

See What I'm Saying? Comparing Intelligent Personal Assistant Use for Native and Non-Native Language Speakers

Yunhan Wu
University College Dublin
yunhan.wu@ucdconnect.ie

Daniel Rough
University College Dublin
daniel.rough@ucd.ie

Anna Bleakley
University College Dublin
anna.bleakley@ucdconnect.ie

Justin Edwards
University College Dublin
justin.edwards@ucdconnect.ie

Orla Cooney
University College Dublin
orla.cooney@ucdconnect.ie

Philip R. Doyle
University College Dublin
philip.doyle1@ucdconnect.ie

Leigh Clark
Swansea University
l.m.h.clark@swansea.ac.uk

Benjamin R. Cowan
University College Dublin
benjamin.cowan@ucd.ie

Abstract

Limited linguistic coverage for Intelligent Personal Assistants (IPAs) means that many interact in a non-native language. Yet we know little about how IPAs currently support or hinder these users. Through native (L1) and non-native (L2) English speakers interacting with Google Assistant on a smartphone and smart speaker, we aim to understand this more deeply. Interviews revealed that L2 speakers prioritised utterance planning around perceived linguistic limitations, as opposed to L1 speakers prioritising succinctness because of system limitations. L2 speakers see IPAs as insensitive to linguistic needs resulting in failed interaction. L2 speakers clearly preferred using smartphones, as visual feedback supported diagnoses of communication breakdowns whilst allowing time to process query results. Conversely, L1 speakers preferred smart speakers, with audio feedback being seen as sufficient. We discuss the need to tailor the IPA experience for L2 users, emphasising visual feedback whilst reducing the burden of language production.

CCS Concepts

• **Human-centered computing** → **User studies**; **Natural language interfaces**; *Accessibility design and evaluation methods.*

Keywords

speech interface; voice user interface; intelligent personal assistants; non-native speakers

ACM Reference Format:
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. In *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '20), October 5–8, 2020, Oldenburg, Germany.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3379503.3403563

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

MobileHCI '20, October 5–8, 2020, Oldenburg, Germany

© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-7516-0/20/10…$15.00
https://doi.org/10.1145/3379503.3403563

1 Introduction

The proliferation of voice based Intelligent Personal Assistants (IPAs) like Google Assistant, on smart speakers and smartphones, has made speech a common interaction modality [8]. Although many IPAs can now be used in languages other than English, coverage and supported functionality is by no means comprehensive, varying by device and assistant (e.g., [28]). This creates a barrier for those whose first language is not fully supported, forcing them to interact in a non-native language or face being excluded from IPA use. Prominent work in HCI looks at IPA user experience [14, 29, 34] almost exclusively from the perspective of first language (L1) English speakers, leaving the experience of users who engage with IPAs in a non-native language (such as L2 speakers) unclear.

The work presented contributes important insight into how L2 speakers experience IPAs. Our work aims to 1) map significant dimensions of L2 user experience whilst also 2) identifying how aspects of the two most popular devices for IPA use (smartphones and smart speakers [39]) support or hinder L2 users. We also compare this to L1 speaker experiences so as to emphasise the contrasting needs of these speaker groups. To achieve this we carried out a study where, following interactions with Google Assistant on both a smartphone and on a smart speaker, L1 and L2 English speakers took part in a semi-structured interview devised to gain insight into their experiences. Our results highlight a number of clear differences between L1 and L2 speakers' perceptions and experiences of IPA use. We found L1 and L2 speakers varied in their interaction approaches, whereby L2 speakers focused heavily on their pronunciation as opposed to L1 emphasising the need for simple, succinct and well planned utterances. Both emphasised the need for speech adaptation, informed by perceived system limitations, yet L2 speakers' adaptation were also driven by their own perceived linguistic limitations. Whereas L1 speakers felt the IPA waited too long to speak after they gave a command, L2 speakers felt the assistant was not sensitive enough to the extra time they needed to produce their command and process the system's utterances. This resulted in the

system regularly interrupting L2 speakers. L2 speakers consistently expressed the desire for IPA design to support lexical retrieval or reduce the need for language production, yet this was not a concern for L1 speakers. We also discovered a clear difference in device preference across the speaker groups. L2 speakers significantly preferred using IPAs on smartphones, because they provided visual feedback that supported their interaction. In contrast, L1 speakers preferred smart speakers, with audio feedback being seen as sufficient to support interaction. Our findings build on recent interest in L2 speakers [35, 36], contributing a deeper insight into L2 IPA experiences across these device types. Our findings emphasise the need to consider L2 user needs so as to be more inclusive of this group. Our work suggests that tailoring IPA interaction by being sensitive to the time needed by L2 users, concentrating on visual feedback and reducing the need for language production in interaction are key design priorities to support L2 users.

2 Related Work

2.1 Interacting with Intelligent Personal Assistants

Recent research in HCI has predominantly focused on understanding IPA user experience from the L1 perspective. It highlights that a major benefit of speech as a modality lies in its facilitation of multitasking, especially in hands-busy/eyes-busy situations such as driving or looking after children [15, 29] (although the benefits of multitasking with speech are heavily dependent on primary task demand [22]). IPAs are commonly used for information search, controlling music applications, setting alarms and timers, and to control IoT (Internet of Things) devices (e.g. smart lights) [4], through limited question-answer type dialogues [23, 34]. Especially through smart speakers, these assistants can be used by multiple people at once, becoming a social focal point [34]. Although clearly useful, previous work has also identified a number of issues. These include users not trusting IPAs to execute more complex or socially sensitive tasks (e.g. sending a message or calling a contact) [29], perceived problems in accurately recognising accented speech [14], and the humanlike design of IPAs poorly signalling actual IPA capabilities [9, 14, 20, 29]. Privacy and data collection practices are also a concern for users [10, 14], driven particularly by the always-listening nature of devices, along with concerns about how speech data is used and stored and who has access to it [4].

