Mental Workload and Language Production in Non-Native Speaker IPA Interaction

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

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

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

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

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

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

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

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

ABSTRACT

Through smartphones and smart speakers, intelligent personal assistants (IPAs) have made speech a common interaction modality. With linguistic coverage and varying functionality levels, many speakers engage with IPAs using a non-native language. This may impact mental workload and patterns of language production used by non-native speakers. We present a mixed-design experiment, where native (L1) and non-native (L2) English speakers completed tasks with IPAs via smartphones and smart speakers. We found significantly higher mental workload for L2 speakers in IPA interactions. Contrary to our hypotheses, we found no significant differences between L1 and L2 speakers in number of turns, lexical complexity, diversity, or lexical adaptation when encountering errors. These findings are discussed in relation to language production and processing load increases for L2 speakers in IPA interaction.

CCS CONCEPTS

• **Human-centered computing** → *User studies*; *Natural language interfaces*; *HCI theory, concepts and models.*

KEYWORDS

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

ACM Reference Format:
Yunhan Wu, Justin Edwards, Orla Cooney, Anna Bleakley, Philip R. Doyle, Leigh Clark, Daniel Rough, and Benjamin R. Cowan. 2020. Mental Workload and Language Production in Non-Native Speaker IPA Interaction. In *2nd Conference on Conversational User Interfaces (CUI '20), July 22–24, 2020, Bilbao, Spain.* ACM, New York, NY, USA, 8 pages.

CUI '20, July 22–24, 2020, Bilbao, Spain

ACM ISBN 978-1-4503-7544-3/20/07.
DOI: 10.1145/3405755.3406118

1 INTRODUCTION

Intelligent personal assistants (IPAs) like Google Assistant have increased the popularity of speech as an interaction modality [9]. Primarily used on smart speakers and smartphones [34], these assistants can be used in a number of different languages, but coverage and functionality across these languages is not comprehensive [27], requiring many users to interact using a non-native language. This includes those using English as a second language, hereby referred to as L2 speakers. Interacting with IPAs in this way is likely to be significantly more challenging than the interaction experienced by those using English as their native language (L1 speakers). For instance, L2 speakers tend to experience difficulty in lexical retrieval [21, 43], because of an incomplete knowledge of the language being used [47], with production being less automatized when compared to L1 users [15]. Alongside increased demands in processing and planning utterances in a second language, this means L2 users may experience a significantly higher mental workload [14, 47] when engaging with IPAs. These factors may also lead them to approach the interaction differently [40, 48]. Our research explores this empirically, by comparing the mental workload and language choices made by L1 and L2 speakers when interacting with IPAs across smart speakers and smartphones.

Our study identified significant differences in cognitive demand between the two speaker groups. Specifically, we found L2 speakers experience significantly higher levels of mental workload when interacting with IPAs in their non-native language compared to L1 speakers. Contrary to expectations, L1 and L2 speakers did not significantly vary in the number of commands needed to complete tasks, number of words used per command, the diversity of their lexicon, nor their levels of adaptation when they experienced errors during interaction. Our findings are the first to focus on the cognitive and linguistic aspects in L2 IPA use. We discuss the findings in relation to the cognitive mechanisms that may be present when interacting with IPAs as an L2 speaker.

2 RELATED WORK

2.1 Language production in speech interface interaction

Current work on language production in speech interface interaction almost universally observes the language choices of L1 speakers. Even then, the volume of work on this topic is limited [9], with a focus on comparing language production in interactions between human-machine and human-human interlocutors. Such work finds that users tend to vary significantly in how they interact with systems compared to how they interact with humans [2], although similar mechanisms may influence language production [11, 12]. People tend to use fewer topic shifts, use more words, as well as use fewer anaphora and fillers when interacting with computers as opposed to human partners. Similarly, people tend to use more basic lexical choices and grammatically simpler utterances [7] when interacting with computers compared to other people [26].

This tendency to vary speech choices based on partner type is thought to be driven by the perception of a computer's competence as a dialogue partner (i.e., a user's partner model), whereby people see voice user interfaces as at-risk listeners [36]. This is similar to the mechanisms for adaptation proposed in psycholinguistics literature, which highlight the tendency for partners to select their language with the perceptions of their audience in mind, termed *audience design* [6]. A similar effect has recently been shown to operate on lexical choice in speech interface interaction, whereby participants interacting with a US-accented system were significantly more likely to use US lexical terms than when interacting with an Irish-accented system [12].

