What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants

Jaewook Lee
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
jaewook4@illinois.edu

Sebastian S. Rodriguez
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
srodri44@illinois.edu

Raahul Natarrajan
Vanderbilt University
Nashville, Tennessee, USA
raahul.natarrajan@vanderbilt.edu

Jacqueline Chen
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
jc27@illinois.edu

Harsh Deep
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
hdeep2@illinois.edu

Alex Kirlik
University of Illinois at Urbana-Champaign
Urbana, Illinois, USA
kirlik@illinois.edu

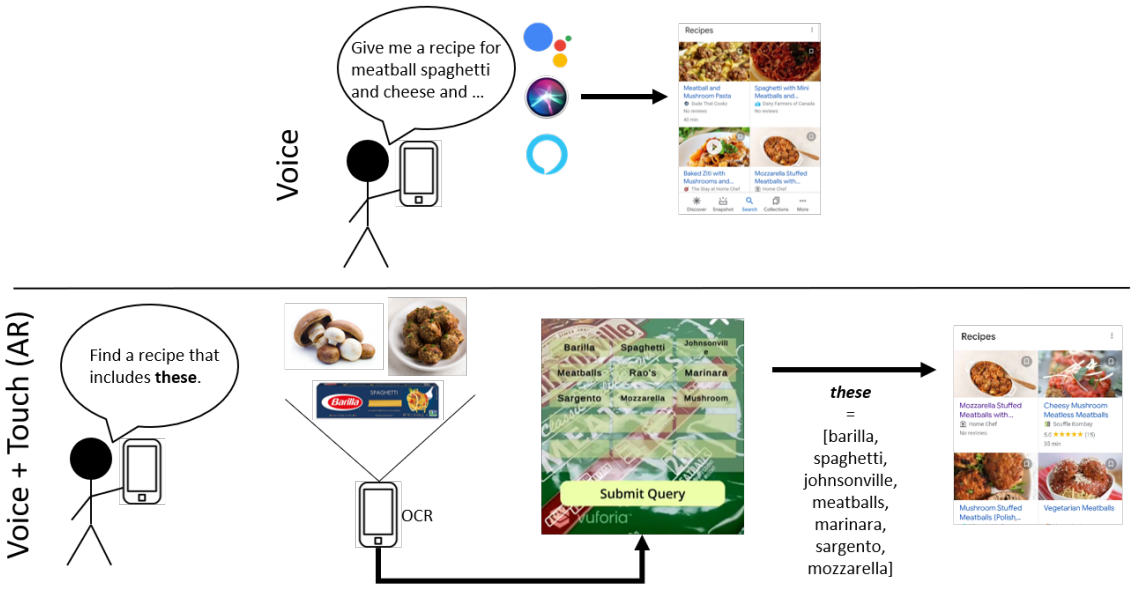

Figure 1: Interaction flow of a standard Voice Assistant and our proposed multimodal (voice + touch) Voice Assistant.

ABSTRACT

Human speech often contains ambiguity stemming from the use of demonstrative pronouns (DPs), such as "this" and "these." While we can typically decipher which objects of interest DPs are referring to based on context, modern day voice assistants (VAs – such as Google Assistant and Siri) are yet unable to process queries containing such ambiguity. For instance, to humans, a question such as "how much is this?" can be clarified through visual reference (e.g., a buyer gestures to the seller the object they would like to purchase). To bridge this gap between human and machine cognition, we built and examined a touch + voice multimodal VA prototype that enables users to select key spatial information to embed as context and query the VA. The prototype converts results of mobile, real-time object recognition and optical character recognition models into augmented reality buttons that represent features. Users can interact with and modify the selected features through a word grid. We conducted a study to investigate: 1) how touch performs as an additional modality to resolve ambiguity in queries, 2) how users use DPs when interacting with VAs, and 3) how users perceive a VA that can understand DPs. From this procedure we found that as the query becomes more complex, users prefer the multimodal VA over the standard VA without experiencing elevated cognitive load. Additionally, even though it took some time getting used to, many participants eventually became comfortable with using DPs

ICMI '21, October 18–22, 2021, Montréal, QC, Canada

ACM ISBN 978-1-4503-8481-0/21/10…$15.00
https://doi.org/10.1145/3462244.3479902

to interact with the multimodal VA and appreciated the improved human-likeness of human-VA conversations.

CCS CONCEPTS

• **Human-centered computing → Mixed / augmented reality**; *User studies*; **Natural language interfaces**; *Sound-based input / output*.

KEYWORDS

intelligent voice assistant, augmented reality, demonstrative pronouns, unparseable query, ambiguous query, mobile interaction, user experience, mixed-methods study

ACM Reference Format:
Jaewook Lee, Sebastian S. Rodriguez, Raahul Natarrajan, Jacqueline Chen, Harsh Deep, and Alex Kirlik. 2021. What's This? A Voice and Touch Multimodal Approach for Ambiguity Resolution in Voice Assistants. In *Proceedings of the 2021 International Conference on Multimodal Interaction (ICMI '21), October 18–22, 2021, Montréal, QC, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3462244.3479902

1 INTRODUCTION

Voice assistants (VAs – also known as conversational agents or intelligent virtual assistants) are software agents that can perform tasks based on user speech, such as commands and queries, as they can recognize and synthesize speech [12]. For instance, users can query a VA for the Earth's radius, and receive an auditory response, much like in conversations with another human. The implementation and integration of VAs are so ubiquitous in today's technological ecosystem that they can be found in mobile devices, smart home speakers, wearable technology, cars, appliances, and even more. The use of VAs is so widespread that an aggregated report by Microsoft states that 72% of the >2000 respondents from developed countries have used an intelligent VA by 2019 [33]. Today, VAs are primarily used for voice-based search and query, music, and internet of things (IoT) applications [1], as they are predominately found on smartphones and smart speakers [33]. With the proliferation of technology and research aiming towards new forms of interaction in the household, voice-based interaction may as well become a new norm.

VAs are not without flaws. When speaking to a VA, it is not uncommon for a device to misinterpret and reply with a canned response (e.g., "Sorry, I don't understand"), leading to frustration from the user when they speak a recognizable query. In fact, an aggregate of VA interaction logs indicate that current VAs cannot process approximately 14% of voice queries [1]. These *unparseable* queries fall under two categories: 1) the VA's inability to understand the user's voice (e.g., accents, noisy environment, mispronunciations) [32], and/or 2) the VA needs contextual information to answer the query correctly (e.g., the query "apple" can refer to either the fruit or the company) [41]. We focus on the latter case, specifically in ambiguous queries, which are questions that can be interpreted in many different ways or lack a clearly defined subject or object.