2.2 Non-native speakers' use of IPAs

Although languages can be changed and added on a number of popular IPAs, coverage and functionality varies across devices and assistants (e.g.[28]). This forces many users to have to speak to IPAs in a language other than their first language. The amount of work on IPA use among L2 speakers is limited, with research being preliminary and questionnaire based in nature [35], or focused on L2 language learning technology experience [2, 18]. Recent quantitative research suggests that L2 speakers find smart speakers harder to use [35, 36] and more difficult to interact with effectively than L1 speakers [35, 36], with language proficiency being related to more positive experience ratings [36]. Although they enjoy the interaction, L2 speakers feel they have to expend considerable effort when planning their utterances [36]. Rephrasing commands also leads L2 speakers to become frustrated [36]. IPA use has also been explored as a tool to help L2 users improve language skills [31] as they afford L2 speakers an opportunity to practise listening to speech output and produce speech input in a stress free context [30]. This work has noted that, although they may not be perceived as such [14], IPAs are adept at recognising accented speech accurately in these contexts, whilst providing a useful tool for L2 language learners [31]. Work has also observed how non-native speech output, particularly the matching of accents with non-native users, can significantly impact user perceptions and behaviour. Non native speakers were more dissatisfied with speech outputs when they used accents that varied from their own [16] whilst non-native speakers also failed to respond to information provided by spoken navigation systems in accents that were dissimilar to their own [26].

2.3 Non-native speaker interaction experiences

Non-native speaker's interaction experiences, such as their experiences of web page readability and internet search, have also been studied, revealing more general HCI difficulties L2 speakers may face. Web page readability studies have shown that vocabulary retrieval and parsing of complex grammatical structures are two major difficulties for L2 speakers when reading English pages [41, 42]. Likewise, a study of online search found L2 users to have particular difficulties in query formation due to both vocabulary limitations and grammatical phrasing. This work also found that repair strategies such as rephrasing are much more difficult for L2 users and add to technology-related stress [7]. While these difficulties are found in non-native use of written language in technology use, we expect that L2 users may experience similar challenges in speech based interactions with technology.

3 Research aims

Most IPA user experience research has focused on identifying issues in L1 speaker interactions. Recent work on L2 speakers has used more quantitative approaches to explore their experiences. Our works builds on this through using qualitative techniques to more deeply investigate the experiences of L2 speakers. To this end, our paper presents research that aims to identify important issues in L2 IPA user experience, emphasising these issues through a contrast with L1 speaker experiences. In particular, we aim to explore how characteristics of the two most popular IPA device platforms, namely: smartphones and smart speakers [39] influence these experiences. Our work is the first to directly compare L1 and L2 users' experiences of an off-the-shelf IPA across these device platforms. Through a deepened understanding of challenges faced by L2 users, we hope to inform IPA design toward more inclusive interactions.

4 Method

4.1 Participants

Thirty three participants (F=14, M=18, Prefer not to say=1; Mean age=28.1 yrs; SD=9.8 yrs) were recruited from a European university via email, posters and flyers displayed across campus, and through

snowball sampling. One participant was removed from the sample due to technical issues in their experiment session. Of the remaining thirty two participants, 16 (F=8, M=7, Prefer not to say=1) were native English speakers, with English as their first language, and 16 (F=6, M=10) were native Mandarin speakers who were non-native English speakers. On a 7 point Likert Scale (1= Not at all proficient; 7 = Extremely proficient) our sample of 16 Mandarin speakers rated their English proficiency as moderate (Mean=4.21, SD=0.7). Across the sample 78.1% (N=25) had used IPAs before, with 9.4% (N=3) reporting frequent or very frequent use. Among participants that reported previous experience with IPAs, 13 were native Mandarin speakers and 12 were native English speakers, with Siri (56%) being most commonly used, followed by Amazon Alexa (36%) and Google Assistant (12%).

4.2 Device conditions

In the experiment, participants interacted with Google Assistant, through both a Moto G6 smartphone (*Smartphone* condition) and a Google Home Mini smart speaker (*Smart speaker* condition) using a within participants design. The order in which these were experienced was counterbalanced across L1 and L2 speaker groups. Using Google Assistant across both conditions ensured that the devices were the only source of difference. Google Assistant was selected because it is commonly used across both device types [32].

4.3 Task

Participants were asked to conduct a total of 12 tasks with Google Assistant (six per device - all tasks are included in supplementary material). Based on research identifying the most common tasks people conduct with IPAs [4, 21] experimental tasks included 1) playing music, 2) setting an alarm, 3) converting values, 4) asking for the time in a particular location, 5) controlling device volume and 6) requesting weather information. Two versions of each task were generated creating two sets of six tasks. Each set of 6-tasks were used in only one of the device conditions. All tasks were delivered to participants as pictograms (see Figure 1). This was so as to eliminate the potential influence of written task instructions on what both L1 and L2 participants might say to the IPAs, to more closely simulate natural query generation and to reduce potential difficulties translating task text for L2 speakers. The task sets were counterbalanced across device and speaker conditions and task order was randomised within sets for each participant.