2.2 L2 speakers and speech interfaces

Recent work comparing IPA use by both L1 and L2 speakers has focused on user experience as opposed to observing their interaction from a cognitive and linguistic perspective. L2 speakers see IPAs as more difficult to use than do L1 speakers [39, 40]. Recent work has also found that L2 speakers perceive difficulties in trying to use the right sentence structures or retrieving the right lexical terms [48] when speaking to IPAs, with L2 speakers feeling they have to rephrase utterances, causing frustration [40]. Research on L2 language production offers potential explanations for these perceived difficulties. It is widely acknowledged that L2 speakers tend to have an incomplete knowledge of the non-native language being used when compared to L1 speakers [15, 47]. Along with a comparative lack of automatisation of the cognitive processes for language production within a second language [15], this means L2 users must resort to specific production strategies to mitigate these production barriers. These include replacing lexical items, reducing message complexity or describing the meaning of words that are hard to retrieve [15]. Paired with the need to process non-native speech when in dialogue, this means L2 speakers experience considerable cognitive load when having to converse in a second language [14, 47].

Accented speech and the need for longer planning time may also lead to L2 users experiencing difficulties in commands being understood, with the system either not recognising commands or interrupting the user before commands are complete [25, 48]. When they encounter communication breakdowns in IPA use, L2 speakers tend to use common strategies to repair commands such as repeating and rephrasing utterances [33]. Yet, the effective planning of error repair may depend on the type of device being used. For example, L2 speakers have emphasised the benefit of using visual feedback [40], allowing them to use further visual information (e.g., transcriptions of the conversation) to diagnose errors in their commands as well as process system prompts, making them more effective when using IPAs [33, 48].

3 RESEARCH AIMS & HYPOTHESES

Although a number of users engage with IPAs in their non-native language, research on cognitive concepts such as the mental workload and the language they produce in interaction is scant. It is therefore critical that we widen research to include the experiences of non-native speakers [39, 40]. Our study focuses on linguistic and cognitive aspects of L2 speaker interaction. We focus on the mental workload experienced by L2 IPA users in comparison to L1 users, while also exploring the differences in language production between the two groups when completing tasks with an IPA.

We hypothesise that, due to planning, generating and processing speech utterances in a different language, L2 speakers are likely to experience significantly higher mental workload in IPA interaction compared to L1 speakers (H1). We also hypothesise that, due to speech recognition and planning time difficulties [48], L2 speakers may need significantly more turns when conducting a task than L1 speakers (H2). Due to lexical retrieval and knowledge constraints compared to L1 speakers, we also hypothesise L2 speakers will have significantly fewer words per utterance (H3), lower lexical diversity than L1 speakers in interaction (H4) and may vary in their levels of adaptation in comparison to L1 speakers when experiencing errors (H5).

Based on work emphasising the importance of visual modalities in supporting L2 speaker IPA use [40, 48], we also hypothesise that these effects may vary significantly by device. Specifically, the visual feedback afforded by Google Assistant on a smartphone may lead to reduced mental workload for L2 speakers due to visual output supporting error diagnosis and system query understanding (H6). As visual support helps users diagnose and correct errors, we also hypothesise that using a smartphone may significantly affect the number of commands per turn (H7) and the number of words per command (H8), while also impacting lexical diversity (H9) and levels of adaptation (H10) for L2 speakers.

4 METHOD

To investigate these hypotheses, we designed a study that enabled us to quantitatively compare the cognitive workload and linguistic properties of L1 and L2 speakers in their interaction with IPAs. The study received ethical approval through the university's ethics procedures for low risk projects.

