by leveraging natural human capabilities [50, 108].

As an external property: In this conception, the term does not refer to the device itself, but rather the experience of using it, i.e. the focus is on what users do and how they feel when using the device [114]. The characteristics that we present in this paper can also be viewed from such an angle, i.e. designers form and utilize characteristics not because they make the VUI more natural, but rather the experience of using it more natural.

The existing usage of the term has drawn heavy criticism from some—Hansen and Dalsgaard [44] find the non-neutral nature of the term to be problematic. In their view, the term has been misused to conflate "novel and unfamiliar" products with "positive associations", akin to marketing propaganda.

Norman [81] contends the distinction between natural and non-natural systems and notes that there is nothing inherently more natural about newer modalities over traditional input methods. With speech, for example, he notes that utterances still have to be learned.

2.3 Characterizing Conversations

In the context of discourse analysis, human conversation can be largely classified into two categories based on its purpose: transactional conversations vs. social (interactional) conversation [23]. Transactional conversation is "message-oriented" and characterized as its function for transmitting factual or propositional information [64]. It is an interaction pursuing practical outcomes, for example, "buying something in a shop and enrolling in a school" [82]. Social (interactional) conversation is for "expressing social relations and personal attitudes" [23]. Examples of social conversations are "greetings, gossip, and social chat or small talk" [48]. The terminology "interactional" conversation has been more widely accepted than "social" conversation in linguistics. However, it was first introduced as social conversation by Clark et al.'s work [27] into our HCI community and since "interaction" is also a heavily overloaded terminology in our community, we use the term social conversation to be consistent with Clark et al.'s work.

Parallel to human spoken conversation having two classifications, conversational agent systems can also be classified into two categories: task-oriented systems vs. chatbot systems [59, 116]. A task-oriented system is "designed to assist users to achieve specific goals (e.g., finding hotels, movies, or bus schedules)" [31, 118], and its domain is focused on the predefined topics for achieving the goals [104], while a chatbot system has an open domain for having general conversations [67, 96]. Due to its goal-oriented nature, task-oriented systems heavily incorporate transactional conversations. Under the constraint of Allen's task-based dialogue hypothesis [4], most conversational agent systems were designed with a task-oriented slant for reasons of tractability [38]. However, cutting-edge conversational agent systems such as Apple Siri and Amazon Alexa combine both aspects of chat and task-based interaction [85, 90].

2.4 Difference Between Spoken Language and Written Language

Previous studies from linguistics have identified the differences between spoken and written languages [19, 92, 115]. Researchers found that people use more complex words for writing compared to when they are speaking [19, 34, 83, 115]. Bennett claims that passive sentences are more frequently used in written texts [13]. Also, more complex syntactic structures are used in written language than spoken language [34]. In our study, we analyzed our interview data based on these previous works to find specific aspects of spoken dialogues that VUI designers deem challenging to mimic when they are writing VUI dialogues.

2.5 Human-likeness in Embodied Agents

A rich body of studies explores issues around human-likeness in embodied agents and the relationship between humans and these agents. They investigate: a plethora of concerns such as ways to transfer human qualities onto machines [24, 80] and ways to maintain trust between users and computers [16, 97, 110], modeling human-computer relationships [1, 17], designing for different user groups such as older adults [98, 109], children [30, 87], understanding how users 'view humanness in dialogue interaction' [33], and examining stereotypes in this domain [26, 78]. Naas et al.'s "Similarity attraction hypothesis" posits that people prefer interacting with computers that exhibit a personality that is similar to their own [77] (e.g., cheerful voice agents can be undesirable to sad users).

In our study, designers reflect on issues that echo the literature by considering factors such as personality, trust, bias, and demographics in their VUI design practice.

2.6 Guidelines and Tools for VUI Design

Many large vendors of commercial voice assistants provide their own separate guidelines for designers [7, 41, 52]. These guidelines offer design advice tailored to developing applications for a specific platform. With regards to platform-independent options, some preliminary effort has been undertaken in the form of principles [95], models [74] and design tools [56] for VUIs. More specifically, Ross et al. provided a set of design principles for the VUI applications taking a role as a faithful servant [95], while Myers et al. analyzed and modelled users' behaviour patterns in interaction with unfamiliar VUIs [74].

Researchers have built several tools in support of VUI design. Klemmer et al.'s SUEDE enabled Wizard-of-Oz style prototyping of VUIs [56]. SPICE and SToNE are toolkits for helping researchers or non-experts develop speech recognizers for VUI applications [57, 100]. In order to help designers modify the synthesized voice in more effective and efficient ways, tech giants such as Amazon and IBM have developed their own high-level SSML (Speech Synthesis Markup Language) tags that comprise the effects from multiple primitive standard SSML tags (e.g., The "Good news" tag from IBM)

[8, 51]. SSML, in general, is an XML-based markup language for speech synthesis applications. Standard SSML enables primitive prosody alteration such as changing volumes to be louder or modifying pitches to be higher [91]. Apart from creating high-level SSML tags, to our knowledge, there hasn't been much support for helping designers to effectively communicate the desired expressions to the TTS engine. Yuan et al. showcased one of the few such supports in their relatively recent poster [36]. They developed a system that visualizes the vocal characteristics of synthesized voices and allows direct modifications of the characteristics.

Our study offers new design guidelines and recommendations for tool support.

3 METHOD

To understand how designers pursue and characterize naturalness in VUIs, we interviewed designers with a variety of VUI design experiences.

3.1 Participants

We recruited 20 VUI designers (7 women, 13 men) through flyers and study invitation messages on social network services such as Facebook and LinkedIn. Participants' ages ranged from 17 to 73 (M = 34.3, Median = 30.5, SD = 14.7). Their nationalities were as follows: 4 Americans, 1 Belgian, 1 Brazilian, 5 Canadians, 1 Dutch, 1 German, 5 Indians, 1 Italian, and 1 Mexican. Each participant received CAD $15/hour as compensation.

The participant pool included designers with diverse occupations and design experiences. 13 were professional VUI designers working full-time on VUI projects. Their job titles included: designer, UX manager, and CEO. Participants' length of professional VUI design experience ranged from 9 months to 20 years (M = 4 years and 2 months, Median = 2 years, SD = 6 years). Our participants worked in companies that ranged considerably in size. Most of the professional VUI designers we recruited (8 out of 13) were working for relatively small companies (2-49 employees), while 2/13 were working for medium-sized companies (50-999 employees) and the rest 3/13 were working for corporations with over 5k employees. The remaining 7 participants were amateur/hobbyist designers.

All of our participants had previously designed at least one conversational voice user interface. In total, we collected data about designers' practices in 38 different VUI projects. There were 27 Intelligent Personal Assistant (IPA) systems for smart home speakers (23 Amazon Alexa, 4 Google Home), 8 Interactive Voice Response (IVR) phone systems, 1 voice agent system for a smart air-conditioner, 1 voice agent system for a mobile application, and 1 voice agent system for a humanoid robot. Most of them were primarily task-oriented systems according to the definition provided by Yu et al. in [119]. We also collected information about their SSML familiarity. About half of the participants considered themselves to be familiar with SSML, while the other half indicated unfamiliarity.

3.2 Semi-structured Interviews

For each participant, we conducted a single-session semi-structured interview. Each interview lasted between 30 min to an hour. All of the interviews were conducted by the lead investigator. We arranged online interviews for 17 of them who could not visit the interview site. Before the interview, our designers answered an online survey that collected demographic data, previous VUI design experiences, and familiarity with VUI design tools such as SSML.

The dialogues in our semi-structured interviews were anchored to four seed questions, each of which was designed to capture the designers' perception of naturalness, specifically trying to get at their practices and challenges in designing natural VUI dialogues: (1) How do you define a natural VUI dialogue? (2) What is the expected value in creating a natural dialogue? (3) How do you create a dialogue to be more natural? (4) Was there any challenge in creating natural dialogues? If you have any, what was the challenge? The interviewer maintained the focus on naturalness. When a participant seemed to deviate, the interviewer actively checked with the participant by asking if their response is related to naturalness.

During the interview, we made an attempt to ask our designers to define naturalness but observed that they tend to ground their responses on their own design practices rather than on a generalizable conceptual description. Hence, we probed further into the design context, such as their design goals, strategies, and challenges, in relation to their pursuit of naturalness in VUIs. To elicit thick descriptions about their design practices, we asked them to ground their response on the one or two most memorable VUI projects that they reported in the pre-interview survey. For example, the participants were first asked what particular design steps they took to create more natural VUIs in one of their past VUI projects and then asked about the most challenging aspects of carrying out those steps.