This work aims to allow users who query VAs to make use of ambiguous speaking patterns along with the ability to provide context and to allow for more natural, anthropomorphic (i.e., human-like) speaking patterns. There can be many sources of ambiguity, but we focus on allowing users to use demonstrative pronouns (DPs – i.e., "this", "that", "these", and "those") because in human speech, referencing objects through demonstratives are especially common [8]. For instance, a query such as "what is *this*?" contains unknown information embedded in context. Context, according to Dey, is "any information that can be used to characterize the situation of an entity." [11] Though context is useful, it is often underused in human-to-machine interactions [4]. Prior work has relied on additional modalities including gaze and visual screens to provide context to machines; we use touch, because it is the prominent method in which we interact with mobile devices today.

To investigate the resolution of ambiguous queries in interactions between humans and VAs, we developed a touch-enhanced multimodal VA that can act upon detecting DPs in a query. We achieved this by modifying an existing VA so that upon detecting the presence of DPs in a given query, it transitions to an augmented reality-based touch interface. The users can then select features on an object of interest to specify the context of the DP, resolving the ambiguity and sending a new contextualized query with complete information. This approach aims to answer the following research questions:

RQ1: How does touch perform as an additional modality to resolve ambiguity in queries that stems from the use of DPs?
RQ2: How do users use DPs when interacting with VAs?
RQ3: How do users perceive a VA that can process DPs?

We conduct a quantitative study and semi-structured interviews to determine the appropriateness of touch and AR as a multimodal avenue to resolve ambiguity in VA queries. From our collected data, we found that 1) preference towards using the multimodal VA increased linearly as the query became more complex, 2) participants found the standard VA to be more usable than the multimodal VA, 3) participants predominately experienced the same cognitive workload when using the multimodal VA compared to the standard VA, and 4) those who liked the multimodal VA enjoyed using it because DPs made interactions with VAs more anthropomorphic. This work contributes to the ongoing efforts within the VA research community to integrate anthropomorphic features with the goal of intuitive vocal interaction and natural speech patterns.

2 RELATED WORK

This section first summarizes how VAs are used today and investigates the line of research regarding development of VAs. We then review the use of DPs in human-to-human speech, their emergence in queries directed at VAs, and attempts to resolve ambiguities using multimodal approaches.

2.1 The Present and Future of VAs

Currently, VAs can respond to voice queries, set delayed execution of a certain task (e.g., alarms and reminders), or control associated devices in a digital ecosystem (e.g., smart homes). Today, users mainly use VAs for voice search, music, IoT applications, alarms, weather, and "small talk" such as jokes [1, 5, 10, 33]. With the number of smartphone users in the world projected to increase from 3.6 billion in 2020 to 4.3 billion by 2023 [37], and the number of smart speakers in the world projected to increase from 320 million units in 2020 to 640 million units by 2024 [46], voice is already a vital part of how we interact with technology. However, there are still

many that are just somewhat satisfied or not satisfied with VAs [33]. Additionally, despite the growing number of users and capabilities, most users still only ask VAs to perform simple tasks [5, 10, 40]. Both phenomena occurred due to various reasons, including lack of features, errors in processing basic queries, limited knowledge of possible capabilities, and privacy concerns [40].

A systematic literature review shows that ongoing research on VAs fall into one of three categories: 1) improving the technology itself by providing better voice recognition, creating human-like speech, adding multilingual support, et cetera, 2) improving user privacy to reinforce trust, and 3) explaining how VA technology is being used today [10]. This suggests that users desire both novel features and privacy when envisioning future VAs. We contribute to the prior effort by prototyping a touch-enhanced multimodal VA that can process ambiguous queries containing DPs.

2.2 Resolving Contextual Ambiguities

2.2.1 What are Demonstrative Pronouns (DPs)? Task-oriented dialogue between two entities (usually people) is often used to complete a particular task, coordinating ideas and actions that will allow all entities to achieve a certain goal [8]. Prior research that studied a corpus of task-oriented dialogues concluded that 51% of object references are demonstratives, specifically using "this", "that", "these", and "those" [8]. DPs are often used to refer to an object depending on proximal ("this", "these") and distal ("that", "those") locations [18]. Despite the ambiguity stemming from DPs, they are used often in everyday dialogue. The likely explanation is that DPs are often used alongside non-linguistic gestures which provide additional forms of context and information beyond speech [17]. Examples of non-linguistic gestures include gaze, facial expressions, and body language. However, for a VA to process the context or referents of DPs, it often requires additional modalities because voice alone may not be sufficient in describing the objects of interest. Thus, we propose touch as an additional modality to study whether DPs can assist human-VA interactions.

2.2.2 Resolving Ambiguities Without Additional Modalities. There have been some approaches that aim to resolve ambiguity in queries without relying on additional interaction modalities. Among those, a popular approach is to design the system to ask relevant follow up questions to the users. For example, Li experimented with multi-turn conversations, which allow VAs to probe for further information to provide a more specific response to the user's query (e.g., if a user asks for a dish, the system can ask for specific features, such as the temperature of the dish) [29]. A similar approach is the System Ask–User Respond (SAUR) paradigm, in which a recommendation system continues to ask questions based on a collection of user reviews until it feels confident enough to recommend a product [48]. While prior research addresses some contexts, such as a user's preference, they do not seem to address the need for spatial context (i.e., objects in the space the user is in), which we focus on in this paper.

2.2.3 Resolving Ambiguities With Additional Modalities. Various approaches have relied on additional modalities to resolve ambiguities. Multimodal approaches often aim to resolve ambiguities stemming from spatial context. Many of the prior approaches have relied on gaze to indicate, detect, and share information about the world with the VA. For example, Prasov and Chai proposed adding gaze detection to speech-based interfaces to improve the accuracy in resolving spatial contexts [39]. Similarly, Elepfandt and Grund used gaze to aid senior citizens and disabled users using VAs, finding that they preferred to speak shorter queries that include DPs when given gaze as an additional input modality [13]. Recently, Mayer proposed WorldGaze, a proof-of-concept that gathers additional information through head gaze to replace proximal DPs (i.e., "this") [31]. For instance, if the user asks: "When does this place open?", "this place" is replaced with the name of the restaurant the user's head is pointing at, providing the VA with a contextualized query.