4.1 Participants

A sample of 33 participants (F=14, M=18, Prefer not to say=1) with a mean age of 28.1 years (SD=9.8 years) took part in the study. These were all recruited from students and staff at a European university via email, campus-wide posters, and snowball sampling. One participant was removed due to technical difficulties in recording their utterances, leaving 32 participants in the sample. 16 (F=8, M=7, Prefer not to say=1) were native English speakers, and 16 (F=6, M=10) were native Mandarin speakers, who used English as their non-native language. These Mandarin speakers self-reported their English proficiency as moderate (7 point Likert Scale: 1 = Not at all proficient; 7 = Extremely proficient; Mean=4.21, SD=0.7). 78.1% (N=25) of our sample had used IPAs before, with 9.4% (N=3) using IPAs frequently or very frequently. For those that had used IPAs before, Siri (56%) was most commonly used, followed by Amazon Alexa (36%) and Google Assistant (12%). Each participant was given a €10 voucher as an honorarium for taking part.

4.2 Device type

The study included two device conditions. Participants interacted with Google Assistant, using both a Moto G6 smartphone (*Smartphone* condition) and a Google Home Mini smart speaker (*Smart speaker* condition) in a within-subjects design. We selected Google Assistant because it is commonly used on both smartphones and on smart speakers [34], minimising potential variation due to differences in the IPAs being used across devices. The order of device interaction was fully counterbalanced across L1 and L2 speaker groups.

4.3 Task

Participants used Google Assistant to complete a total of 12 tasks (6 with each device) across the experimental session. Experimental tasks focused on 6 common IPA tasks [3, 17]: 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. To reduce practice effects, two versions of each task were generated, creating two sets of six tasks. Each set of tasks was used in only one of the device conditions. To eliminate the influence of written tasks on user utterances, and the potential confound of written tasks increasing L2 speaker cognitive load, all tasks were delivered to participants as pictograms (see Figure 1 - all pictograms are included in supplementary material). The order of task sets were arbitrarily assigned, ensuring they were counterbalanced as much as possible across device and speaker conditions. Task order was randomised within sets for each participant.

4.4 Measures

4.4.1 Mental Workload: To assess participants' mental workload during interaction with each of the devices, participants completed the NASA-TLX [24] after completing each task set. The NASA-TLX is a 6-item Likert scale (20 point scale per item) questionnaire, measuring 6 constituent factors of mental workload: *Mental Demand, Physical Demand, Temporal Demand, Performance, Effort*, and *Frustration.* Scores on the questionnaire were summed to create an overall workload (Raw TLX) score (Range: 0-120, see [23]).

4.4.2 Language production in interaction: To assess language production in interaction, user task commands were transcribed. From these transcripts, a number of measures were derived. These measures include: *Number of commands per task*, *Lexical complexity*, *Lexical diversity per task*, *Dynamic lexical adaptation*, *Lexical adaption from initial command.*

Number of commands per task is defined as the number of utterances, starting with a wake phrase (i.e. "Hey Google" or "OK Google"), that a participant used to complete a task.

Lexical complexity (measured through word count per command) was derived by dividing the total word count used to complete a task by the number of turns taken. This measure represents the complexity of the utterance, and follows measures of L2 linguistic complexity used in text-based research [35]. As commands to speech interfaces tend to be concise, formulaic statements [19, 26], we used word count per command rather than measuring numbers of clauses as is done in other L2 complexity research [35].

Guiraud's index of lexical diversity [22] was also calculated to identify the number of unique words used when completing a task (*Lexical diversity per task*). This measure compares unique words in a command to the root of total words in a command. It is considered to be a robust alternative to diversity measures that use a direct ratio of unique words to total words, as these measures tend to inflate diversity as utterance lengths increase [45].

To gauge levels of lexical adaptation for tasks that required multiple utterances to complete, we measured the Guiraud index of lexical diversity for each pair of consecutive commands within a task (*Dynamic lexical adaptation*). We also measured the Guiraud index of lexical diversity for each utterance paired with the first utterance of a task to determine how much participants varied their lexical choices away from their initial command (*Lexical adaption from initial command*). Both measures of adaptation were used so that different styles of adaptation would be detected. For instance, participants may make a command, try a different phrasing, then return to their original phrasing. This would result in high dynamic lexical adaptation but low lexical adaptation from initial command. Participants may alternatively adapt by changing few words across many commands, resulting in low dynamic lexical adaptation but high lexical adaptation from initial command as each utterance increasingly departs from the first attempt. Using both measures allow us to detect these differences.