3.3 Data Analysis

Audio recordings of all 20 interviews were fully transcribed before being analyzed. We used Braun and Clarke's approach for reflexive thematic analysis [20] for analyzing the interview data. Their approach was particularly suited for our study because of its theoretical flexibility and rigour. As advised in [21], we checked during the analysis phase if our interview data from 20 participants produced a compelling story answering our research questions. Three members of the research team had one-hour weekly meetings where we developed the themes over the course of several months. As Clarke and Braun suggested in [21], we took "researcher subjectivity as not just valid but a resource", thus actively discussing our own interpretations of the data during the meetings to form a "coherent story about the coded data" [28]. We took both inductive and deductive approaches for coding the data and developed a set of coherent themes that form the basis of our findings. Our deductive coding was derived from the previous works on "the classification of human conversation" [23, 35, 99].

4 FINDINGS: HOW DESIGNERS PURSUE NATURALNESS IN VUIS.

Our designers articulated their design practices, goals, and challenges when they strive to enhance naturalness in their VUIs. The designers expressed their assumptions of what conversational characteristics constitute naturalness in VUIs. In the analysis, we identified 12 different characteristics, each of which was associated to

Social	
1	Express Sympathy
2	Be Interesting, Charming, and Lovable
3	Express Interest to Users
Transactional	
4	Proactively Help Users
5	Be Capable of Handling a Wide Range of Topics in the Task Domain
6	Present a Task-appropriate Persona
7	Deliver Information With Machine-Like Speed and Accuracy*
8	Maintain User Profiles to Deliver Personalized Services*
Core	
9	Use Spoken Language Rather Than Written Language
10	Use Appropriate Prosody
11	Understand Variations in Human Language
12	Collaboratively Repair Conversation Breakdowns

Table 1: 12 characteristics of VUIs' naturalness in different conversational settings. Beyond-human characteristics are denoted with an asterisk (*).

specific design goals in pursuit of naturalness and accompanied practices.

It is worth noting that these characteristics are not monolithic. A thematic analysis revealed three categories for classifying these characteristics according to the varying conversational context in which the VUI is supposed to operate: (1) ones that promote desirable social dialogues, (2) those that help the user to accomplish their tasks, and (3) common traits which are generic to any natural conversations. The first two categories largely resemble classifications for human-to-human conversations in existing literature [23, 35, 99], labeled as "social conversation" and "transactional conversation". To be consistent with that literature, we adopted those labels; we label the third type as "core conversation." This categorization is summarized in Table 1.

We emphasize that the three categories are types of *conversations* in which the dialogue between a human and voice assistant (VA) is situated but not necessarily types of VUI applications. For example, there are needs for augmenting task-oriented VUI applications (e.g., IVR systems) with the traits beneficial for friendly and sympathetic conversations (see Section 5.1.1 for details.) We construed such cases as naturalness characteristics for *social* conversations. Our study framed *the characteristics of naturalness* using these three categories, which should not only help readers to conceptualize the 12 characteristics found in our study but can also serve as a lens to understand why designers pursue a particular subset of these characteristics for a VUI application and how these characteristics can often conflict with each other. The following section provides detailed descriptions of each characteristic.

4.1 Characteristics of Natural VUIs in *Social* Conversation

Our designers emphasized the importance of incorporating proper social conversation as part of harmonious and positive human-agent interactions. To increase user engagement on their services, they endeavor to provide the user with a realistic conversation and a feeling of being heard, just like humans building a positive relationship with one another by having social conversations [94]. Our designers especially highlighted the three conversational characteristics as follows.

4.1.1 Express Sympathy. Ten participants mentioned the importance of providing sympathetic responses to users' sentiments to maintain harmonious interactions. Most of the designers' elaborations on this part were focused on showing sympathy when the user experiences *negative* sentiments. They try to make the VAs console the user when the user feels negative or upset: *"If they respond negatively, [then] Alexa responds, 'Oh, I'm sorry to hear that.'"* (P4) Beyond being sympathetic, our designers even actively try to soothe the user's feelings in situations when they feel heightened emotions such as anger: *"You have a calm reassuring voice when they're upset because there's traffic."* (P9) One designer went further and suggested that VAs show empathy. P2 indicated that once voice tone technology becomes sufficiently advanced, he will want the VA's voice tone to adapt to empathize with users' moods: *"I would change [the voice tone] if I know your favorite team won, I'd have a happy voice. If I know your favorite team lost, I'd have a sad voice."*

To find out if the user feels negative, our designers use user responses, their profile information (e.g., how well their favorite sports teams perform), and the location of the conversations (e.g., hospital). Yet, several designers reported that incorporating sentiment analysis into the design process can require too much time commitment: *"I don't have time to know the APIs that can do sentiment detection."* (P11) If there was no way to detect users' real-time sentiments, then our designers chose to use a *"flat voice"* (P3) to prevent the happy voice of a VA from upsetting the user who is currently feeling down, as suggested in [79]: *"You have to control the tone of voice, because you can't sound very enthusiastic, things like that, because you never know the situation of the person on the other side."* (P3)

4.1.2 Be Interesting, Charming, and Lovable. Social conversations include humour and gossip which fulfill hedonic values [14, 84] that transactional conversations do not contain. In order to bring more user engagement for task-oriented applications, 4 participants reported trying to write more intriguing dialogues to create a charming persona: "[Being] *interactive means using some good words. Something which sounds interesting to the user."* (P17)

Depending on the VUI application contexts, using gentle and kind language may not be the best way to portray the VA as a charming persona. P7, who created an Alexa application for resolving the conflicts between children, mentioned that a charming VA can embody a charismatic persona by being sarcastic and funny, rather than being loving and nice: *"She's not loving and caring, but she's maybe a little sarcastic. She makes fun of what they say, and I would say she's lovable, not loving."* (P7) The importance of being entertaining was emphasized, especially when the target users are children: *"So, when it's a kid's application, you respond back in a very funny way. You use, terms like 'Okie Dokie'."* (P13)

4.1.3 Express Interest to Users. Four participants said that they incorporated greetings, compliments, and welcoming words that express interest to the user. These words make the conversations

appear *"real-ish"* (P11), and make the user feel their request is acknowledged: *"I think the benefit of providing this type of response, instead of just blank ones, is that it actually helps the person feel like their responses actually got heard."* (P4) P11 and P6 mentioned that VUIs can even make users feel as if they have personal connections to the applications by providing daily greetings or feedback on the user's actions: *"We could just say the recipe steps and all of that, and not have to ask questions like, 'How's the spice?' and all of that, but if we do, then there's some kind of personal connection."* (P6)

4.2 Characteristics of Natural VUIs in *Transactional* Conversation

For transactional conversations, our designers considered that the user experience will be the most natural when VUIs best accommodate the users' need for *getting things done*. It is worth noting that some of the conversation characteristics in this category were different from the qualities of conversation people normally expect from human-to-human conversations, as the designers had to prioritize task performance over conversational realism. To specify, it can be desirable for a VUI to exhibit *machine-like* speed and memory that extend beyond what human agents can offer. We label such characteristics as "Beyond-human characteristic".