4.4 Post interaction interview

After interacting with both devices, participants took part in a semi-structured interview. These interviews lasted approximately 20 minutes and focused on 3 key topics: 1) general views towards IPAs; 2) experiences with the IPAs in the experiment; and 3) reflections on how they spoke to each system. Participants were also asked to identify which of the devices they preferred and explain their preference. So as to ensure that there were no linguistic barriers when expressing their opinions, L2 speaker interviews were conducted in Mandarin. All data was audio recorded and transcribed, with the Mandarin interviews being translated back into English by a Mandarin speaking member of the research team. The interview data was then analysed using inductive thematic analysis [6]. Initial coding was conducted independently by two of the research team with experience in qualitative data analysis using thematic approaches. After initial codes and themes were generated, these were discussed by both coders and further refined based on these discussions. The results of this analysis are presented in the Results section below.

4.5 Procedure

The research received ethical approval through the University's low risk project ethics procedures [HS-E-19-127]. English and Mandarin speakers were recruited from staff and students at a European University via email, posters and through snowball sampling. Upon arriving at the lab, participants were fully briefed about the nature of the study before being asked to provide written consent. Next, they were asked to complete a demographic questionnaire, giving information about their age, sex, profession, nationality, native language and English proficiency. Both groups were asked to give details regarding whether they had used IPAs before, and if so how frequently. They were also asked to select which they had used most frequently from a list of IPAs (Siri, Alexa, Google Assistant, Microsoft Cortana, Samsung Bixby or other).

Participants were then introduced to the task pictograms, before interacting with Google Assistant. In order to ensure participants did not encounter significant difficulties interpreting the pictograms, they were first presented with a trial set of images (on paper) and asked to write down what they would say to the agent to complete each task. These were similar in topic and layout to the pictograms used during their smartphone and smart speaker Google Assistant interactions, but varied in the information being requested (e.g. varying the city for which the time is requested). Participants' interpretations were checked by the experimenter prior to interacting with the devices.

After completing this, participants were presented with a set of six pictograms on a laptop, one at a time, and asked to complete the task represented in the image using Google Assistant on one of the devices. Participants self-reported the completion of tasks when they either thought they had accomplished the goal of the task or felt stuck and unable to complete the task. Tasks were deemed complete by participants rather than experimenters in order to avoid influencing interaction strategies. Once participants felt they had completed a task, they were asked to confirm by clicking a checkbox on the screen. By clicking a button, participants then revealed the next task. This process was then repeated until all six tasks were completed for that device. The participant was then asked to complete a further six tasks with Google Assistant using the next device. The tasks were delivered using the same process. An experimenter was present, only to ensure the tasks were being engaged with and to help with any technical issues.

After finishing the tasks with both smartphone and smart speaker devices, participants then completed a short post-interaction interview. This was recorded using a Blue Yeti microphone and audio capture software (Audacity v. 2.3.0). After completing the interview, participants were fully debriefed as to the aims of the study, uses of the data they had provided, and were given contact information for any further questions. They were then thanked for

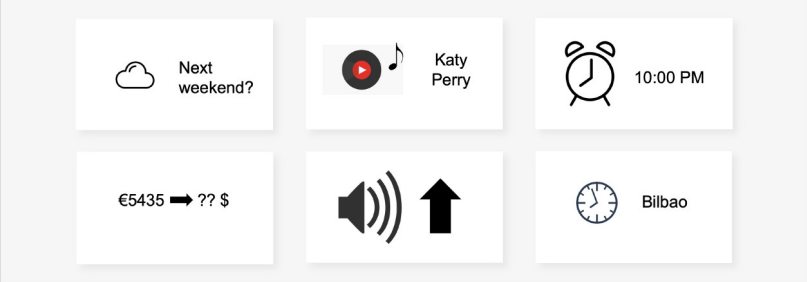


Figure 1: Example set of task pictograms

participation and given a €10 voucher as an honorarium. The study lasted approximately 40 minutes.

5 Results

5.1 General speaker differences

Irrespective of device type, clear differences emerged between L1 and L2 speakers' experiences when using IPAs. These revolved around how each group approached the interaction, issues in turn taking, and the desire for design to reduce the effort involved in language production.

5.1.1 Interaction approaches: Echoing previous literature on language production in speech interface [27] and IPA [14, 29, 34] use, L1 speakers prioritised vocal clarity, using correct English, brevity and planning when approaching interaction with both IPAs:

"I suppose it's breaking it down and thinking about what the question actually is. I suppose what it is you want to know, how you should ask that if you were using proper English." [P17-L1]

"[A]rticulate clearly and you know, think about how was, what was the simplest way possible to ask the question you know." [P4-L1]

L2 speakers also heavily emphasised adaptations aimed at increasing the likelihood of being understood, but this was due to their own perceived language or speech limitations. For instance, rather than adapting vocabulary to try and improve system performance (as is commonly mentioned in previous work [14, 29]) L2 speakers instead altered their vocabulary based on whether or not they knew a particular word:

"I tend to change the words I used. For example, when I asked the devices to change the volume, I didn't know the word 'volume' so I changed it to 'voice' or 'sound'." [P9-L2]

L2 speakers tried to adopt strategies to overcome this, especially when the IPAs repeatedly did not recognise their utterances, although these were unsuccessful:

"The recognition of people's names and place names has a low success rate. And when I don't know a word, I would spell it. But the devices cannot understand that." [P9-L2]