4.5 Procedure

Upon arrival, participants were welcomed by the experimenter, given an information sheet with details about the experiment and asked to give written consent to take part in the study. Participants then completed a demographic questionnaire, giving information about their age, sex, nationality, native language, experience with IPAs and speech interfaces, and their self-reported level of English proficiency. Participants were then given instructions for the study. Within these, they were asked to also look at 6 practice pictograms with the same visual structure as those in the experimental session but different in the information requested, and write what they would say to the IPA to complete the task depicted. From these responses, experimenters ensured they were interpreting the pictograms correctly before conducting the experimental tasks. They were then asked to complete a number of tasks with Google Assistant on two devices - a smartphone and a smart speaker. These tasks were displayed on a laptop, one at a time. Participants were asked to complete a task using the Assistant and once they felt they had done so, were asked to move to the next task. After completing

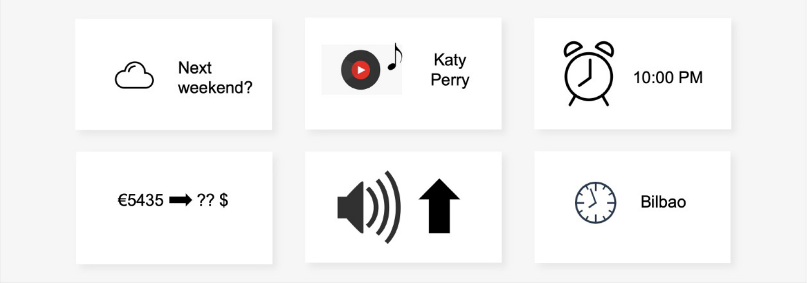

Figure 1: Example set of task pictograms

a set of 6 tasks with one of the devices, participants then completed the NASA-TLX. This was then repeated for the next 6 tasks, wherein they interacted with Google Assistant through the other device. After finishing all tasks with both devices, participants then completed a short post-interaction interview and were then fully debriefed as to the aims of the study, and thanked for participation. To capture participant utterances, the sessions were recorded using Audacity v. 2.3.0.

5 RESULTS

Out of the total of 384 tasks, 315 were successfully completed (82%) with 14 partially completed (3.6%) (i.e., participants completed the task but varied the information requested). 45 tasks (11.7%) were not successfully completed, of which 24 (6.2%) were not completed due to technical errors. Unsuccessful and technical error tasks were excluded from the dataset analysed. Before analysis, all data was screened for outliers, with these being replaced by values of the mean ± 3 SDs as suggested in [18]. Descriptive statistics for all measures included in the study are shown in Table 1.

5.1 Mental Workload

Due to violation of the assumption of normal distribution (p<.05), a robust mixed ANOVA with 10% trimmed means was run using the WRS2 package (Version 1.0) [30] in R (Version 3.6) [41]. There was a statistically significant main effect of speaker on the mental workload experienced, whereby L1 speakers reported significantly lower NASA-TLX scores (Mean=27.0; SD=19.07) than L2 speakers (Mean=42.0; SD=14.37) [Q=11.74, p=.002] (see Figure 2). This supports our first hypothesis (H1). However, there was no statistically significant main effect of device type [Q=0.28, p=.60] or interaction between speaker type and device type [Q=0.81, p=.37] on mental workload. H6 was therefore not supported.

5.2 Language production in interaction

5.2.1 Analysis Approach: To analyse the language production data, linear mixed-effects models (LMM) were run using the *lme4* package (Version 1.1.21) [5] in R (Version 3.6) [41]. This type of analysis allows for the modelling of fixed (i.e., device and speaker type) and random (i.e., participant and task variations) effects on specific outcomes such as lexical diversity. LMMs also allow us to model individual differences through random intercepts, as well as differences in how the fixed effects vary by participant and by task through modelling random slopes. We take the approach of modelling the maximal random effect structure determined by the experiment [4], reducing the complexity of random effects by removing higher order random slopes to facilitate convergence. We report LMM results in the text, following recent best-practice guidelines [31] by also reporting all LMM analyses fully. These appear in the supplementary material. We include fixed and random effect results as well as reporting all model syntax to improve model reproducibility.