4.2.1 Proactively Help Users. Eleven participants mentioned that a natural VUI should be efficient, and proactively *"detect or even ask for the things that* [the user] *needs."* (P12) In other words, a natural VUI should understand the meaning behind the statement and take action proactively to help users. P18's example is particularly illustrative of this point: *"From a linguistic perspective, 'Could you help me with my software?' is a yes-no question. 'I have a problem with my software.' is not even a question yet. So for 'I have a problem', bots need to be more proactive and ask a question, 'Could I help you with the software?'"* (P18)

VAs can proactively lead the conversation with the user to minimize the number of conversation turns, which helps reduce the user's overhead in responding to the series of queries. *"You should not overload the user with a lot of information. You should try to cut down as many decisions for the user as possible."* (P13) To do this, VA should be *"asking them* [users] *less and less and assuming more..."* (P13) To ask fewer questions, a natural VUI should make decisions based on contextual information: *"If the user tells me the zip code correctly, I don't ask him for city and state, I use some libraries to find the name of the city and state...We need to have a record of the entire conversation from top to bottom."* (P13)

Even though minimizing the number of questions is important, if the consequence of failing the task is considerable, a natural VUI should ask the user to confirm: *"...so if* [VA] *says things like 'You wanted your checking account. Is that correct?' and I say 'No, I want my savings account' then that to me, that confirm-and-correct* [strategy] *is a very important part in making it more conversational."* (P10)

4.2.2 Be Capable of Handling a Wide Range of Topics in the Task Domain. Nine participants mentioned that a natural VUI should not only be able to respond to the questions directly related to its task, but also be able to handle a wide range of topics within the domain of its task: *"I would think it* [a natural VUI] *would need to handle anything that is specific to that institution, right? If I call Bank of America and ask about my Bank of America go-card, you know you need to understand me."* (P10)

When a user brings a topic that is beyond the task domain handled by the voice agent, a natural VUI should still continue the conversation and remind the user about the task domain in which it can help with: *"...if a person says 'I want to order a pizza', and your skill*[2] *has no idea what that is...Give them a helpful prompt saying 'This is the senior housing voice assistant. I can help you with finding when the next bus is, or finding when the next garbage day is, or this or this."* (P2)

To help users be aware of the boundaries of the serviceable topic domain, the designers recommended preemptively providing context to users to help them understand what they can do with the application: *"A lot of people make a mistake in the design by saying 'Welcome to Toyota. How can I help you?' And it's like you're going to fail right there because that's so open-ended. No one will have an idea of what they can or can't say. They will probably fail. So you have to be really clear...like 'Welcome to Toyota's repair center! Would you like to schedule an appointment?'"* (P9)

4.2.3 Present a Task-appropriate Persona. Four participants said that a natural VUI application should present an appropriate persona for its target task. The tone of voice should match the application's purpose to increase user trust and elicit proper user responses. For example, P4 mentioned that the VA in financial applications should sound serious so as to portray a reliable persona: *"If you're talking about your wealth management, you don't want to have a fun guy. It has to be serious."* (P5) As another example, P4 designed an application for collecting elders' health status. He tried to make the VA sound like a real doctor to ensure that users take the conversation seriously and report their status correctly. *"...as if someone was visiting their doctor and asking the questions...it was better than making it seem like you were having a conversation with a friend, because it was kind of a serious topic dealing with...people would take it more seriously if they felt that it was a natural doctor, something like that."* (P4)

4.2.4 Beyond-human characteristic #1: Deliver Information With Machine-Like Speed and Accuracy. Our designers mentioned that, to accomplish its transactional tasks in an efficient manner, a natural VUI should incorporate machine-specific attributes such as high processing powers, and only selectively mimic certain parts of human conversation instead of pursuing every aspect of a natural human conversation. Specifically, P12 suggested that a natural VUI should attain the human-level ability to maintain conversational context while being able to deliver accurate information in a blazing fast manner: *"So it's just super-fast processing times, being able to deliver information while maintaining conversational context."* (P12)

This is where the designers' conceptions of naturalness in VUIs depart from what it means to be natural in human-to-human conversations. Our designers described natural human speech as often being indirect and inefficient, so these aspects of human conversation should be left out when designing for a natural VUI: *"Oh, no less conversational, because you don't want...something that you're using every day. You don't want to have that be chatty and friendly*

[2] Skill refers to a voice application that runs on Amazon's smart speakers.

right? You want to get your work done. So you know concentrating on being efficient and giving them the information and exactly the way that they want it." (P10)

4.2.5 Beyond-human characteristic #2: Remember User Profiles to Deliver Personalized Services. Human memories are volatile in contrast to machine memories. Designers mentioned that a natural VUI offering a transactional service should store a vast amount of information specific to the user, such as personal profiles or preferences to "customize all the knowledge of the user" (P8) and "personalize things and make things fit each user." (P6): *"Suppose you have an allergy or specific dietary requirements, then we could filter out all of those recipes and only suggest you the recipes that fit your needs."* (P6) Designers are aware that storing personal information comes with concerns about privacy. They highlighted the importance of *"be[ing] transparent to the user about the collected and stored data."* (P8)

4.3 Core Characteristics

Among the conversation characteristics mentioned by our designers, there was a set of basic elements that a natural VUI should have, regardless of whether the aim is to support social conversations or transactional conversations. We also noticed that most of these core characteristics echo suggestions in existing VUI design guidelines. We list these core characteristics briefly below for completeness, citing the relevant prior works.

4.3.1 Use Spoken Language Rather Than Written Language. Six participants mentioned that utterances of a natural VUI should have characteristics of spoken language as opposed to written text. For example, people tend to use more abstract words and complex sentence structures when writing [3, 106].

4.3.2 Use Appropriate Prosody. Eleven participants were aware that a natural VUI should convey non-verbal meaning with the appropriate prosody, including intonations, pauses, and stress. Their elaboration of design practices largely echoed the guidelines suggested by Cohen et al. [29] and Pearl [86].

4.3.3 Understand Variations in Human Language. Human language is immensely flexible, and we can express the same request in countless ways. Thirteen participants mentioned that a natural VUI should understand various synonymous expressions spoken by users. Harris's VUI design textbook [45] suggests that "judiciously maxing synonyms and metonyms in the vocabulary" allows more flexibility to the user.

4.3.4 Collaboratively Repair Conversation Breakdowns. During verbal communication, we often encounter small conversation breakdowns when people do not respond in a timely way or do not understand what each other said. Four participants mentioned that a natural VUI should solve these kinds of conversation breakdowns in a similar way to how humans collaboratively solve them by asking each other. It seems that the strategies of our designers largely echo conversation repair and error recovery strategies suggested by the existing VUI design textbooks [29, 70].

Challenges Specific to Designing for Social Conversations	
1	Augmenting Task-Oriented VUIs with Social Conversations is Difficult to Balance
2	Synthesized Voice Lacks Expressivity
3	SSML is Time-Consuming to Use While Not Producing the Desired Results
4	Difficult to Capture the Users' Emotions
Common Challenges	
5	Existing VUI Guidelines Lack Concrete and Useful Recommendations on How to Design for Naturalness
6	Writing for Spoken Language Is Difficult
7	Handling Various User Inputs and Conversational Context is Difficult

Table 2: 7 Primary Challenges in Designing Natural VUIs

5 DESIGNERS EXPERIENCE CHALLENGES

We asked our designers what was most challenging about designing for a natural VUI. In response, they described their challenges in the context of design practices for attaining the specific naturalness characteristics presented in Section 4. This enabled us to map their challenges to the categories of naturalness characteristics (see Table 2). In the end, there are challenges that are common to designing any type of conversation and ones that relate more specifically to *social* conversations. We note that challenges unique to transactional conversations were not prevalent.

5.1 Challenges Specific to Designing VUIs for Social Conversations

Our designers reported frequent struggles with imbuing VUIs with sympathetic and humane dialogues. These challenges pertained to design practices for enhancing VUI's naturalness characteristics in social conversations (Section 4.1). We illustrate the four prevailing challenges as follows.

5.1.1 Augmenting Task-Oriented VUIs with Social Conversations is Difficult to Balance. Five designers wanted to add characteristics of social conversations, such as expressing sympathy and maintaining an intriguing persona, to their task-oriented VUI applications. However, in an attempt to do so, they found that the dialogue gets longer and it conflicts with the overarching goal of the task-oriented applications to complete the tasks efficiently: *"So obviously I wanted to write the dialogues that felt* [like a] *human* [and] *didn't feel robotic, but I soon realized that things are more complicated. The more you want to add personality to things, then the longer becomes your dialogue."* (P20)

Sets of desired characteristics between task-oriented applications and social conversations tend to conflict: efficiency and simplicity for getting things done vs. friendly personality, interactivity, and familiarity for a human-like presence: *"...efficient, but it has to come up as like friendly* [and] *conversational."* (P5) *"Challenges are keeping it simple, yet interactive. It should sound familiar. It should sound friendly. It should not go out of the voice, so like that."* (P17)

This dilemma put our designers in a quandary and often made them give up incorporating social characteristics to their VUIs:

"But again we're still thinking 'Should we actually put in those little sentences [for having social interactions] *or not?'"* (P6) *"I would prefer, right now, to focus more on helping people achieve their goals and move on with their lives more than a kind of having these artificial entities talking to me in slang."* (P20)