L2 speakers were also highly sensitive to their pronunciation or need to retrieve the correct words in interaction, which took up considerable time when interacting with the IPAs:

"... sometimes it cannot understand some pronunciation by non-native English speakers, and it cannot help you pronounce the words you don't know." [P9-L2]

"As a non-native English speaker, I have a hard time on proper nouns ... Maybe the devices cannot recognize due to this issue, and it may waste a lot of time." [P4-L2]

5.1.2 Waking and turn taking: There were also clear differences between L1 and L2 speakers when it came to *waking* the device and managing turn taking with the IPAs. L2 speakers regularly felt like they struggled to wake the Assistants in both device conditions:

"I feel that I wake up a few times and it ignored me ... like you have finished the commands but it maybe didn't get you." [P14-L2]

During interactions, L2 speakers suggested they sometimes needed extra time to formulate an utterance and that this was not taken into consideration by the system. Subsequently, the system would either reset or barge in before they had finished their request:

"I need some time to think ... if I think for a long time, I have to wake up the machine again. The command I said before is wasted." [P4-L2]

"[W]hen I didn't finish my sentences and the machine 'thinks' that I'm finished. It's like, the IPAs can only analyze the sentences you said." [P14-L2]

In contrast, L1 speakers perceived the delay between speaking to the device and it responding as too long, making the interaction seem slow:

"[A]gain it would be the delayed interaction. What I think is going on there is it is trying to figure out whether or not you finished a sentence?" [P9-L1]

"It's too slow to react, because it can't keep up with me. I'd much rather type..." [P1-L1]

This led L1 speakers to question whether the devices were working correctly, increasing frustration:

"[T]here were times when I wasn't sure if you know it was just not working or something? Because I really clearly said what I wanted it to do and it didn't, you know, react." [P4-L1]

"It's more frustrating I suppose, the lack of proper response to it, there was no, 'don't understand the questions' ... the feedback didn't seem adequate." [P17-L1]

5.1.3 Reducing the burden of production: L2 speakers also tended to emphasise ways that IPAs could be improved, with a focus on reducing the level of language production needed as well as ways to support word recall and gaps in lexical knowledge. Frustration with having to reproduce or reformulate queries from scratch was

Table 1: Frequency of participant device preferences

Group	Smart Speaker	Smartphone	No Preference
L1	12	4	0
L2	4	10	2

common, with L2 speakers suggesting that it would be helpful if IPAs were aware of previous attempts to make a query:

"It's, maybe I can meet some words problems. We are not native speakers after all. So I hope that it can, you, know… recognize this part more intelligently. For example, I use some simple words to describe my commands and the devices can understand the meaning. The devices can try to figure out my requirements. That will be better." [P7-L2]

"After I finished the question, for example, when I asked the time of a city, and got an answer for a wrong city, then there is no need to ask the whole question again. I should only need to emphasize the city name." [P1-L2]

Similarly, to support L2 issues with word selection, participants suggested that contextual options could be provided in cases where the intent of the query was recognised but users were struggling with the specific noun required:

"When I don't know that word I can say 'please transfer Celsius degree to another temperature unit'...for example, they can list all temperatures for you." [P8-L2]

"For example, if you say some unclear words and the devices can show you the possible choices or match the most possible option or something." [P16-L2]

This desire to support the interaction with suggestions was not identified by L1 users, who instead suggested better recognition of wider forms of language such as colloquialisms:

"[T]hey're interesting I suppose in how they try to make you speak in a different way that's not natural to you. They make your colloquialism sound strange and they make you pronounce things in that curious kind of way." [P17-L1]

"Improved? I'm not too sure, I suppose it involves listening to more conversations and getting a colloquial idea than just proper English." [P5-L1]

5.2 Device-specific differences

Our analysis discovered a marked difference in the way that both speaker types experienced IPA use through the smart speaker and smartphone respectively. There was a clear difference in preference between the speaker groups, driven by the benefits of visual feedback and output for supporting the interaction.

5.2.1 L1 and L2 device preference: A Chi-squared test revealed a statistically significant difference in device preference between L1 and L2 speakers [$\chi^2(1, N = 30) = 4.74, p < 0.05$]. Among the L1 speakers, 75% (N=12) preferred using Google Assistant on the smart speaker, whilst 63% of L2 speakers (N=10) preferred using the Assistant on the smartphone. Two participants reported having no preference. Preference frequencies are shown in Table 1.

5.2.2 Visual confirmation: Both L1 and L2 speakers commented on the role of visual feedback in IPA interactions using a smartphone. However, L2 speakers placed much greater emphasis on the benefits of this feature in allowing them to build confidence in the efficacy of the system's speech recognition capabilities, and in supporting them to identify exact reasons for miscommunication. Having visual feedback on the smartphone when using the Assistant also reduced the burden of having to interpret, translate and retain information for L2 speakers, which was an issue when responses and feedback were solely delivered using speech (i.e. when using the smart speaker). As shown in the *General Speaker Differences* section, it was common for L2 speakers to doubt whether they had pronounced words correctly. The on-screen feedback with the smartphone allowed them to identify exactly what the system did and did not recognise:

"When I interacted with the second one [smartphone], I feel good. Because sometimes I know that my pronunciation is not accurate. But it can recognize the words accurately." [P7-L2]

"I feel I didn't pronounce accurately for some location names. However, the interface can recognize my pronunciation correctly." [P6-L2]