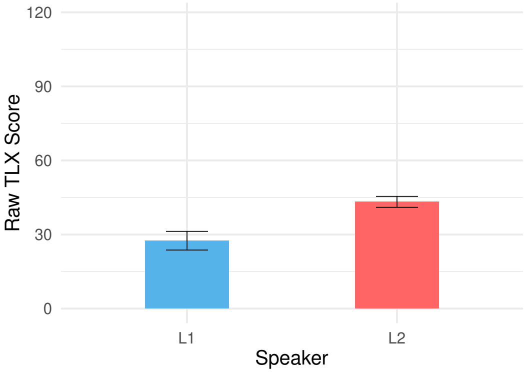

Figure 2: Mean Raw TLX scores (10% trimmed means with trimmed standard error) for each speaker group

5.2.2 Number of commands per task: Across the data set there was a total of 933 user commands. The LMM run showed no statistically significant effect of speaker [Unstandardized β=-0.39, SE β=0.37, 95% CI [-1.12,0.34], t=-1.06, p=.29], device [Unstandardized β=0.12, SE β=0.27, 95% CI [-0.41,0.63], t=0.43, p=.67] or speaker and device interaction [Unstandardized β=0.33, SE β=0.38, 95% CI [-0.41,1.07], t=0.88, p=.38] on the number of user commands per task. This means that our hypotheses (H2 and H7) were not statistically supported.

5.2.3 Lexical complexity: Across the dataset there were 7112 words used to command the IPAs, with an average of 7.62 words per command. There was no statistically significant effect of speaker [Unstandardized β=-0.65, SE β=0.59, 95% CI [-1.83,0.53], t=-1.11, p=.27], device [Unstandardized β=0.50, SE β=0.34, 95% CI [-0.17:1.18], t=1.45, p=.15] or speaker and device interaction [Unstandardized β=-0.45, SE β=0.49, 95% CI [-1.41,0.51], t=-0.92, p=.36] on the number of words used per command. Therefore our

Table 1: Descriptive statistics by speaker and device type

		NASA-TLX score (10% trimmed)		Number of commands per task		Lexical complexity		Lexical diversity per task		Dynamic lexical adaptation		Lexical adaptation from initial command	
Speaker	Device Type	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
L1	Smart speaker	27.36	18.59	2.24	2.18	8.08	2.87	2.61	0.53	2.12	0.67	0.98	0.96
	Smartphone	29.00	13.38	2.35	2.14	8.57	3.07	2.58	0.58	2.24	0.61	0.80	0.91
	Total	27.50	14.13	2.29	2.15	8.32	2.97	2.60	0.56	2.19	0.66	0.88	0.93
L2	Smart speaker	47.00	12.20	1.84	1.28	7.39	2.58	2.45	0.60	2.05	0.66	0.71	0.93
	Smartphone	40.64	8.69	2.30	2.02	7.43	2.11	2.55	0.53	1.96	0.71	0.88	0.89
	Total	43.23	8.11	2.07	1.70	7.42	2.55	2.50	0.57	2.00	0.68	0.80	0.91
Total	Smart speaker	36.89	16.16	2.04	1.79	7.74	2.74	2.53	0.57	2.09	0.66	0.85	0.95
	Smartphone	35.27	10.25	2.33	2.07	8.01	2.69	2.57	0.56	2.10	0.67	0.84	0.90

hypotheses in relation to lexical complexity (H3 and H8) were not statistically supported.

5.2.4 Lexical diversity per task: The LMM model showed no statistically significant effect of speaker type on levels of lexical diversity per task [Unstandardized β=-0.15, SE β=0.11, 95% CI [-0.38,0.07], t=-1.38, p=.18], speaker type [Unstandardized β=-0.01, SE β=0.08, 95% CI [-0.16,0.14] ,t=-0.19, p=.85] and speaker device interaction [Unstandardized β=0.12, SE β=0.11, 95% CI [-0.09,0.33] ,t=1.14, p=.26]. Therefore our hypotheses in relation to lexical diversity (H4 and H9) were not statistically supported.