5.1.2 Synthesized Voice Lacks Expressivity. To make a VUI *sound* natural in a social conversation, the designers wanted to have control over the way the speech synthesizer would narrate their dialogues to the user. However, 9 participants reported that current speech synthesis technology lacks the *expressivity* to interpret the intended meaning of the dialogue text and convey it to the user via rich paralanguage. They felt that even the best speech synthesizer still sounds like *"just a robo-voice"* (P4) or like *"just putting the sounds together"* (P18) rather than *"really meaning it* [the script]." (P18) They think that the voice synthesis technology has a large gap to bridge, saying *"there's a long way to go for it to become very expressive."* (P7)

They reported two specific cases where limited expressivity of the synthesized voices demotes naturalness of social dialogues. (1) Conveying emotion: Our designers found that the currently available synthesized voices were not good enough to express nuanced emotions that they desired to express for his storytelling application: *"...there are some subtleties that I couldn't get Alexa to feel nostalgic about, you know, there is no command like nostalgia about the house party that you first met this guy that you are still in love with at, you know?"* (P1) (2) Injecting non-lexical words: P8 reported that synthesized voices do not produce proper tones for non-lexical words (i.e., words do not have a defined meaning) such as laughter. Her design intention was to make her robot laugh with a happy tone, but it had a sarcastic tone instead: *"The robot can not laugh, because if the robot laughs, and it just says, 'Ha-ha-ha', it sounds sarcastic."* (P8) The other cases of mispronunciation include putting proper breaks in long sentences and pronouncing contractions, proper nouns (e.g., names of products and people), and interrogative sentences ending with a question mark.

Hiring voice actors who can narrate the script in a natural tone and flow of a "real voice" was reported by many participants as a common solution to make a VUI sound natural. However, recorded audio is considered to be significantly limited in flexibility when there is a need to change the narration and in scalability when handling a wide variety of data and conversation context: *"...if we discover during research there are more words, then we have to hire that actor again to speak those words again. So it was not practical at all."* (P8)

5.1.3 SSML is Time-Consuming to Use While Not Producing the Desired Results. While SSML is intended to address the expressivity challenge (see Section 5.1.2), it is largely failing to do so. Nine of our designers pointed out that writing and editing SSML tags is "time digging" (P13) and that using SSML frequently fails to yield the desired result as it sounds *"still too mechanical"* (P5) and *"it* [SSML] *doesn't come close to what it would be if you use a voice actor."* (P7) Due to these limits of the current SSML, most participants had abandoned using SSML except for making simple changes such as slowing the speeds, inserting breaks, and correcting mispronunciations.

At the heart of SSML's problem is its reductionist approach where it only offers control of each prosodic element at a time *separately* while it is the *holistic* experience of a sentence-level flow that modifies the meaning of the sentence and conveys nuanced emotion. Our designers found it difficult to make the whole sentence flow naturally, even after fine-tuning speech timings and prosody features by meticulously editing SSML tags: *"I think it's not very natural, like another 0.5-second break here, another somewhat slower here, all those things."* (P15) Designing for rich and expressive non-verbal prosody requires holistic control of all prosody features at the same time.

Most of our designers were using a simple text editor or generic XML mark-up tools for writing SSML tags. The lack of a quick and lightweight validation feature in these tools was another source of frustration. They had to (re)write and (re)listen to the whole sentence or paragraph even when only making a small change to their dialogues: *"Let's say you listen to a prompt, you decided that you wanted to change one thing by using SSML. You change that thing. You listen to it again. [...] right there you just spent [...] a couple minutes maybe, and if you have a hundred prompts to do, it's just not worth it for the small benefit you'll get."* (P9) Also, it was hard to evaluate when the SSML tag reached the optimal level of expressiveness. Our designers often spent a lot of time iteratively modifying SSML tags without knowing when to stop: *"Hard to stop, like, I'm not satisfied with what I got there, so I just keep on changing something here and there."* (P15)

The designers reported that different VUI platforms (e.g., Google Home and Amazon Alexa) can interpret the same SSML tags differently, hence the resulted voices may sound different. This requires designers to test their SSML tags for each platform, which takes a lot of time: *"Different speech synthesizers are going to have different packages, so I want to be able to play with the SSML before I decide on how this is going to work."* (P10)

5.1.4 Difficult to Capture the Users' Emotions. Our designers found it challenging to write VUI dialogues that are sympathetic to VUI users' emotions due to the lack of a way to capture emotions of VUI users and incorporate the detected emotions into their VUI dialogue designs. Our participants wished for a VUI design tool where they can write VUI conversation flows depending on the detected emotions of the user: *"I think it would be good to identify emotions... More useful* [thing] *would be to detect emotional content on an utterance* [from the user] *to give you the context."* (P2)

As a stopgap solution, our designers embraced the emotion-agnostic strategy that avoids making their VAs sound too excited or happy in case the user is experiencing negative feelings: *"You have to control the tone of voice because you can't sound very enthusiastic, things like that... because you never know the situation of the person on the other side. You don't know if the person is really emotionally ill or something more serious is happening at the time that the person is calling and interacting with the system."* (P3)

5.2 Common Challenges in Designing Natural VUIs

We identified three major sources of designers' struggles that can be commonly applied when designing for naturalness in both social and transactional conversations.

5.2.1 Existing VUI Guidelines Lack Concrete and Useful Recommendations on How to Design for Naturalness. Six designers mentioned three types of problems in applying the existing VUI guidelines to their design process.

First, our designers found that the existing design guidelines do not apply to certain VUIs depending on the context of the project: *"At the same time, I think every company will have its own set of these* [design guidelines] *...I mean, some apps are made to comfort people and make them feel less alone, and those* [generic] *guidelines are completely irrelevant, so it does depend on the context."* (P5)

Second, they found some existing VUI guidelines easy to dismiss as cliche and easy to let go: *"somewhat common sense in terms of avoiding using technical language, try making it casual and simple. [...] I feel like it's kind of obvious and you know that when you're creating something like a voice skill* [application for Amazon Alexa]... *I probably read* [the design guideline] *once, and I just left it."* (P6)

Third, some guidelines were useful, but validating the design with respect to them was effortful: *"So for example, I need to work on confirmations. Let me go to refresh my memory on how to do confirmation style... I don't have to like constantly go back to them* [the design guidelines], *but I certainly do go back in and look* [at them]." (P9)

5.2.2 Writing for Spoken Language Is Difficult. Existing guidelines recommend that designers write VUI dialogues in spoken form rather than written form [10, 29, 41]. However, for our designers, writing scripts for spoken language was a non-trivial challenge. Our designers reported that they write dialogues by typing on the keyboard instead of speaking them first, and it is often hard to detect unnaturalness of the dialogues just by reading: *"A lot of times, the conversation sounds good on paper, but you really have to just say it."* (P12) P18 offered an illustrative example; even though several guidelines ask designers to avoid putting too much information in one line [42], they often make the mistake as it seems fine when they just *read* the script.

When conversing with others, people *subconsciously* use the features of spoken language, such as filler words, colloquial words, and personal pronouns [19, 42, 92, 115]. Our designers mentioned that the unnaturalness of written scripts is hard to detect due to its subconscious nature, and many designers often treat this problem as something insignificant and hence do not put the effort in enhancing it: *"people* [VUI designers] *feel like because they can talk, because they speak English, so they can write one of these interfaces."* (P10)

5.2.3 Handling Varied User Inputs and Conversational Context is Difficult. Four of our designers acknowledged that VUIs afford user inputs with considerable flexibility. However, conversation breakdowns from unexpected inputs can jeopardize the naturalness of dialogue. Hence, the designers reported that it is difficult to expect and prepare for all possible conversational scenarios that can occur during user interaction (e.g., "[Users] *say something completely different from what I expected."* (P2) Part of this problem stemmed from limitations of the current natural language understanding (NLU) engine in comprehending every possible expression in our language.

Our designers reported using two strategies to prevent conversational breakdowns from unexpected inputs. First, they often collect synonymous expressions through fieldwork with potential end-users and train the NLU engine, but they usually found that the collected data do not cover all possible inputs. (P9) Second, they narrow down a set of available conversation pathways *in advance* by making the VA guide the user to talk about matters in the service domain only: *"We can't handle all those things. So you really need to know how to guide the conversation to get the person to know what they can say, and help them say it in a way that your technology can actually handle."* (P9) The range of spoken inputs can still be difficult to predict for certain user groups. P8, P13, and P16 mentioned that designing for children is particularly challenging due to the wide range of possible inputs they can generate: *"People can say anything. Children can say more than anything... they don't follow instructions, usually. So they can go randomly into anything."* (P16)

6 DISCUSSION

We reflect on our findings and their implications for understanding the meaning of naturalness in human-agent interactions, incorporating factors of the conversational context in the VUI design process, and developing better tools and guidelines in pursuit of naturalness in VUIs.