In addition to building user confidence in the system's recognition capabilities, this transparency also allowed non-native speakers to localise specific lexical items that were the cause of miscommunication:

"However, for the smartphone, for example, I don't know how to ask the question for temperature. Then I asked 'How to describe the temperature in two systems' and the system replied to me. Although I still cannot pronounce those two words, I can receive some information about that. The smartphone can show the text on the screen." [P9-L2]

In line with this observation, L2 speakers noted having the opposite experience with the smart-speaker:

"For the first interface [smart speaker], I even have no idea about where I got it wrong. It's like...maybe you have misarticulation, however, you never know which words you have the pronunciation issues with'.' [P3-L2]

5.2.3 Visual output: In addition to transcriptions of the speech being recognised by the IPA, L2 speakers also benefited from supplementary information, displayed on-screen in the smartphone condition, in response to queries. Here, the pairing of speech and visual output from the system was found to alleviate difficulties in interpreting audio feedback whilst simultaneously trying to retain the information given by the system:

"[A]s a non-native English speaker, reading is much easier than listening. That device [smartphone] has a screen for people to read the text. I can gather more information using that." [P4-L2]

"If I didn't pay attention on hearing the answers I can check the details on the screen." [P10-L2]

Whilst the visual output provided in the smartphone condition was viewed positively by a number of L2 speakers, it was often negatively perceived by L1 speakers. These users stated that it reduced the usefulness of the experience when visual output was provided as an alternative to speech-based responses:

"[I]t brought up a list of web resources and I thought I can do that myself, you know. I expected it to give me a response rather than leave me to look at sites where I can get the information myself" [P14-L1]

"I suppose, it's easy enough to check things on your phone anyways so like, I don't feel I need to say it. Like I don't know." [P2-L1]

6 Discussion

IPAs are useful, especially in facilitating interaction in hands busy/eyes busy contexts [14, 29]. Although language coverage has recently been broadened to accommodate non-English speaking users, functionality and device coverage across these languages is highly variable (e.g. [28]), excluding these users from the full benefits of IPA use. Building on recent preliminary questionnaire-based work [35, 36], our research, the first to compare IPA experiences across devices for L1 and L2 users, identifies in detail the key aspects of L2 user experience, how this compares to L1 users, and whether this varies across smartphone and smart speaker based IPA interactions. Through thematic analysis, we identified a number key themes that highlight differences between L1 and L2 users. We found clear differences in the approaches taken to language production when generally interacting with the IPAs. L1 speakers perceived themselves to focus on structuring their commands as succinctly and as simply as possible. Although L2 speakers also tended to plan and adapt their natural speech, they focused more on how this was driven by perceived language limitations, in particular pronunciation, lexical knowledge and retrieval. Poor IPA waking and turn taking was also a problem experienced by L2 speakers. They felt that Google Assistant was not sensitive to the time they needed to produce their utterances, meaning they regularly experienced system barge-in. In contrast, L1 speakers felt that the time between production and recognition was too long, disrupting the interaction. When considering ways the interaction with IPAs might be improved, L2 speakers consistently emphasised approaches that would reduce the need for language production or support lexical retrieval. This was not a concern aired by L1 speakers. Our work adds richer detail to our current understanding of the difficulties faced by L2 users in IPA interactions [35, 36] along with a deeper understanding of differing potential causes of difficulty and frustration faced by L1 and L2 users respectively.

Our research also discovered a clear difference in device preference between L1 and L2 speakers, with L2 speakers preferring smartphone based IPA interactions. This was in contrast to L1 speakers, who preferred using the smart speaker. In particular, L2 speakers found the visual feedback provided through the smartphone critical in supporting their interaction, helping them diagnose reasons for interaction breaking down, whilst also giving them confidence in the system's ability to understand their utterances. L1 speakers, on the other hand, felt that audio-only feedback was sufficient. This also extends to system output, whereby L2 speakers benefited from the display of query results through text or through onscreen information, with the L1 speakers highlighting this as unnecessary. Thus, our work highlights how accessing IPAs through screen-based rather than speech-only devices, can be useful in supporting L2 speakers; reducing both the level of effort required for successful interaction and the level of frustration they experience. We explore the reasons and design implications of these findings below.

6.1 Designing IPAs to be sensitive to L1 and L2 interaction differences

Our findings highlight some important differences between how L1 and L2 speakers interact with IPAs. L1 speakers emphasised the importance of succinct and short utterances. This supports previous observations of user behaviour, whereby users tend to simplify or adapt their language choices when interacting with speech interfaces [3, 27, 29] to increase the likelihood of interaction success [5, 13, 33]. This is thought to be driven by users seeing speech interfaces as *at risk listeners* [33] or poor interlocutors [5], whereby adaptation to perceived system limitations is required to ensure interaction success. Although L2 speakers may also perceive speech systems in the same way, L2 speakers seemed to more heavily place the burden of potential interaction failure on themselves, seeing their pronunciation and lack of linguistic knowledge as significant barriers. This should be considered when designing IPA interactions. L2 users may need to be given more time to produce utterances, and more opportunities to clarify perceived misinterpretations of speech output, requiring multiple turns to repair and negotiate miscommunications [25]. These are not currently afforded by the one size fits all approach of current IPAs. Future design of IPAs should look to tailor the experience if the system identifies a user as a non-native speaker. Without these changes, L2 speakers may be at risk of abandoning IPA use more readily. Recent work focusing on language learning contexts, suggests that abandonment, along with direct repetition and rephrasing of queries, are common L2 strategies when faced with miscommunication with IPAs [31]. Future research should more deeply explore ways to tailor the IPA experience based on this, and how that may influence long term IPA engagement.