5.2.5 Dynamic lexical adaptation: Over the 315 successful tasks, 116 required more than one command to complete. Tasks that participants only used one turn to complete (N=199) were excluded from the dataset. There was no statistically significant effect of speaker [Unstandardized β=-0.04, SE β=0.16,95% CI [-0.36,0.28], t=-0.28, p=.78], device [Unstandardized β=0.14, SE β=0.14, 95% CI [-0.14,0.42], t=0.98, p=.32] or speaker and device interaction [Unstandardized β=-0.24, SE β=0.20, 95% CI [-0.64,0.16], t=-1.20, p=.23] on the level of lexical diversity from a preceding turn. Therefore, L1 and L2 speakers did not vary in their levels of lexical adaptation from a previous utterance when having to use more than one command to complete a task. There was also no impact of device type on levels of lexical adaption from previous command, so H5 and H10 were not supported.

5.2.6 Lexical adaptation from initial command: Again, tasks where participants only used one utterance to complete the task were excluded from analysis. The LMM showed no statistically significant effect of speaker [Unstandardized β=-0.26, SE β=0.18, 95% CI [-0.61,0.10], t=-1.43, p=.16], device [Unstandardized β=-0.17, SE β=0.17, 95% CI [-0.51,0.17], t=-1.01, p=.32] or speaker and device interaction [Unstandardized β=0.33, SE β=0.25, 95% CI [-0.15,0.82], t=1.35, p=.18] on the level of lexical diversity from the first turn. It seems that both L1 and L2 speakers tend to use similar levels of lexical adaptation from their first turn, with this adaptation not being influenced by device type. This means that again H5 and H10 were not supported.

6 DISCUSSION

Our work set out to identify how using IPAs in a non-native language impacted mental workload and language production. We found L2 speakers experienced significantly higher mental workload than L1 speakers in IPA interactions across both smart speakers and smartphone devices. Although there were significant levels of workload for L2 users, there were no significant differences between L1 and L2 speakers in terms of the number of turns, words used and diversity of lexical choice. They also did not vary in the level of lexical adaptation from their initial utterances. They also did not vary in their level of lexical adaptation when comparing to a preceding turn. We discuss the interpretations of these findings below.

6.1 Linguistic retrieval, synthesis processing & workload

Our work highlights that, even though they may show similar types of language use, L2 speakers experience significantly higher mental workload than L1 users in IPA interaction. Reasons for this are likely to involve the increased load in producing and processing utterances in a non-native language [14, 15]. Efforts needed for lexical retrieval in production and processing may be of particular influence. Multilingual speakers store significantly more words in their mental lexicon when compared to monolinguals, to facilitate accurate word retrieval in processing and production when using other languages. This is thought to lead to less frequent access of words across their lexicon, making activation lower and thus leading to difficulties in recall and retrieving these lexical items [15, 21, 43]. The lack of automatisation of language production processes [15], is also likely to contribute to this load.

In addition to production issues, many L2 speakers also find it more cognitively challenging to process and understand non-native synthetic speech [46]. Non-native speakers find synthesis in a non-native language significantly less intelligible than do native speakers [1, 42, 46]. This is proposed to derive from L2 speakers' comparative unfamiliarity with their non-native language's phonological system, common syntactic structures and lexicon, which may increase cognitive load when interpreting and processing speech output [46]. In real world IPA use, this mental workload may be even higher as background noise negatively affects non-native speakers' ratings of intelligibility compared to native speakers [42]. A challenge for future HCI research is to investigate ways to mitigate this load for L2 users.

6.2 Lexical adaptation and limited potential for diversity

Contrary to our hypotheses, number of commands, lexical adaptation, complexity and diversity did not vary across speaker groups or device types. There may be a number of reasons for this. Although L2 users may experience more load in lexical retrieval, IPA interaction still tends to be lexically constrained. Consequently, complex and diverse lexical choices may not be a priority, as IPAs are often seen as basic dialogue partners [8, 10, 32]. This is contrasted by more open-ended interactions in which people have been shown to use conversational and complex linguistic structures (e.g., with automotive interfaces [28, 29]). L1 and L2 speaker variance may be more stark in these types of interaction.

The opportunity for lexical variation may be further limited by the requirement to use the wake word at the start of commands, reducing the potential for variability. Additionally, although adaptation has been noted as a common strategy for error repair in human-machine dialogue [26, 36], it may be that lexical adaptation in this instance is not the primary adaptation strategy for users. Although L2 users have suggested they may use lexical strategies in IPA use (e.g., substitution or describing the meaning of words they cannot retrieve) [48], adaptation of pronunciation is much more strongly emphasised by L2 speakers in previous work [40, 48]. L2 speakers tend to vary significantly from L1 speakers in other speech dimensions like tempo, rate of hesitations (e.g., filled pauses, repetitions and corrections [47]) while also adapting syntactically or semantically [37]. Our findings suggest that, at a lexical level, L2 speakers and L1 speakers do not vary in the limited lexical context of IPA interaction. Future work should look to explore other forms of adaptation as well as other linguistic cues in language production with IPAs across these user groups.