While gaze has shown great promise in resolving ambiguities, it is not without flaws. The main concerns with gaze is that it is inaccurate [31] and it may trigger interface elements even when users have no such intentions (i.e., the Midas touch problem) [24]. To circumvent these issues, we chose touch as the additional modality, as it is more accurate and familiar to the users than gaze.

Besides gaze, some researchers have actively studied a visual screen as a potential modality for VAs. Because humans have limited short-term memory [28], many researchers found that shorter responses coupled with visual elements (e.g., images) increase the perceived usability of VAs [2]. For instance, Naik used an Alexa device with a screen to show a list of selections resulting from a given query [34]. For example, when a user asks for movie recommendations, instead of saying a list of names, a VA displays the results on a screen, which users can select from through touch rather than having to remember all of the different film titles. If a VA has access to a visual screen, it can display the results onto a space reachable by the user, thus resolving the need for spatial contexts. Similar to these prior works, we use touch as the additional modality because mobile devices have a screen, which users already have experience interacting with.

3 METHODOLOGY

3.1 Voice + Touch Multimodal VA Prototype

We implement a VA that recognizes ambiguities in queries and aims to resolve them through augmented reality (AR). As discussed, we focus on DPs and attempt to retrieve their related spatial contexts. To accomplish this, we modified the Google Assistant SDK [16] such that if it detects a DP in a query, it transitions to our AR-based touch interface instead of searching immediately. DP detection was achieved by performing a substring search on the query.

We used Unity [45] and Vuforia [23] to train a simple computer vision model to detect predetermined objects for our study and to prototype the AR touch interface. We chose to combine features of an object recognition model and an optical character recognition (OCR) model in our system because most common objects can be identified either by their appearance or labels on them. To train the object recognition model, we took images of our objects from multiple angles and uploaded them to a Vuforia database. We then mimicked results of an OCR model by manually placing outline buttons over features on objects (e.g., labels). We chose to use outline buttons to prevent them from occluding any features. Through this process, we built an AR touch interface that, upon scanning an

object, highlights that object's features. The users can then touch to select features they want to provide as context for the query.

When the user touches a highlighted feature, a word grid is populated, which will serve as a record-keeping interface. At any point, the user can open the grid to 1) check that all of the desired features have been stored, 2) rearrange the ordering of the features, or 3) delete any of the features. Once the grid contains all of the desired features in the correct ordering, the user can then press the "Submit Query" button to search. The interaction flow and related images of the prototype can be seen in Figure 1.

3.2 Design

For our study, we chose everyday tasks aided with VAs. According to a prior report, the leading use case of VAs is to ask queries to retrieve general information (e.g., "Who is Roger Federer?") [25]. The same source reports the most common types of scenarios, from which we chose two categories: querying "products" and "how-to instruction", because both cases are highly dependent on spatial context. We refer to these as the Toy task (i.e., participants will search for information about a toy) and the Recipe task (i.e., participants will search how to prepare a recipe).

In the Toy task, we asked participants to find information about a specific toy. We provided participants with a Lego set (Cole's Speeder Car) and asked them to query for the following information: 1) release date, 2) price, and 3) the instructional manual. In the Recipe task, we asked participants to find recipes that use: 1) one ingredient, 2) three ingredients, and 3) five ingredients. We presented the participants with five ingredients and asked them to select any of the ingredients when forming their queries. This resulted in 6 queries across the 2 tasks.

We conduct this study as a within-subjects design. The independent variable of interest is the VA (either a standard VA – we used Google Assistant – or our multimodal VA). Every participant completes the Toy and Recipe tasks with each VA, resulting in 12 queries per participant. We also note an additional independent variable in the Recipe task: the number of contexts requested (i.e., ingredients), as this affects the complexity of the query. To account for ordering effects, the VA used and tasks were counterbalanced.

The recorded measures during the study are: 1) the participant's preference for using a particular VA, 2) the participant's preference for a particular VA's query result, 3) the usability of the VAs, measured by the System Usability Scale (SUS) [6, 7], and 4) the cognitive workload while using the VAs, measured by the NASA Task Load Index (NASA-TLX) [19, 20]. Upon completion of both tasks, participants were led through a semi-structured interview for 15 minutes where they shared their experience using the VAs, discussing 1) overall preference, 2) differences of features and interactions with each VA, and 3) changes they wish they could make to the VA. We establish the following hypotheses:

> $\mathbf{H}_1$: As the complexity of a query increases, users will prefer using the multimodal VA.
> $\mathbf{H}_2$: Users will prefer the response given by the multimodal VA more than the standard VA.
> $\mathbf{H}_3$: Users will find the multimodal VA to be as usable as a standard VA.
> $\mathbf{H}_4$: Users will experience an equivalent cognitive load using the multimodal VA compared to the standard VA.

For H_3 and H_4, we chose to measure similarity because our system requires users to go through an additional step before receiving an answer. Because of this, we determined that an increase in usability and a reduction in cognitive load when using our system are unlikely and wanted to assure that both are not being sacrificed for the addition of an AR touch interface.

3.3 Procedure

Participants were recruited from a graduate computer science class. Participants first signed the consent form and filled out a brief demographic survey. Participants unfamiliar with a standard VA were instructed in how to query a VA. All participants then completed a tutorial on how to use the multimodal VA. Then, participants completed the two tasks using both of the VAs. After each task, the participants completed the SUS and NASA-TLX surveys, and chose which VA they preferred to use and which responses they liked most. Upon completion of both tasks, participants completed the semi-structured interview. Participants were then debriefed. The study procedure is visualized in Figure 2.

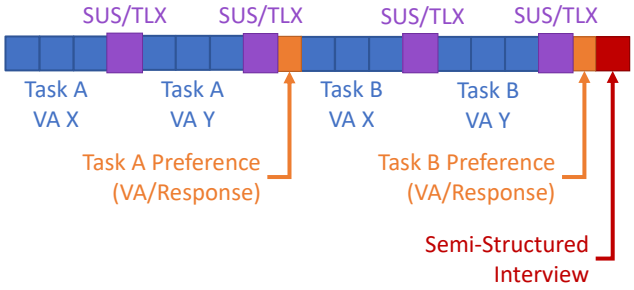

Figure 2: Timeline of the study procedure. Tasks and VA used were counterbalanced per block: A/B = either Toy task or Recipe task; X/Y = either standard VA or multimodal VA.

4 RESULTS

We collected 20 participant interactions and interviews for our analysis. The participants were between 18 and 27 years of age (μ = 21 years, σ = 2 years), with 55% being male. The participants' races were distributed between Asian (70%) and White (30%). Out of the 20 participants, 7 were non-native English speakers.