6.2 Screens are integral for L2 IPA users

Screen-based feedback was clearly important in supporting L2 users' IPA experience. For instance, speech recognition transcriptions displayed on screen were found to play a role in developing L2 speakers' confidence in the system's recognition capabilities, whilst also helping them to diagnose specific reasons for communication breakdown. Using the screen to support speech output from the system with supplementary information, such as links to websites or maps, was also seen as a non-trivial benefit among L2 users. Previous work illuminates potential reasons for this. L2 speakers find non-native synthesis less intelligible than L1 speakers [1, 37], particularly in noisy environments [37]. The interpretation of non-native speech also significantly increases cognitive load for L2 interlocutors during dialogue interactions [19, 38] as linguistic dimensions (e.g. sound system and common linguistic structures) may vary significantly from their native language [40]. Although reading non-native language text is also likely to increase cognitive load, the permanence of supplementary visual feedback [31], may give L2 users extra time to process and comprehend the information and refer back to it at a later time, improving their experience. This is not possible with speech only smart speakers. Research on the cognitive implications of screens in voice user interface (VUI)

design is needed to support this further, yet from our study it is clear that incorporating well designed screen based feedback is a design imperative for improving L2 speaker interaction with IPAs. Future studies should build on this, by concentrating on identifying specific ways that this screen based feedback can be improved to further support L2 speakers.

6.3 The need to relieve the burden of production

L2 speakers also regularly mentioned difficulties in producing what they felt were the correct words, or pronunciations of words, to interact effectively with the IPA. Suggestions for IPA improvements echoed these difficulties. For instance, it was felt that systems should allow the user to correct a single lexical item - the root cause of a miscommunication - whilst preserving the context of a query, or that they should list a small range of options associated with specific common tasks. This was so as to save L2 users the effort of formulating multiple queries for a task. Again, this is likely to be due to the additional difficulties and increase in cognitive load associated with retrieving appropriate lexical items across multiple languages. This retrieval difficulty is because, compared to monolinguals, bilinguals experience less frequent word activation through processing and production [24], making the word needed at a specific moment hard to access compared to monolingual users [38]. This makes retrieval for production particularly cognitively taxing [19], with our L2 participants suggesting that the IPA needs to be sensitive to the time this takes. Literature within psycholinguistics and HCI also suggests other ways in which IPAs could be designed to reduce this production burden. For instance, system speech output could be used to increase activation of relevant, in-domain lexicon and syntax, by using priming [5, 11, 12]; essentially priming keywords, nouns and structures the system can effectively manage. This could be facilitated through speech only and multimodal interactions and may be particularly useful in error management, assisting L2 speaking users in reformulating queries. Indeed, such techniques may reduce language production load for all users, and thus lead to significant user experience benefits across L1 and L2 users. Future work needs to focus more specifically on how these techniques may influence user experience.

6.4 Limitations

This work examines how L2 speakers of English interact with IPAs, highlighting how this compares to L1 English speakers. To ensure that we could conduct interviews in L2 speakers' first language, all of the L2 participants in the present work were native speakers of Mandarin. These participants were students of University College Dublin living in a country where English is a primary language and thus are more frequently exposed to English than other L2 English speakers might be. Although this may limit generalisability, our findings may be conservative, as people with less frequent exposure to English would likely have even greater difficulty in interacting effectively in IPAs. Further, the level of dissimilarity between Mandarin and English may also impact the generalisability of our findings. English and Mandarin vary significantly on a number of dimensions (e.g. relative importance of structure and tone in defining meaning, frequency of abstract and concrete nouns [17]). L2 speakers with a more similar first language to English (e.g. native Germanic language speakers) may vary in their experiences compared to the L2 users in this study. Future work should look to explore this in more detail. In addition, our work uses Google Assistant for both interactions. Future work in this topic should look to include a wider and more diverse selection of L2 speakers, as well as explore L2 experiences with other assistants and contexts.

To compare participant experiences with smartphone and smart speaker based systems, all participants interacted with both in each experiment. Through these interactions they completed two sets of six tasks, with similarities in the way they were graphically presented. Participants' experiences with the IPA through the first device may of course influence their interaction with the second device, as they may use similar commands or alter commands based on their initial experience. They may also be less hesitant in engaging with the IPA and completing the tasks second time around. To reduce the potential for practice effects we ensured that the task sets were randomised and were counterbalanced across the experiment conditions. Likewise, device conditions were also counterbalanced within groups.

Rather than using text or audio to deliver the tasks, we constructed a set of 12 pictograms, with two versions of each task. This was specifically used to reduce the likelihood of participants using the exact wording used in task instructions had the instructions been written or spoken. Using pictograms, although adding another cognitive demand on users through the need to interpret the image, was designed to encourage users to generate queries in a way that is more in-keeping with 'real-world' interactions. The use of text in this experiment might also have unduly increased the cognitive load of L2 speakers as they would have to translate the task text and then generate the query. The pictogram method used meant they could interpret the images without the need for task translation. So as to ensure that these were interpreted accurately, we also made participants report how they would word a query to the IPA for that particular image, with the experimenter ensuring that these were accurately interpreted and any issues in interpretation could be clarified before commencing the experiment. Due to the nature of the participants in this type of work, it is important for future work to consider the nature of task delivery and the potential interference that using linguistic means to communicate tasks may have on the findings of future research.