6.1 Naturalness is context dependent

The context-dependent characteristics of naturalness hinge primarily on the role that a VUI is expected to play. For the present, we discovered two primary roles, transactional and social, as the prevalent types that designers identified. However, given the pervasiveness of computing to people's everyday lives [18] and the near-ubiquity of VUI-enabled devices, it's entirely possible that the role of VUIs will be extended and diversified beyond these two types. For instance, a VUI in an interactive learning system can serve the role of an *instructor* [55], and not that of a task-oriented assistant nor a social companion. Designers will need to adjust their conception of naturalness in VUI to the new roles appropriately.

We categorized the characteristics of naturalness according to the role of VUIs, drawing heavily on Clark et al.'s categories of conversation types [27]. Another researcher may take a different but complementary approach by leveraging dimensions other than the role of VUIs, such as labeling each found characteristic with different qualities of natural VUI experiences, such as anthropomorphism, transparency, efficiency, pleasantness, etc.

Our study reveals that naturalness characteristics may not transfer across the different roles a VUI plays. For instance, answering a question with high accuracy and speed may not be conducive to the naturalness of a social conversation, but may be highly desirable and feel *natural* in a transactional one. As such, designers should not blindly follow core naturalness characteristics; they need to start with the role of the conversation and selectively incorporate related characteristics, and sometimes even demote a certain naturalness characteristic when it mismatches the target context.

The target user demographic is another factor to consider when designing for naturalness. For example, children tend to anthropomorphize VUIs [75] and the designers in our study expect their VUIs to offer children an entertaining experience. They focused on tailoring designs to the specific needs of the user. This aligns with Wigdor et al.'s conception of naturalness [114] that whether an

interaction is natural or not is contingent on the user's subjective experience.

6.2 There are inherent tradeoffs in designing for naturalness in transactional vs. social agents

We found that there are characteristics that are challenging to pursue together. Most of the designers in our study were experienced primarily in designing for task-oriented VUIs. When they try to enrich dialogues of those systems with social courtesy, such as sympathetic responses or small chit-chat, they face difficulty in finding an appropriate tradeoff between designing for an effective assistant vs. an affable companion; P20 adds that *"the more you want to add personality to things, the longer become your dialogue"*. This challenge uncovers a significant design dilemma facing designers. Fundamentally, such agents offer a transactional service that is structured in a social format only at the surface level [88]. However, they still have to reconcile the dual role of helpful assistant and pleasant social interlocutor.

6.3 Naturalness goes beyond behavioral realism

When asked what is natural in VUIs, designers tend to think of naturalness as a quality of the system that can provide a positive experience in the given application context. This concept differs from the existing notion of naturalness that equates it to behavioural realism (being like a human). Our results provide a conceptual departure from the notion of naturalness as an imitation of the qualities in human-to-human interactions. In their seminal work on mediated communication [49], Hollan and Stornetta claim that interactive media technologies should be designed to go "beyond being there", such that new tools offer uniquely beneficial interaction that humans in the flesh cannot offer, rather than simply mimicking face-to-face reality. Given that we found 'beyond-human' characteristics within transactional types, it is intriguing to contemplate what machine-specific characteristics might enhance natural *social* encounters. There might, for example, be aspects of non-human agents that are beneficial for offering consolation and emotional support.

6.4 Implications for VUI design tools

Our findings indicate that designers seeking naturalness want to control VUI characteristics in ways that are not typically accessible to a 'mere' user of the platforms (e.g., acquiring more sophisticated voice control than SSML allows). Hence, to promote naturalness, the platform providers must either: (1) develop and distribute design tools that help designers create natural VUIs or (2) give the designers direct access to control over VUI behaviours on their platforms. Delving into the designers' practices for naturalness in the design process revealed two significant critical gaps in the way existing VUI related technologies facilitate their jobs. In this section, we propose design implications for the platform providers to help them fill these gaps.

6.4.1 Towards More Natural Sounding Narration. Although fine-tuning the way VUIs *sound* was given great importance in providing natural user experience with VUIs, the VUI scene is lacking in design tools for producing narrations that sound rich and nuanced. Our designers regard SSML and voice talents to be the only possible two approaches available to them, but the pros and cons of the two were complementary: working with voice talents gives designers a great deal of control over para-language (i.e. non-lexical attributes of speech, such as intonation, timing, etc.), but hiring them is expensive and recorded audio clips are not flexible nor scalable. SSML lacks sufficient control, as detailed in Section 5.1.3. VUI designers need a solution that is both scalable and allows a great degree of control.

There is a lot to learn from the way VUI designers guide voice talents to narrate a given script in intended prosody and timing. Their directions are primarily demonstration-based, such as ghost narration. Hence, a demonstration-based prosody editing, leveraging our own voices [111], is a promising approach worth investigating. Also, our participants mentioned that they often use multi-modal cues, such as hand gestures, to convey and highlight intended changes to the voice talents. Similarly, graphical or direct manipulation of intonation and timing can enable faster and easier creation of natural-sounding narration.

6.4.2 Writing for More Natural Spoken Dialogues. One of the primary challenges for VUI designers is to write dialogues with spoken language characteristics, such as frequent use of filler words and fewer big words. In the field, there exist several dialogue design tools that offer many beneficial features like dialogue mapping and instant speech-based testing, but they do not recommend or validate linguistic properties for a dialogue script. Given that designers tend to dismiss this kind of naturalness characteristic easily, a proper scripting interface should warn the user, similarly to proof-reading tools, when the dialogue has too many traits of written language or should suggest alternative phrasing in spoken language. Also, such tools can offer editing suggestions that are tailored to the target users or the purpose of the application as the required naturalness varies by such design contexts.

Implementing such tools will require linguistic modelling of spoken vs. written language, computational prediction for evaluating the given script, and alternative searching for recommending different expressions. Natural language processing techniques are becoming increasingly sophisticated. For example, F-Score [47] is a linguistic measure of how formal a given text is and ConvoKit [25] can identify linguistic markers that are indicative of politeness.

7 CONCLUSION

To inform those who create tools and guidelines to support the VUI design process, we conducted 20 interviews with VUI designers with the key focus on how they strive to attain naturalness in their VUIs, how they integrate such goals into their design practice, and the challenges they face in doing so. Through a reflexive thematic analysis, we uncovered 12 characteristics of natural VUIs and introduced 3 categories for these characteristics: 'Social', 'Transactional,' and 'Core'. While many of the traits we found are human-like in essence, some designers mentioned that they saw naturalness in VUIs as a quality that is *beyond-human*—in their conception, such as exhibiting machine-like speed, accuracy, and memory. We also

uncovered 7 challenges that designers face in incorporating naturalness characteristics in VUIs. Most importantly, there is a design dilemma of attaining social characteristics in task-oriented applications, which our designers expect will become more acute as the role of modern VUIs expands. SSML, specifically designed to support the creation of expressive prosody, fails to do so. In addition, incorporating spoken language features into written VUI scripts is difficult. We end by providing implications for future tool support for VUI design.

8 LIMITATIONS & FUTURE WORK

As we didn't observe designers in their workplace, future studies should conduct direct observation or contextual inquiry of designers, which may bring a deeper understanding of richer context and social dynamics in creating natural VUIs. Also, our 12 characteristics show how designers perceive naturalness, but some of these may not be aligned with how end-users perceive naturalness. Therefore, future studies should investigate potential mismatches.

ACKNOWLEDGMENTS

This work was partially supported by the NSERC Discovery Grant and CREATE programs, as well as a KEIT ATC+ Grant. We appreciate the participants who helped us with this study, and thank MUX lab members for their valuable feedback.