4.1 Quantitative Results

$\mathbf{H}_1$: As the complexity of a query increases, users will prefer using the multimodal VA. As the preference for the multimodal VA is a binary outcome, we use logistic regression to model preference given with the query complexity as the predictor variable. Since the Toy task features the same complexity over every query (only one context), we focus our analysis on the Recipe task. We employed a multiple logistic regression to predict preference for the multimodal VA using query complexity, familiarity with VAs, and age as predictors. The regression indicates that complexity (β = 1.94, p <0.001) and age (β = -0.085, p <0.05) strongly predict

preference for the multimodal VA. Chi-squared tests additionally show distinct preferences for 1 object ($\chi^2(1, 20) = 9.8$, $p < 0.01$) and 5 objects ($\chi^2(1, 20) = 12.8$, $p < 0.001$), with no distinct preference for 3 objects ($\chi^2(1, 20) = 3.2$, $p = 0.07$). Preferences for the multimodal VA are visualized in Figure 3.

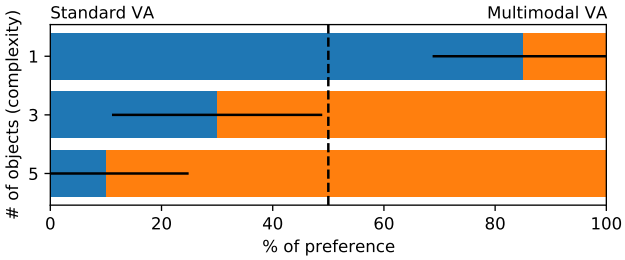

Figure 3: Aggregated VA preferences against complexity in the Recipe task. Blue and orange denote the percentage of participants who preferred the standard or multimodal VA, respectively.

H_2: Users will prefer the response given by the multimodal VA more than the standard VA. The response variable for this hypothesis captures whether participants liked the response given by the standard VA, the multimodal VA, or both. Since we are focused on the participant preferring the multimodal VA, we note a positive response (1) whether the participant liked multimodal VA exclusively, and a negative response (0) otherwise. We use a series of chi-square tests to determine whether there was a difference in exclusively preferring the multimodal VA over all other options. For the Toy task, all queries were non-significant for exclusive preference (query 1: $\chi^2(1, 20) = 1.8$, $p = 0.18$; query 2: $\chi^2(1, 20) = 0.8$, $p = 0.37$; query 3: $\chi^2(1, 20) = 3.2$, $p = 0.07$). For the Recipe task, all numbers of query objects were non-significant for exclusive preference (1 object: $\chi^2(1, 20) = 0.2$, $p = 0.65$; 3 objects: $\chi^2(1, 20) = 3.2$, $p = 0.07$; 5 objects: $\chi^2(1, 20) = 1.8$, $p = 0.18$). Aggregated preferences are visualized in Figure 4.

We additionally use multiple logistic regression to predict if the task, query complexity, familiarity with VAs, and age predict preference towards multimodal VAs. Modeling the positive response indicates that the task ($\beta = 1.02$, $p < 0.01$) and age ($\beta = -0.05$, $p < 0.05$) significantly predict the user's exclusive preference for the multimodal VA, with complexity and familiarity being non-predictive.

H_3: Users will find the multimodal VA to be as usable as a standard VA. We calculate the interpreted SUS score given by responses after every task and intervention, and use equivalence testing to determine the same level of usability between conditions through the two one-sided t-tests (TOST) procedure. The upper and lower bounds for the TOST was set at 6 ($-\Delta_L = \Delta_U = 6$), as prior research demonstrates that grounding SUS scores towards adjective ratings provides a categorical average range of 12.56 [3], rounded down to 12 for conservativeness. For the Toy task, the TOST indicates that there is no significant equivalence in usability between VAs, further suggesting superiority for the standard VA ($\mu_s = 77.13$, $\mu_m = 67.86$, difference 95% CI: [1.07, 17.427], $p = 0.79$). For the Recipe task, the TOST indicates that there was no significant equivalence in usability between VAs ($\mu_s = 72.38$, $\mu_m = 71.5$, difference 95% CI: [-7.87, 9.62], $p = 0.12$). SUS confidence intervals are visualized in Figure 5. Thus, we fail to reject the H_3 TOST null hypothesis that there is no significant equivalence and conclude that participants found the standard VA to be more usable than the multimodal VA.

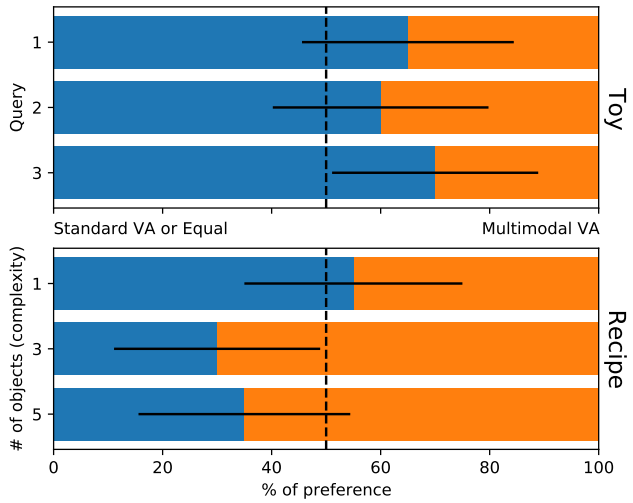

Figure 4: Aggregated VA response preferences for the Toy task (top) and the Recipe task (bottom). Blue denotes negative response (participant preferred the standard VA responses exclusively or both responses equally) and orange denotes positive response (participant preferred the multimodal VA response exclusively).

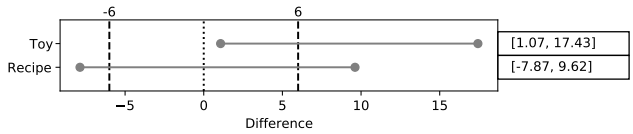

Figure 5: SUS mean difference 95% CIs plotted against TOST bounds (dashed vertical lines) for the Toy task and the Recipe task. No significant equivalences were found.