7 Conclusion

IPAs, although accessible to native English speakers, are not universally accessible in all languages. Language coverage varies by device and/or assistant. This means it may not be feasible for all users to interact in their native tongue. Our study has outlined key themes related to how L1 and L2 speakers vary in their user experience, and how aspects of IPAs may benefit or impede L2 users. We find that using IPAs through smartphones, which afford visual feedback to support the user, are significantly preferred by L2 users. We also identified important differences in the strategies L2 and L1 users had when planning their utterances to ensure communication success, with L2 users looking to identify ways to reduce their levels of language production in interaction. Importantly, by comparing L1 and L2 users, the work highlights specific areas that

may be leveraged to support L2 speakers in future IPA use. It also demonstrates the importance of expanding the types of users being researched to ensure that IPAs are designed to be more inclusive and accessible to a wider audience.

Acknowledgments

This work was conducted with the financial support of the UCD China Scholarship Council (CSC) Scheme grant No. 201908300016, Science Foundation Ireland ADAPT Centre under Grant No. 13/RC/2106 and the Science Foundation Ireland Centre for Research Training in Digitally-Enhanced Reality (D-REAL) under Grant No. 18/CRT/6224.

References

[1] Diane Mayasari Alamsaputra, Kathryn J. Kohnert, Benjamin Munson, and Joe Reichle. 2006. Synthesized speech intelligibility among native speakers and non-native speakers of English. *Augmentative and Alternative Communication* 22, 4 (2006), 258–268. https://doi.org/10.1080/00498250600718555

[2] Minoo Alemi, Ali Meghdari, and Maryam Ghazisaedy. 2015. The impact of social robotics on L2 learners' anxiety and attitude in English vocabulary acquisition. *International Journal of Social Robotics* 7, 4 (2015), 523–535.

[3] René Amalberti, Noëlle Carbonell, and Pierre Falzon. 1993. User representations of computer systems in human-computer speech interaction. *International Journal of Man-Machine Studies* 38, 4 (April 1993), 547–566. https://doi.org/10.1006/imms.1993.1026

[4] 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). https://doi.org/10.1145/3311956

[5] Holly P. Branigan, Martin J. Pickering, Jamie Pearson, Janet F. McLean, and Ash Brown. 2011. The role of beliefs in lexical alignment: Evidence from dialogs with humans and computers. *Cognition* 121, 1 (Oct. 2011), 41–57. https://doi.org/10.1016/j.cognition.2011.05.011

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Peng Chu, Anita Komlodi, and Gyöngyi Rózsa. 2015. Online search in english as a non-native language. *Proceedings of the Association for Information Science and Technology* 52, 1 (2015), 1–9.

[8] Leigh Clark, Philip Doyle, Diego Garaialde, Emer Gilmartin, Stephan Schlögl, Jens Edlund, Matthew Aylett, João Cabral, Cosmin Munteanu, Justin Edwards, and Benjamin R Cowan. 2019. The State of Speech in HCI: Trends, Themes and Challenges. *Interacting with Computers* (2019). https://doi.org/10.1093/iwc/iwz016

[9] Leigh Clark, Abdulmalik Ofemile, and Benjamin R. Cowan. In press. Exploring verbal uncanny valley effects with vague language in computer speech. In *Voice Attractiveness: Studies on Sexy, Likable, and Charismatic Speakers*, Melissa Barkat-Defradas, Benjamin Weiss, Jürgen Trouvain, and John J. Ohala (Eds.). Springer.

[10] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Christine Murad, Cosmin Munteanu, Vincent Wade, and Benjamin R. Cowan. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. *arXiv:1901.06525 [cs]* (Jan. 2019). https://doi.org/10.1145/3290605.3300705 arXiv: 1901.06525.

[11] Benjamin R Cowan and Holly P Branigan. 2015. Does voice anthropomorphism affect lexical alignment in speech-based human-computer dialogue?. In *Sixteenth Annual Conference of the International Speech Communication Association*. 155–159.

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

[13] Benjamin R Cowan, Philip Doyle, Justin Edwards, Diego Garaialde, Ali Hayes-Brady, Holly P Branigan, João Cabral, and Leigh Clark. 2019. What's in an accent? The impact of accented synthetic speech on lexical choice in human-machine dialogue. In *Proceedings of the 1st International Conference on Conversational User Interfaces*. 1–8.

[14] Benjamin R Cowan, Nadia Pantidi, David Coyle, Kellie Morrissey, Peter Clarke, Sara Al-Shehri, David Earley, and Natasha Bandeira. 2017. What can i help you with?: infrequent users' experiences of intelligent personal assistants. In *Proceedings of the 19th International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 43.

[15] Anna L. Cox, Paul A. Cairns, Alison Walton, and Sasha Lee. 2008. Tlk or txt? Using voice input for SMS composition. *Personal and Ubiquitous Computing* 12, 8 (2008), 567–588. https://doi.org/10.1007/s00779-007-0178-8

[16] Nils Dahlbäck, Seema Swamy, Clifford Nass, Fredrik Arvidsson, and Jörgen Skågeby. 2001. Spoken Interaction with Computers in a Native or Non-native Language-Same or Different. In *Proceedings of INTERACT*. 294–301.

[17] Antonella Devescovi and Simonetta D'Amico. 2004. The competition model: Crosslinguistic studies of online processing. In *Beyond Nature-Nurture*. Psychology Press, 215–242.