REFERENCES

[1] Marshall D. Abrams, George E. Lindamood, and Thomas N. Pyke. 1973. Measuring and Modelling Man-Computer Interaction. In *Proceedings of the 1973 ACM SIGME Symposium (SIGME '73)*. Association for Computing Machinery, New York, NY, USA, 136–142. https://doi.org/10.1145/800268.809345

[2] Aaron Adler and Randall Davis. 2007. Speech and Sketching for Multimodal Design. In *ACM SIGGRAPH 2007 Courses* (San Diego, California) *(SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 14–es. https://doi.org/10.1145/1281500.1281525

[3] F Niyi Akinnaso. 1982. On the differences between spoken and written language. *Language and speech* 25, 2 (1982), 97–125.

[4] James Allen, Donna Byron, Myroslava Dzikovska, George Ferguson, Lucian Galescu, and Amanda Stent. 2000. An architecture for a generic dialogue shell. *Natural Language Engineering* 6, 3-4 (2000), 213–228.

[5] Amazon. 2018. *Things Every Alexa Skill Should Do: Pass the One-Breath Test.* Amazon. https://developer.amazon.com/blogs/alexa/post/531ffdd7-acf3-43ca-9831-9c375b08afe0/things-every-alexa-skill-should-do-pass-the-one-breath-test

[6] Amazon. 2020. *Conversational AI.* Amazon. https://developer.amazon.com/en-US/alexa/alexa-skills-kit/conversational-ai

[7] Amazon. 2020. Design Process: The process of thinking through the design of a voice experience. https://developer.amazon.com/fr/designing-for-voice/design-process/ Accessed: 2020-01-31.

[8] Amazon. 2020. *Supported SSML Tags.* Amazon. https://docs.aws.amazon.com/polly/latest/dg/supportedtags.html

[9] Amazon. 2020. *Voice Design for Alexa Experiences: Be Adaptable.* Amazon. https://developer.amazon.com/en-US/docs/alexa/alexa-design/adaptable.html

[10] Amazon. 2020. *Write Out a Script with Conversational Turns.* Amazon. https://developer.amazon.com/en-US/docs/alexa/alexa-design/script.html

[11] AmazonAlexa. 2019. Amazon Alexa Live Design Track - Free Online Conference for Voice Developers. https://www.twitch.tv/videos/418170087?t=0h50m26s

[12] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How People (Really) Use Voice Assistants. *ACM Trans. Comput.-Hum. Interact.* 26, 3, Article 17 (April 2019), 28 pages. https://doi.org/10.1145/3311956

[13] T Bennett. 1977. Verb voice in Unplanned and Planned narratives. *Keenan, EO yT. Bennett (ed.): Discourse across time and space. SCOPIL* 5 (1977).

[14] Gregory S Berns. 2004. Something funny happened to reward. *Trends in cognitive sciences* 8, 5 (2004), 193–194.

[15] Li Bian and Henry Holtzman. 2011. Qooqle: Search with Speech, Gesture, and Social Media. In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 541–542. https://doi.org/10.1145/2030112.2030203

[16] Timothy Bickmore and Justine Cassell. 2001. Relational Agents: A Model and Implementation of Building User Trust. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '01)*. Association for Computing Machinery, New York, NY, USA, 396–403. https://doi.org/10.1145/365024.365304

[17] Timothy W. Bickmore and Rosalind W. Picard. 2005. Establishing and Maintaining Long-Term Human-Computer Relationships. *ACM Trans. Comput.-Hum. Interact.* 12, 2 (June 2005), 293–327. https://doi.org/10.1145/1067860.1067867

[18] Susanne Bødker. 2006. When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles*. Association for Computing Machinery, New York, NY, USA, 1–8.

[19] Gladys Borchers. 1936. An approach to the problem of oral style. *Quarterly Journal of Speech* 22, 1 (1936), 114–117.

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. https://doi.org/10.1191/1478088706qp063oa arXiv:https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp063oa

[21] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2010. *Thematic Analysis.* Springer Singapore, Singapore, Chapter 48, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103

[22] Patricia Braunger and Wolfgang Maier. 2017. Natural Language Input for In-Car Spoken Dialog Systems: How Natural is Natural?. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*. Association for Computational Linguistics, Pennsylvania, PA, USA, 137–146.

[23] Gillian Brown, Gillian D Brown, Gillian R Brown, Brown Gillian, and George Yule. 1983. *Discourse analysis.* Cambridge university press, Cambridge, UK.

[24] J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in Conversational Interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Pittsburgh, Pennsylvania, USA) *(CHI '99)*. Association for Computing Machinery, New York, NY, USA, 520–527. https://doi.org/10.1145/302979.303150

[25] Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. ConvoKit: A Toolkit for the Analysis of Conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, 57–60. https://www.aclweb.org/anthology/2020.sigdial-1.8

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

[27] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, et al. 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*. Association for Computing Machinery, New York, NY, USA, 1–12.

[28] Victoria Clarke and Virginia Braun. 2014. *Thematic Analysis.* Springer, New York, NY, 1947–1952. https://doi.org/10.1007/978-1-4614-5583-7_311

[29] Michael H Cohen, Michael Harris Cohen, James P Giangola, and Jennifer Balogh. 2004. *Voice user interface design.* Addison-Wesley Professional, Boston, USA.

[30] Ron Cole, Dominic W Massaro, Jacques de Villiers, Brian Rundle, Khaldoun Shobaki, Johan Wouters, Michael Cohen, Jonas Baskow, Patrick Stone, Pamela Connors, et al. 1999. New tools for interactive speech and language training: using animated conversational agents in the classroom of profoundly deaf children. In *MATISSE-ESCA/SOCRATES Workshop on Method and Tool Innovations for Speech Science Education*. ISCA, BAIXAS, FRANCE.

[31] Lucie Daubigney, Matthieu Geist, Senthilkumar Chandramohan, and Olivier Pietquin. 2012. A comprehensive reinforcement learning framework for dialogue management optimization. *IEEE Journal of Selected Topics in Signal Processing* 6, 8 (2012), 891–902.

[32] Ken H Davis, R Biddulph, and Stephen Balashek. 1952. Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America* 24, 6 (1952), 637–642.

[33] 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*. Association for Computing Machinery, New York, NY, USA, 1–12.

[34] Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study. *Acta Psychologica* 20 (1962), 78–100.

[35] Suzanne Eggins and Diana Slade. 2005. *Analysing casual conversation.* Equinox Publishing Ltd., Sheffield, United Kingdom.

[36] Yuan-Yi Fan, Soyoung Shin, and Vids Samanta. 2017. Contour: An Efficient Voice-enabled Workflow for Producing Text-to-Speech Content. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 133–135.

[37] Piedad Garrido, F Martinez, and Christian Guetl. 2010. Adding semantic web knowledge to intelligent personal assistant agents. In *Proceedings of the ISWC*

2010 Workshops. Philippe Cudre-Mauroux, Cambridge, MA, USA, 1–12.

[38] Emer Gilmartin, Francesca Bonin, Loredana Cerrato, Carl Vogel, and Nick Campbell. 2015. What's the game and who's got the ball? genre in spoken interaction. In *2015 AAAI Spring symposium series*. AAAI, Menlo Park, CA, US.

[39] Przemysław Głomb, Michał Romaszewski, Sebastian Opozda, and Arkadiusz Sochan. 2012. Choosing and Modeling the Hand Gesture Database for a Natural User Interface. In *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*, Eleni Efthimiou, Georgios Kouroupetroglou, and Stavroula-Evita Fotinea (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 24–35.

[40] Google. 2020. *Conversation design*. Google. https://developers.google.com/assistant/actions/design

[41] Google. 2020. *Conversation Design Guideline*. Google. https://designguidelines.withgoogle.com/conversation/conversation-design/welcome.html

[42] Google. 2020. *Conversational components–Informational statements*. Google. https://designguidelines.withgoogle.com/conversation/conversational-components/informational-statements.html

[43] David Griol, Javier Carbó, and José M Molina. 2013. An automatic dialog simulation technique to develop and evaluate interactive conversational agents. *Applied Artificial Intelligence* 27, 9 (2013), 759–780.

[44] Lone Koefoed Hansen and Peter Dalsgaard. 2015. Note to Self: Stop Calling Interfaces "Natural". In *Proceedings of The Fifth Decennial Aarhus Conference on Critical Alternatives* (Aarhus, Denmark) *(CA '15)*. Aarhus University Press, Aarhus N, 65–68. https://doi.org/10.7146/aahcc.v1i1.21316

[45] Randy Allen Harris. 2004. *Voice interaction design: crafting the new conversational speech systems*. Elsevier, Amsterdam, Netherlands.