H_4: Users will experience an equivalent cognitive load using the multimodal VA compared to the standard VA. We run our statistical analysis with the raw TLX scores, as reporting these are acceptable according to a meta-analysis of NASA-TLX usage [19]. We used the TOST procedure for each scale to determine equal cognitive load between the VAs per task. The upper and lower bounds for the TOST was set at 18 ($-\Delta_L = \Delta_U = 18$), following a meta-analysis of NASA-TLX scores determining that the average standard deviation across all scales for applications in handheld devices is 18.66 [22], rounded down to 18 for conservativeness. For the Toy task, the TOST showed equivalence on all TLX scales ($p < 0.05$), with the average of all TLX scales showing equivalence ($p < 0.001$). For the Recipe task, the TOST showed equivalence for all TLX scales ($p < 0.05$), except for physical ($p = 0.11$), with the averages of all TLX scales showing equivalence as well ($p < 0.001$). TLX scales

and confidence intervals are visualized for both tasks in Figure 6. Thus, we predominately reject the H_4 TOST null hypothesis and determine a significant workload similarity between the standard VA and our multimodal VA.

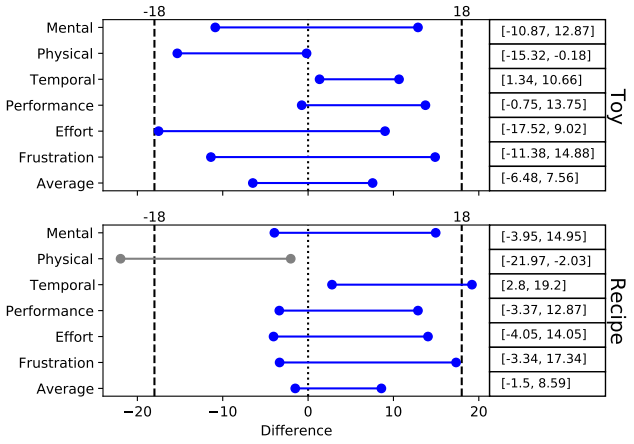

Figure 6: NASA-TLX mean difference 95% CIs plotted against TOST bounds (dashed vertical lines) for the Toy task (top) and the Recipe task (bottom). Blue CIs indicate significant equivalence.

4.2 Interview Results

The interview data was collected and analyzed through thematic analysis, which was done using affinity diagramming, a design method where researchers summarize patterns within responses by iteratively clustering quotes [30]. Below are the themes that we found by using this process:

4.2.1 Usage Patterns of Demonstrative Pronouns. During the study, we observed that 18 out of the 20 participants relied solely on "this" and "these" in their queries, even though they were told that "that" and "those" are viable DPs as well. Many stated that all of the objects were proximal, so they had no reason to say DPs that refer to distal objects. When we asked the two participants who used "that" and "those" as to why they used them, P5 said: "...because it understood 'this' as 'dis' or 'bees'," while P15 said it was "...simply because [they were] curious."

We also observed differences in preference between "this" and "these." The majority of the participants found using the word "these" to be helpful because there was no need to vocalize many objects' names and any additional features for their queries. However, the majority of the participants found the word "this" to be not effective other than in the following circumstances: 1) words on the object are difficult to pronounce, 2) the object is unfamiliar to the user, or 3) the object contains no words. The first scenario especially troubled many participants. For example, P2, a native English speaker, pronounced "arrabbiata" as "array-bbiata." In addition, P1 noted that "...[touch input] is good if you are a non-native speaker and so you cannot pronounce some things." Many objects contain brand names, labels, and other words that can be challenging to pronounce for both native and non-native speakers. The second scenario also posed challenges. 4 out of 20 participants pronounced "NinjaGo" as "Ninja G Q" because the font used on the box makes the "O" look like a "Q". While some were able to correct themselves because of prior knowledge in either Legos or the Ninjago television series, these 4 participants lacked the expertise to do so. Lastly, 3 participants commented that the modified VA would have been better than the standard VA if an unfamiliar object has no labels on it because then, it would be impossible to be specific.

4.2.2 Simplicity vs. Detailedness. Participants preferred to do the Toy task using the standard VA. They reasoned that when there is only one object of interest, it is simpler to query directly for that object's name than to use the word "this" and invoke the AR solution. Those that preferred to use the multimodal VA reasoned that they prefer detailedness despite the cost (i.e., additional time needed for touch). Alternatively, participants preferred to do the Recipe task using the multimodal VA. For the one ingredient scenario, the participants still preferred to use the standard VA. For the three and five ingredients scenarios, the participants reasoned that: "when looking for a recipe, it needs to contain the exact ingredients I have in front of me" (P3), but "it is difficult to say everything I want to say all at once" (P20). Those that still preferred to use the standard VA reasoned that "it is still faster to say the names of all of the ingredients than to use touch" (P15). Across both tasks, on average, the participants stated that they would use the multimodal VA if the query needed more than 3 to 5 features and the standard VA if otherwise.

In general, preference between the two systems seemed to be determined by whether participants prefer simplicity or detailedness. Those who created simple queries using just a one-word descriptor per object prioritized receiving an answer as quickly as possible over accuracy and relevance of the results, hence they preferred the standard VA. Many others formed queries using more than a one-word descriptor per object because they wanted to create a detailed query to receive a result most relevant to their current situation. Participants who formed long and complex queries found them tedious and difficult to say aloud, and therefore preferred touch.

4.2.3 Anthropomorphic VA. One key aspect of this study is whether a VA that can process DPs is more anthropomorphic and whether users like to interact with it. 13 out of 20 participants found using DPs a more natural, human-like way of speaking to VAs. Their rationale is well summarized by P7: "... so we should be able to speak to devices just like how we talk to people." Others initially found DPs to be awkward. This seemed to be especially true for those with prior experience using VAs. For example, P15 mentioned: "I am not used to speaking to a voice assistant this way...I always enunciate every word clearly and make sure that the questions are as specific as possible." However, 5 of them added that as they gained additional experience using DPs, they were able to use the multimodal VA as comfortably as the standard VA. This suggests that while DPs may require some time to become accustomed to, with additional experience, they can assist in creating more natural human-to-VA interactions.

5 DISCUSSION

5.1 Processing Both Specific and Ambiguous Queries

From our data, it is apparent that the participants increasingly prefer the multimodal VA over the standard VA as the complexity of the query (i.e., the number of features/objects it needs) increases. Participants prefer to use the standard VA for simpler queries because vocalizing a few features takes less time and effort than touch. Conversely, participants prefer to use the multimodal VA for complex queries because memorizing the names and features of multiple objects and forming a coherent query using all of them is difficult. When saying a long query aloud, many participants had to repeat themselves several times, because the standard VA executed the search as soon as they paused to think which word comes next. The multimodal VA helps to allow participants to take as long as they want when forming complex queries. Additionally, the participants stated that while touch is much slower than voice for one object, the gap reduces as the number of objects increases. The pros and cons of each system suggest that while the standard VA needs additional features, the multimodal VA is unnecessary in some scenarios. This finding reinforces long-established design principles in human-centered design, such as valuing simplicity without giving up interesting or novel features [35, 36]. An ideal multimodal VA would simplify the interaction process by keeping features of a standard VA while providing a new modality to provide context.