[18] Gilbert Dizon. 2017. Using intelligent personal assistants for second language learning: a case study of Alexa. *TESOL Journal* 8, 4 (2017), 811–830.

[19] Zoltán Dörnyei and Judit Kormos. 1998. Problem-solving mechanisms in L2 communication: A psycholinguistic perspective. *Studies in second language acquisition* 20, 3 (1998), 349–385.

[20] 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 (MobileHCI '19)*. Association for Computing Machinery, New York, NY, USA, Article 5, 12 pages. https://doi.org/10.1145/3338286.3340116

[21] Mateusz Dubiel, Martin Halvey, and Leif Azzopardi. 2018. A Survey Investigating Usage of Virtual Personal Assistants. *CoRR* abs/1807.04606 (2018). arXiv:1807.04606 http://arxiv.org/abs/1807.04606

[22] Justin Edwards, He Liu, Tianyu Zhou, Sandy J. J. Gould, Leigh Clark, Philip Doyle, and Benjamin R. Cowan. 2019. Multitasking with Alexa: How Using Intelligent Personal Assistants Impacts Language-Based Primary Task Performance. In *Proceedings of the 1st International Conference on Conversational User Interfaces (CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 4, 7 pages. https://doi.org/10.1145/3342775.3342785

[23] Emer Gilmartin, Marine Collery, Ketong Su, Yuyun Huang, Christy Elias, Benjamin R. Cowan, and Nick Campbell. 2017. Social talk: making conversation with people and machine. In *Proceedings of the 1st ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents - ISIAA 2017*. ACM Press, Glasgow, UK, 31–32. https://doi.org/10.1145/3139491.3139494

[24] Tamar H Gollan and Lori-Ann R Acenas. 2004. What is a TOT? Cognate and translation effects on tip-of-the-tongue states in Spanish-English and tagalog-English bilinguals. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30, 1 (2004), 246.

[25] Barbara Hoekje. 1984. Processes of Repair in Non-Native-Speaker Conversation. (March 1984). https://eric.ed.gov/?id=ED250922

[26] Marie Jonsson and Nils Dahlbäck. 2011. I can't hear you? drivers interacting with male or female voices in native or non-native language. In *International Conference on Universal Access in Human-Computer Interaction*. Springer, 298–305.

[27] Alan Kennedy, A Wilkes, L Elder, and Wayne Murray. 1988. Dialogue with machines. *Cognition* 30 (1988), 37–72. https://doi.org/10.1016/0010-0277(88)90003-0

[28] Bret Kinsella. 2019. Google Assistant Now Supports Simplified Chinese on Android Smartphones. http://bit.ly/30Yg8qN. Accessed 27th Jan 2020.

[29] 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*. ACM, 5286–5297.

[30] Souheila Moussalli and Walcir Cardoso. 2016. Are commercial 'personal robots' ready for language learning? Focus on second language speech. *CALL communities and culture–short papers from EUROCALL* (2016), 325–329.

[31] Souheila Moussalli and Walcir Cardoso. 2019. Intelligent personal assistants: can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning* 0, 0 (2019), 1–26. https://doi.org/10.1080/09588221.2019.1595664

[32] Christie Olson and Kelli Kemery. 2019. *2019 Voice report: Consumer adoption of voice technology and digital assistants.* Technical Report. Microsoft.

[33] Sharon Oviatt, Jon Bernard, and Gina-Anne Levow. 1998. Linguistic Adaptations During Spoken and Multimodal Error Resolution. *Language and Speech* 41, 3-4 (July 1998), 419–442. https://doi.org/10.1177/002383099804100409

[34] 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*. ACM, 640.

[35] Aung Pyae and Paul Scifleet. 2018. Investigating Differences between Native English and Non-Native English Speakers in Interacting with a Voice User Interface: A Case of Google Home. In *Proceedings of the 30th Australian Conference on Computer-Human Interaction (OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 548–553. https://doi.org/10.1145/3292147.3292236

[36] Aung Pyae and Paul Scifleet. 2019. Investigating the Role of User's English Language Proficiency in Using a Voice User Interface: A Case of Google Home Smart Speaker. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3290607.3313038

[37] Mary Reynolds, ZS Bond, and Donald Fucci. 1996. Synthetic speech intelligibility: Comparison of native and non-native speakers of English. *Augmentative and Alternative Communication* 12, 1 (1996), 32–36.

[38] Norman Segalowitz and Jan Hulstijn. [n. d.]. Automaticity in bilingualism and second language learning. *Handbook of bilingualism: Psycholinguistic approaches* ([n. d.]), 371–388.
[39] Voicebot.ai. 2018. voice Assistant Consumer Adoption Report. http://bit.ly/2TZE0sL. Accessed 27th Jan 2020.
[40] Catherine Watson, Wei Liu, and Bruce MacDonald. 2013. The effect of age and native speaker status on synthetic speech intelligibility. In *Eighth ISCA Workshop on Speech Synthesis*.
[41] Chen-Hsiang Yu and Robert C Miller. 2010. Enhancing web page readability for non-native readers. In *Proceedings of the sIGCHI conference on human factors in computing systems*. 2523–2532.
[42] Chen-Hsiang Yu, Jennifer Thom-Santelli, and David Millen. 2011. Enhancing blog readability for non-native english readers in the enterprise. In *CHI'11 extended abstracts on human factors in computing systems*. 1765–1770.