[46] Thomas Kevin Harris and Roni Rosenfeld. 2004. A universal speech interface for appliances. In *Eighth International Conference on Spoken Language Processing*. ISCA, BAIXAS, FRANCE.

[47] Francis Heylighen and Jean-Marc Dewaele. 2002. Variation in the contextuality of language: An empirical measure. *Foundations of science* 7, 3 (2002), 293–340.

[48] Katri Hiovain, Päivi Kristiina Jokinen, et al. 2016. Acoustic features of different types of laughter in North Sami conversational speech. In *Proceedings of the LREC Workshop Just talking–casual talk among humans and machines, Portorož, Slovenia, 2016*. European Language Resources Association (ELRA), France.

[49] Jim Hollan and Scott Stornetta. 1992. Beyond being there. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 119–125.

[50] Terence KL Hui and R Simon Sherratt. 2017. Towards disappearing user interfaces for ubiquitous computing: human enhancement from sixth sense to super senses. *Journal of Ambient Intelligence and Humanized Computing* 8, 3 (2017), 449–465.

[51] IBM. 2019. *Expressive SSML*. IBM. https://cloud.ibm.com/docs/text-to-speech?topic=text-to-speech-expressive

[52] Apple Inc. 2020. Siri Style Guide - SiriKit. https://developer.apple.com/siri/style-guide/

[53] SoundHound Inc. 2020. *Natural Dialogue and Conversation with Voice UI*. SoundHound Inc. https://www.soundhound.com/vui-guide/enabling-natural-conversations-in-vui

[54] INTUITY. 1996. *INTUITY™ CONVERSANT® System Version 6.0*. Lucent Technologies, Bell Labs Innovations Bell Labs Innovations, NJ, USA.

[55] Hyunhoon Jung, Hee Jae Kim, Seongeun So, Jinjoong Kim, and Changhoon Oh. 2019. TurtleTalk: an educational programming game for children with voice user interface. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–6.

[56] Scott R. Klemmer, Anoop K. Sinha, Jack Chen, James A. Landay, Nadeem Aboobaker, and Annie Wang. 2000. Suede: A Wizard of Oz Prototyping Tool for Speech User Interfaces. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology* (San Diego, California, USA) *(UIST '00)*. Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/354401.354406

[57] A. Kumar, F. Metze, and M. Kam. 2014. Enabling the Rapid Development and Adoption of Speech-User Interfaces. *Computer* 47, 1 (2014), 40–47.

[58] Unseok Lee and Jiro Tanaka. 2013. Finger Identification and Hand Gesture Recognition Techniques for Natural User Interface. In *Proceedings of the 11th Asia Pacific Conference on Computer Human Interaction* (Bangalore, India) *(APCHI '13)*. Association for Computing Machinery, New York, NY, USA, 274–279. https://doi.org/10.1145/2525194.2525296

[59] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 801–809.

[60] Vice Media LLC. 2020. *Google Duplex Puts AI Into a Social Uncanny Valley*. Vice Media LLC. https://www.vice.com/en_us/article/d3kgkk/google-duplex-assistant-voice-call-dystopia

[61] B. Loureiro and R. Rodrigues. 2011. Multi-touch as a Natural User Interface for elders: A survey. In *6th Iberian Conference on Information Systems and Technologies (CISTI 2011)*. IEEE, New Jersey, NJ, United States, 1–6.

[62] Bruce T Lowerre. 1976. *The HARPY speech recognition system*. Technical Report. CARNEGIE-MELLON UNIV PITTSBURGH PA DEPT OF COMPUTER SCIENCE.

[63] Ewa Luger and Abigail Sellen. 2016. "Like Having a Really Bad PA": The Gulf between User Expectation and Experience of Conversational Agents. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI '16)*. Association for Computing Machinery, New York, NY, USA, 5286–5297. https://doi.org/10.1145/2858036.2858288

[64] John Lyons. 1977. Semantics Cambridge University Press. *Cambridge, UK* 1 (1977).

[65] Speech Technology Magazine. 2018. Q&A: Designing Natural Language Home Assistants Using Amazon Alexa. https://www.speechtechmag.com/Articles/ReadArticle.aspx?ArticleID=123486

[66] Alessio Malizia and Andrea Bellucci. 2012. The Artificiality of Natural User Interfaces. *Commun. ACM* 55, 3 (March 2012), 36–38. https://doi.org/10.1145/2093548.2093563

[67] Michael L Mauldin. 1994. Chatterbots, tinymuds, and the turing test: Entering the loebner prize competition. In *AAAI*, Vol. 94. AAAI, Menlo Park, CA, US, 16–21.

[68] Michael F McTear. 2016. The rise of the conversational interface: A new kid on the block?. In *International Workshop on Future and Emerging Trends in Language Technology*. Springer, New York, NY, USA, 38–49.

[69] Michael F. McTear. 2017. The Rise of the Conversational Interface: A New Kid on the Block?. In *Future and Emerging Trends in Language Technology. Machine Learning and Big Data*, José F Quesada, Francisco-Jesús Martín Mateos, and Teresa López Soto (Eds.). Springer International Publishing, Cham, 38–49.

[70] Robert J Moore and Raphael Arar. 2019. *Conversational UX Design: A Practitioner's Guide to the Natural Conversation Framework*. Morgan & Claypool, CA, USA.

[71] Christine Murad and Cosmin Munteanu. 2019. " I don't know what you're talking about, HALexa" the case for voice user interface guidelines. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. Association for Computing Machinery, New York, NY, USA, 1–3.

[72] Christine Murad and Cosmin Munteanu. 2020. Designing Voice Interfaces: Back to the (Curriculum) Basics. 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, 1–12. https://doi.org/10.1145/3313831.3376522

[73] Christine Murad, Cosmin Munteanu, Leigh Clark, and Benjamin R Cowan. 2018. Design guidelines for hands-free speech interaction. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*. Association for Computing Machinery, New York, NY, USA, 269–276.

[74] Chelsea M. Myers, David Grethlein, Anushay Furqan, Santiago Ontañón, and Jichen Zhu. 2019. Modeling Behavior Patterns with an Unfamiliar Voice User Interface. In *Proceedings of the 27th ACM Conference on User Modeling, Adaptation and Personalization* (Larnaca, Cyprus) *(UMAP '19)*. Association for Computing Machinery, New York, NY, USA, 196–200. https://doi.org/10.1145/3320435.3320475

[75] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56, 1 (2000), 81–103. https://doi.org/10.1111/0022-4537.00153 arXiv:https://spssi.onlinelibrary.wiley.com/doi/pdf/10.1111/0022-4537.00153

[76] Clifford Nass, Youngme Moon, B. J. Fogg, Byron Reeves, and Chris Dryer. 1995. Can Computer Personalities Be Human Personalities?. In *Conference Companion on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. Association for Computing Machinery, New York, NY, USA, 228–229. https://doi.org/10.1145/223355.223538

[77] Clifford Nass, Youngme Moon, B. J. Fogg, Byron Reeves, and Chris Dryer. 1995. Can Computer Personalities Be Human Personalities?. In *Conference Companion on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. Association for Computing Machinery, New York, NY, USA, 228–229. https://doi.org/10.1145/223355.223538

[78] Clifford Nass, Youngme Moon, and Nancy Green. 1997. Are Machines Gender Neutral? Gender-Stereotypic Responses to Computers With Voices. *Journal of Applied Social Psychology* 27, 10 (1997), 864–876. https://doi.org/10.1111/j.1559-1816.1997.tb00275.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.1997.tb00275.x

[79] Clifford Ivar Nass and Scott Brave. 2005. *Wired for speech: How voice activates and advances the human-computer relationship*. MIT press Cambridge, MA, USA.

[80] R. S. Nickerson. 1976. On Conversational Interaction with Computers. In *Proceedings of the ACM/SIGGRAPH Workshop on User-Oriented Design of Interactive Graphics Systems* (Pittsburgh, PA) *(UODIGS '76)*. Association for Computing Machinery, New York, NY, USA, 101–113. https://doi.org/10.1145/1024273.1024286

[81] Donald A. Norman. 2010. Natural User Interfaces Are Not Natural. *Interactions* 17, 3 (May 2010), 6–10. https://doi.org/10.1145/1744161.1744163

[82] Ulin Nuha. 2014. Transactional and Interpersonal Conversation Texts in English Textbook. *Register Journal* 7, 2 (2014), 205–224.