This principle proposes that a working implementation of the multimodal VA should behave differently based on the complexity of the scenario. For example, if the task is to search for the price of a furniture, many users will form a simple query, as many participants did during the Toy task. However, if the task is to find the cheapest furniture from a list of 5, the users may find it easier to search using the DP "these." The final VA should work for both of these scenarios, meaning that one system should be able to process both specific and ambiguous queries.

To answer whether touch as multimodality could successfully complement standard VAs, we compared the usability and cognitive load of both VAs. We conclude that the standard VA is more usable than the multimodal VA, but the cognitive load across the usage of both systems is equivalent except for physical effort (participants spent more physical effort with the multimodal VA in the Recipe task). Sacrifices in both usability and physical effort to add a novel feature support the need for a combined system. Users should be able to use the standard VA without any sacrifices and only rely on ambiguity when necessary. Equivalent cognitive load implies that if one user chooses to say an ambiguous query, and another user does not, they will both be able to complete the same task without their preference of queries impacting their cognitive load.

5.2 Benefits of Anthropomorphic VAs

The term anthropomorphism describes "the tendency to imbue the real or imagined behavior of nonhuman agents with human-like characteristics, motivations, intentions, or emotions" [14]. In recent years, many researchers looked into creating an anthropomorphic VA with a focus on human-like voice [38]. For example, Chérif and Lemoine gave a VA a human-like voice and found that an anthropomorphic VA left a stronger social presence and formed stronger bonds with the users than a VA with a synthetic voice [9]. During this study, we enabled the participants to use DPs when interacting with VAs to mimic crucial characteristics of human-to-human conversations (i.e., ambiguity). From this experience, most participants expressed satisfaction with using the multimodal VA because of the improved human-likeness of their interactions. For instance, P5 and P15 stated that when they interact with a standard VA, they have to speak without any ambiguity, which can sometimes be difficult. P15 provided a hypothetical scenario: if one was visiting a certain landmark (e.g., a statue), and it was unlabeled, a standard VA would not be able to query information about the landmark, beyond asking "where am I?" Both further commented that because they did not have to be specific when interacting with the multimodal VA, some interactions felt more natural, as if they are "talking to their friends." This suggests that a VA can be perceived as more anthropomorphic if it can understand more human-like queries (e.g., understand ambiguity).

5.3 VAs for Non-Native English Speakers

An ongoing problem within the automatic speech recognition (ASR) domain is the inability of ASR systems (including VAs) in understanding different accents, some of which arise from non-native English speakers [21, 26, 43, 44]. When prior research studied state-of-the-art ASR systems that were developed by Amazon, Apple, Google, IBM, and Microsoft, they noticed that all five exhibited substantial racial disparities, where the average word error rate was 35% for African-American speakers compared with 19% for White speakers [27]. While some VAs now support different English accents and languages other than English, coverage is not comprehensive, which forces some to interact using a non-native language or excludes them from using VAs completely [47]. This is often seen in technologies that rely on machine learning because they are often trained on data that does not represent all races equally. A commonly pitched solution to this problem is collecting non-biased data, but this can take a very long time. Additionally, work by Gebru and Denton argues that doing so will not "reduce harms caused by machine learning to dataset bias" as individuals who are in marginalized communities did not perceive much changes in VAs after they have been trained on "non-biased" data [15]. We propose DPs as a partial solution to addressing the average word error rate disparities across non-native English speakers.

The participant pool included 7 non-native English speakers, which allowed us to note problems caused by language barriers and accents. For instance, participants mispronounced words such as "arrabbiata." In addition, homophones sometimes troubled the non-native English speakers. For instance, the standard VA sometimes comprehended "Cole's" as "Kohl's," which has a very similar pronunciation but is a distinct word. These participants all agreed that using DPs helped because they were relatively easier to say than some of the more complicated words found on products. However, even DPs had their limitations. When participants interacted with the multimodal VA using DPs, the VA had no trouble understanding pronunciations of 6 out of 7 non-native English speaking participants. However, it interpreted P5's pronunciation of "this" as "dis" or "bees." Fortunately, the multimodal VA had no trouble

understanding the pronunciations of "these" from the 7 participants. In many cases, interactions were enhanced thanks to the capability of introducing words easier to recognize instead of words misheard due to accents.

5.4 Design Recommendations

When designing a touch-enhanced multimodal VA, two takeaways drive our recommendations: 1) touch is not a universal modality, and 2) users should have the autonomy to choose whether to use any additional modalities in VAs.

Participants overall preferred the multimodal VA over the standard VA in scenarios requiring queries involving more than 3 to 5 features ("these"), but tedious otherwise ("this"). Additionally, some participants complained that some of the AR touch areas were too small (i.e., the fat finger problem [42]). To address this, we suggest providing users with other modalities in addition to touch. During the interview, many participants expressed that while touch worked well for plural DPs, they wanted the multimodal VA to automatically pull features out from the object of interest if they said either of the singular DPs. They reasoned that when using singular DPs, there should only be one referent, which the mobile device should be able to decipher based on 1) which objects are within the camera frame, 2) which objects are closer to the center of the screen, and 3) whether each object is nearby ("this") or far away ("that") from the camera. This notion is best summarized by P19, who stated: "When I say 'this,' I am talking about one object in front of me... so maybe it should just search using the words it found instead of showing them to me." Many participants noted that they would be pointing their phones at the object of interest when using a singular DP. Because that object is likely to be aligned with the center of the camera frame, mobile devices should be able to capture the referent using a raycast extending from the center of the screen. This approach still relies on AR, but it does not depend on touch, making the interactions more automatic. Alternatively, the authors of WorldGaze aimed to resolve ambiguity in queries by allowing the system to extract features located in the direction of the user's head gaze [31]. Although gaze tracking on mobile devices is inaccurate, it is still a reliable way of knowing the object of interest because when users say "this," they will likely be looking at the referent [31]. However, while gaze works well with singular DPs, touch may be a better solution for plural DPs because selecting multiple contexts through gaze would require shifts in head rotation, which can be more tedious than tapping on a screen. By providing either of these additional modalities in addition to touch, we are hopeful that we can create a more satisfactory experience for users as they use both singular and plural DPs.

Finally, as discussed in Section 5.1, additional modalities, including touch, may not be necessary for simpler queries. To address this, instead of creating a VA that can only process queries with or without ambiguities, it seems best to design a system that can process both and letting users decide how they want to interact with the VA.

5.5 Limitations and Future Work

The presented work has its limitations. The system is merely a Wizard-of-Oz prototype using a simplistic computer vision model to detect predetermined characters and objects and do not extend further to other objects. This limits the freedom users could have when querying, as we cannot derive insight into what type of objects or situations users would contextualize using a multimodal VA. Additionally, our resulting participant sample was biased towards Asian users because of the reduced sampling availability due to COVID-19. This resulted in a fruitful discussion about the needs of non-native English speakers, but it may not represent issues affecting all users.

Nevertheless, this specific sample has presented a potential line of research for the inclusivity of users who may not speak the primary operating language of the VA. Analogous to human-human interactions where two people who do not speak the same language communicate with contextual cues and body language, multimodal interactions may serve as the contextual bridge for users who have a challenging time speaking the VA's operating language. This work reinforces the use of anthropomorphism as a central characteristic of human-VA interaction, and research should continue investigating other methods to incorporate it as such.

6 CONCLUSION

We investigated whether touch is a useful modality for resolving ambiguity in VA queries and how users take advantage of DPs as they interact with VAs. Since current VAs are not equipped to handle ambiguous queries, we created a Wizard-of-Oz prototype to address DPs and ran a comparison study. We observed that while users preferred to use the standard VA for simpler queries (e.g., a query about one object of interest), users preferred the modified VA more as the complexity of the query rose. Additionally, the modified VA performed extremely well in some specific scenarios, such as when the user wanted to include a lot of detail into their queries or if they were a non-native English speaker. However, we note that touch is not a universal modality for resolving ambiguities. Touch was too slow in resolving ambiguities of simpler queries, and many participants complained that some of the AR touch areas were too small to interact with. While touch is a useful modality for resolving ambiguity, using it alongside other modalities (e.g., gaze) and creating a system that provides users with the options for either using DPs or not is recommended. Furthermore, participants were comfortable with using DPs to interact with VAs, even though many agreed that such situations were unfamiliar to them at first. Finally, even though the participants found the standard VA to be more usable than the multimodal VA, they predominantly experienced the same cognitive load (i.e., not statistically different) across both systems. These discoveries illustrate the effectiveness of touch-based multimodality in resolving ambiguities engendered by DPs within queries, to someday allow VAs to be perceived as another human entity.

ACKNOWLEDGMENTS

This research was funded by the Grainger College of Engineering Illinois Scholars Undergraduate Research (ISUR) Program at the University of Illinois at Urbana-Champaign. We would also like to thank David Lindlbauer at Carnegie Mellon University for his feedback on earlier drafts of this paper.

REFERENCES

[1] Tawfiq Ammari, Jofish Kaye, Janice Y. Tsai, and Frank Bentley. 2019. Music, Search, and IoT: How people (really) use voice assistants. *ACM Transactions on Computer-Human Interaction* 26 (2019). Issue 3. https://doi.org/10.1145/3311956

[2] Rianna R Baeza and Anil R Kumar. 2019. Perceived Usefulness of Multimodal Voice Assistant Technology. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 63, 1 (nov 2019), 1560–1564. https://doi.org/10.1177/1071181319631031

[3] Aaron Bangor, Philip Kortum, and James Miller. 2009. Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *J. Usability Studies* 4, 3 (May 2009), 114–123.

[4] Bruno G. Bara. 2010. *Cognitive Pragmatics: The Mental Processes of Communication.* The MIT Press.

[5] Frank Bentley, Chris Luvogt, Max Silverman, Rushani Wirasinghe, Brooke White, and Danielle Lottridge. 2018. Understanding the Long-Term Use of Smart Speaker Assistants. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–24. https://doi.org/10.1145/3264901

[6] J. Brooke. 1996. SUS: A 'Quick and Dirty' Usability Scale.

[7] John Brooke. 2013. SUS: a retrospective. *Journal of usability studies* 8 (2013). Issue 2.

[8] D K Byron and J F Allen. 1998. Resolving Demostrative Anaphora in the TRAINS93 Corpus. In *Proceedings of Second Colloquium on Discurse Anaphora and Anaphor Resolution (DAARC2).* http://hdl.handle.net/1802/1456

[9] Emna Chérif and Jean-François Lemoine. 2019. Anthropomorphic virtual assistants and the reactions of Internet users: An experiment on the assistant's voice. *Recherche et Applications en Marketing (English Edition)* 34, 1 (2019), 28–47. https://doi.org/10.1177/2051570719829432

[10] Allan de Barcelos Silva, Marcio Miguel Gomes, Cristiano André da Costa, Rodrigo da Rosa Righi, Jorge Luis Victoria Barbosa, Gustavo Pessin, Geert De Doncker, and Gustavo Federizzi. 2020. Intelligent personal assistants: A systematic literature review. *Expert Systems with Applications* 147 (2020), 113193. https://doi.org/10.1016/j.eswa.2020.113193

[11] Anind K. Dey. 2001. Understanding and using context. *Personal and Ubiquitous Computing* 5 (2001). Issue 1. https://doi.org/10.1007/s007790170019

[12] IBM Cloud Education. 2020. Conversational AI. https://www.ibm.com/cloud/learn/conversational-ai.

[13] Monika Elepfandt and Martin Grund. 2012. Move it there, or not?: The design of voice commands for gaze with speech. In *Proceedings of the 4th Workshop on Eye Gaze in Intelligent Human Machine Interaction, Gaze-In 2012.* https://doi.org/10.1145/2401836.2401848

[14] Nicholas Epley, Adam Waytz, and John T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. *Psychological Review* 114, 4 (Oct 2007), 864–886. https://doi.org/10.1037/0033-295x.114.4.864

[15] Timnit Gebru and Emily Denton. 2020. Fairness Accountability Transparency and Ethics in Computer Vision. https://sites.google.com/view/fatecv-tutorial/home?authuser=0.

[16] Google. 2017. Google Assistant SDK. https://developers.google.com/assistant/sdk/.