[83] Roy C O'Donnell. 1974. Syntactic differences between speech and writing. *American Speech* 49, 1/2 (1974), 102–110.

[84] Shigehiro Oishi, Selin Kesebir, Casey Eggleston, and Felicity F Miao. 2014. A hedonic story has a transmission advantage over a eudaimonic story. *Journal of Experimental Psychology: General* 143, 6 (2014), 2153.

[85] Ioannis Papaioannou, Christian Dondrup, Jekaterina Novikova, and Oliver Lemon. 2017. Hybrid chat and task dialogue for more engaging hri using reinforcement learning. In *2017 26th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, New Jersey, NJ, United States, 593–598.

[86] Cathy Pearl. 2016. *Designing voice user interfaces: principles of conversational experiences*. "O'Reilly Media, Inc.", Massachusetts, USA.

[87] Diana Pérez-Marín and Ismael Pascual-Nieto. 2013. An exploratory study on how children interact with pedagogic conversational agents. *Behaviour & Information Technology* 32, 9 (2013), 955–964.

[88] 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* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 640, 12 pages. https://doi.org/10.1145/3173574.3174214

[89] Aung Pyae and Tapani N. Joelsson. 2018. Investigating the Usability and User Experiences of Voice User Interface: A Case of Google Home Smart Speaker. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct* (Barcelona, Spain) *(MobileHCI '18)*. Association for Computing Machinery, New York, NY, USA, 127–131. https://doi.org/10.1145/3236112.3236130

[90] Silvia Quarteroni. 2018. Natural Language Processing for Industrial Applications. *Spektrum* 41 (2018), 105.

[91] W3C Recommendation. 2010. *SSML stardard set up by W3C Recommendation*. W3C Recommendation. https://www.w3.org/TR/speech-synthesis11/#:~:text=SSML%20is%20part%20of%20a,in%20Web%20and%20other%20applications.

[92] Gisela Redeker. 1984. On differences between spoken and written language. *Discourse processes* 7, 1 (1984), 43–55.

[93] Byron Reeves and Clifford Ivar Nass. 1996. *The media equation: How people treat computers, television, and new media like real people and places*. Cambridge university press, Cambridge,UK.

[94] Jack C Richards, Jack C Richards, et al. 1990. *The language teaching matrix*. Cambridge University Press, Cambridge,UK.

[95] Steven Ross, Elizabeth Brownholtz, and Robert Armes. 2004. Voice User Interface Principles for a Conversational Agent. In *Proceedings of the 9th International Conference on Intelligent User Interfaces* (Funchal, Madeira, Portugal) *(IUI '04)*. Association for Computing Machinery, New York, NY, USA, 364–365. https://doi.org/10.1145/964442.964536

[96] James L Ryan, Richard L Crandall, and Marion C Medwedeff. 1966. A conversational system for incremental compilation and execution in a time-sharing environment. In *Proceedings of the November 7-10, 1966, fall joint computer conference*. Association for Computing Machinery, New York, NY, USA, 1–21.

[97] Tracy Sanders, Kristin E Oleson, Deborah R Billings, Jessie YC Chen, and Peter A Hancock. 2011. A model of human-robot trust: Theoretical model development. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 55. SAGE Publications, Los Angeles, CA, 1432–1436.

[98] Sergio Sayago, Barbara Barbosa Neves, and Benjamin R Cowan. 2019. Voice Assistants and Older People: Some Open Issues. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 7, 3 pages. https://doi.org/10.1145/3342775.3342803

[99] Klaus P Schneider. 1988. *Small talk: Analyzing phatic discourse*. Vol. 1. Hitzeroth, New York, NY, USA.

[100] Tanja Schultz, Alan W Black, Sameer Badaskar, Matthew Hornyak, and John Kominek. 2007. Spice: Web-based tools for rapid language adaptation in speech processing systems. In *Eighth Annual Conference of the International Speech Communication Association*. ISCA, BAIXAS, FRANCE.

[101] Sanni Siltanen and Jouko Hyväkkä. 2006. Implementing a Natural User Interface for Camera Phones Using Visual Tags. In *Proceedings of the 7th Australasian User Interface Conference - Volume 50* (Hobart, Australia) *(AUIC '06)*. Australian Computer Society, Inc., AUS, 113–116.

[102] Gabriel Skantze. 2007. *Error handling in spoken dialogue systems-managing uncertainty, grounding and miscommunication*. Gabriel Skantze, Stockholm, Sweden.

[103] Alessandro Soro, Samuel Aldo Iacolina, Riccardo Scateni, and Selene Uras. 2011. Evaluation of User Gestures in Multi-Touch Interaction: A Case Study in Pair-Programming. In *Proceedings of the 13th International Conference on Multimodal Interfaces* (Alicante, Spain) *(ICMI '11)*. Association for Computing Machinery, New York, NY, USA, 161–168. https://doi.org/10.1145/2070481.2070508

[104] Pei-Hao Su, Milica Gasic, Nikola Mrksic, Lina Rojas-Barahona, Stefan Ultes, David Vandyke, Tsung-Hsien Wen, and Steve Young. 2016. On-line Active Reward Learning for Policy Optimisation in Spoken Dialogue Systems. arXiv:arXiv:1605.07669

[105] Bernhard Suhm. 2003. Towards best practices for speech user interface design. In *Eighth European Conference on Speech Communication and Technology*. ISCA, BAIXAS, FRANCE.

[106] Deborah Tannen. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58, 1 (1982), 1–21.

[107] Alan M Turing. 2009. Computing machinery and intelligence. In *Parsing the Turing Test*. Springer, New York, NY, USA, 23–65.

[108] Andries Van Dam. 2001. User interfaces: disappearing, dissolving, and evolving. *Commun. ACM* 44, 3 (2001), 50–52.

[109] Laura Pfeifer Vardoulakis, Lazlo Ring, Barbara Barry, Candace L. Sidner, and Timothy Bickmore. 2012. Designing Relational Agents as Long Term Social Companions for Older Adults. In *Intelligent Virtual Agents*, Yukiko Nakano, Michael Neff, Ana Paiva, and Marilyn Walker (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 289–302.

[110] Ning Wang, David V Pynadath, Susan G Hill, and Aberdeen Proving Ground. 2015. Building trust in a human-robot team with automatically generated explanations. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC)*, Vol. 15315. National Training and Simulation Association, Orlando, Florida, 1–12.

[111] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. arXiv:arXiv:1703.10135

[112] Zhuxiaona Wei and James A Landay. 2018. Evaluating speech-based smart devices using new usability heuristics. *IEEE Pervasive Computing* 17, 2 (2018), 84–96.

[113] Mark Weiser. 1999. The Computer for the 21st Century. *SIGMOBILE Mob. Comput. Commun. Rev.* 3, 3 (July 1999), 3–11. https://doi.org/10.1145/329124.329126

[114] Daniel Wigdor and Dennis Wixon. 2011. *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[115] Charles H Woolbert. 1922. Speaking and writing—A study of differences. *Quarterly Journal of Speech* 8, 3 (1922), 271–285.

[116] Rui Yan. 2018. "Chitty-Chitty-Chat Bot": Deep Learning for Conversational AI.. In *IJCAI*, Vol. 18. IJCAI, California, CA, USA, 5520–5526.

[117] Nicole Yankelovich, Gina-Anne Levow, and Matt Marx. 1995. Designing SpeechActs: Issues in speech user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. Association for Computing Machinery, New York, NY, USA, 369–376.

[118] Steve Young, Milica Gašić, Blaise Thomson, and Jason D Williams. 2013. Pomdp-based statistical spoken dialog systems: A review. *Proc. IEEE* 101, 5 (2013), 1160–1179.

[119] Zhou Yu, Leah Nicolich-Henkin, Alan W Black, and Alexander Rudnicky. 2016. A wizard-of-oz study on a non-task-oriented dialog systems that reacts to user engagement. In *Proceedings of the 17th annual meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Pennsylvania, PA, USA, 55–63.

[120] J. K. Zao, T. T. Gan, C. K. You, S. J. R. Méndez, C. E. Chung, Y. T. Wang, T. Mullen, and T. P. Jung. 2014. Augmented Brain Computer Interaction Based on Fog Computing and Linked Data. In *2014 International Conference on Intelligent Environments*. IEEE, New Jersey, NJ, United States, 374–377. https://doi.org/10.1109/IE.2014.54