[17] Jeanette K Gundel, Nancy Hedberg, and Ron Zacharski. 2004. Demonstrative pronouns in natural discourse. In *Proceedings of the 5th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC 2004).*

[18] M A K Halliday and C.M.I.M. Matthiessen. 2013. *Halliday's Introduction to Functional Grammar.* Taylor & Francis. https://books.google.com/books?id=odUqAAAAQBAJ

[19] Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9 (2006), 904–908. https://doi.org/10.1177/154193120605000909 arXiv:https://doi.org/10.1177/154193120605000909

[20] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Human Mental Workload*, Peter A. Hancock and Najmedin Meshkati (Eds.). Advances in Psychology, Vol. 52. North-Holland, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

[21] Drew Harwell. 2018. The Accent Gap. https://www.washingtonpost.com/graphics/2018/business/alexa-does-not-understand-your-accent/.

[22] Morten Hertzum. 2021. Reference values and subscale patterns for the task load index (TLX): a meta-analytic review. *Ergonomics* 0, 0 (2021), 1–10. https://doi.org/10.1080/00140139.2021.1876927 arXiv:https://doi.org/10.1080/00140139.2021.1876927 PMID: 33463402.

[23] PTC Inc. 2015. Vuforia. https://developer.vuforia.com/.

[24] Robert J. K. Jacob. 1991. The use of eye movements in human-computer interaction techniques: what you look at is what you get. *ACM Transactions on Information Systems* 9, 2 (Apr 1991), 152–169. https://doi.org/10.1145/123078.128728

[25] Bret Kinsella. 2018. What People Ask Their Smart Speakers. https://voicebot.ai/2018/08/01/what-people-ask-their-smart-speakers/.

[26] Fedor Kitashov, Elizaveta Svitanko, and Debojyoti Dutta. 2018. Foreign English Accent Adjustment by Learning Phonetic Patterns. arXiv:1807.03625 [cs.SD]

[27] Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America* 117, 14 (2020), 7684–7689. https://doi.org/10.1073/pnas.1915768117

[28] Ludovic Le Bigot, Loïc Caroux, Christine Ros, Agnès Lacroix, and Valérie Botherel. 2013. Investigating memory constraints on recall of options in interactive voice response system messages. , 106–116 pages. https://doi.org/10.1080/0144929X.2011.563800

[29] Toby Jia Jun Li. 2020. Multi-Modal Interactive Task Learning from Demonstrations and Natural Language Instructions. In *UIST 2020 - Adjunct Publication of the 33rd Annual ACM Symposium on User Interface Software and Technology.* Association for Computing Machinery, Inc, 162–168. https://doi.org/10.1145/3379350.3415803

[30] Bella Martin and Bruce M. Hanington. 2012. *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions* (digital ed ed.). Rockport Publishers.

[31] Sven Mayer, Gierad Laput, and Chris Harrison. 2020. Enhancing Mobile Voice Assistants with WorldGaze. In *Conference on Human Factors in Computing Systems - Proceedings.* Association for Computing Machinery. https://doi.org/10.1145/3313831.3376479

[32] Rich McCormick. 2017. Please don't make me talk to voice assistants anymore. https://www.theverge.com/2017/6/6/15744106/voice-assistants-siri-dont-make-me-talk.

[33] Microsoft. 2019. Voice Report. https://advertiseonbing-blob.azureedge.net/blob/bingads/media/insight/whitepapers/2019/04%20apr/voice-report/bingads_2019_voicereport.pdf.

[34] Vishal Ishwar Naik, Angeliki Metallinou, and Rahul Goel. 2018. Context aware conversational understanding for intelligent agents with a screen. In *32nd AAAI Conference on Artificial Intelligence, AAAI 2018.* 5325–5332. www.aaai.org

[35] Donald A. Norman. 2007. Simplicity is highly overrated. *Interactions* 14, 2 (Mar 2007), 40–41. https://doi.org/10.1145/1229863.1229885

[36] Donald A. Norman. 2013. *The design of everyday things.* Basic Books.

[37] S. O'Dea. 2020. Smartphone users worldwide 2016-2023. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/.

[38] Michela Patrizi, Maria Vernuccio, and Alberto Pastore. 2021. "Hey, voice assistant!" How do users perceive you? An exploratory study. *Sinergie Italian Journal of Management* 39, 1 (Feb 2021), 173–192. https://doi.org/10.7433/s114.2021.10

[39] Zahar Prasov and Joyce Y. Chai. 2008. What's in a Gaze? The role of eye-gaze in reference resolution in multimodal conversational interfaces. In *International Conference on Intelligent User Interfaces, Proceedings IUI.* 20–29. https://doi.org/10.1145/1378773.1378777

[40] PwC. 2018. Consumer Intelligence Series: Prepare for the voice revolution. https://www.pwc.com/us/en/advisory-services/publications/consumer-intelligence-series/voice-assistants.pdf.

[41] Tony Russell-Rose and Tyler Tate. 2013. Chapter 6 - Displaying and Manipulating Results. In *Designing the Search Experience*, Tony Russell-Rose and Tyler Tate (Eds.). Morgan Kaufmann, 129–166. https://doi.org/10.1016/B978-0-12-396981-1.00006-9

[42] Katie A. Siek, Yvonne Rogers, and Kay H. Connelly. 2005. Fat Finger Worries: How Older and Younger Users Physically Interact with PDAs. In *Human-Computer Interaction - INTERACT 2005*, Maria Francesca Costabile and Fabio Paternò (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 267–280.

[43] Rachael Tatman. 2017. Gender and Dialect Bias in YouTube's Automatic Captions. (2017), 53–59. https://doi.org/10.18653/v1/w17-1606

[44] Rachael Tatman and C. Kasten. 2017. Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *INTERSPEECH.*

[45] Unity Technologies. 2005. Unity. https://unity.com/.

[46] Lionel S. Vailshery. 2020. Smart speaker installed base worldwide 2020 and 2024. https://www.statista.com/statistics/878650/worldwide-smart-speaker-installed-base-by-country/.

[47] Yunhan Wu, Daniel Rough, Anna Bleakley, Justin Edwards, Orla Cooney, Philip R. Doyle, Leigh Clark, and Benjamin R. Cowan. 2020. See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers. *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services* (2020). https://doi.org/10.1145/3379503.3403563

[48] Yongfeng Zhang, Xu Chen, Qingyao Ai, Liu Yang, and W. Bruce Croft. 2018. Towards conversational search and recommendation: System Ask, user respond. In *International Conference on Information and Knowledge Management, Proceedings*, Vol. 10. ACM, 177–186. https://doi.org/10.1145/3269206.3271776