6.3 Proficiency and automaticity

Although we found no significant difference between speakers in lexical diversity and complexity, this may be due to proficiency of the participants recruited. L2 participants rated themselves as moderately proficient and all attended an English-speaking university. These factors, together with the relative simplicity of the commands required for IPA use, may explain the lack of effect in our analysis. Increased proficiency significantly improves IPA user experience for L2 language speakers [39, 40]. Increased fluency in a second language is also linked to the proceduralisation of syntactic and lexical knowledge of that language [44]. Although we found no effect in our sample, there may be differences between beginner and more advanced L2 speakers. Future work should look to identify the role that this proficiency has on language production within IPA interactions.

7 LIMITATIONS

Along with L2 users being recruited from a European university where English is the primary language, all L2 users were native Mandarin speakers, which may influence the wider generalisability of results to other native and non-native language combinations. It may be that cognitive effects seen in our work vary based on similarities and differences of the languages being used, such as the phonetic or structural similarity of a non-native language to participants' native tongue. This means that L2 speakers whose native languages are more closely related to English may experience even less evident language production effects than Mandarin speakers. It is therefore important that future work explores whether similar effects are seen for L2 speakers with different native languages, as well as differing levels of language ability mentioned above. It is also important to note that future work should look to increase sample size so as to identify whether the findings are replicated across larger samples of users.

To increase ecological validity, participants were able to control when to move on to the next task. This meant that participants could complete the tasks at their own pace and may more accurately reflect how many attempts participants are willing to give a task before abandoning it. Individual differences in this willingness are likely to influence the number of commands users made. Some were willing to try several times in order to successfully complete tasks, whereas others preferred to skip to the next task after relatively few attempts, even if they were not successful at completing the task (although we note only 5.5% of tasks in our data were abandoned by participants). Although the experimenters encouraged participants to try as many times as necessary, they had the freedom to move on before a successful response, which could have influenced the number of commands recorded per task.

In relation to ecological validity, it is also important to note that our research was lab based. This allowed us to minimise potential confounds such as background noise and user distraction. Yet this context may have also made users aware that they were being recorded. Real-world IPA use is likely to vary on these dimensions in comparison to a lab based environment. Future work should therefore aim to replicate our findings in a real-world deployment.

Rather than using text based task instructions, we used pictograms to inform participants what to complete during the study. This was to ensure that the processing of non-native language in task instructions for L2 users did not confound any mental workload effects. The use of pictograms also ensured that text-based instructions did not influence subsequent language used when making commands. Future studies with L2 speakers should investigate the mental workload and language production impact of delivering written tasks experienced by speakers in such studies.

Our findings are limited to a relatively constrained linguistic task of IPA interaction. IPAs are generally designed to perform simple tasks [3, 13] through question-answer adjacency pair dialogues [20, 38], rather than being designed for more conversational or open-ended speech tasks [10, 16]. It is important that future research considers the nature of L2 speech behaviours in these more open-ended scenarios.

8 CONCLUSION

Although IPA use has grown, fuelled by their inclusion on smart speakers and smartphones, not all languages are fully supported, leading some users to interact in a non-native language. Our study focused on these non-native (L2) speakers to understand differences in their experience of IPAs from native (L1) speakers from a cognitive and linguistic perspective. We found that L2 speakers experienced significantly higher mental workload than L1 speakers, irrespective of the device they are using. Even though they experience higher load in producing and interpreting the language from the IPA, they did not vary in the way they interacted linguistically with the IPAs, showing similar number of commands, lexical complexity, lexical diversity and lexical adaptation to L1 speakers. Our work sheds light on this under-researched set of users. CUI-based research needs to study this group in more detail to identify ways to support their IPA interactions, reducing the cognitive burden they experience.

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] 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

[3] 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 Article 17 (April 2019), 28 pages. https://doi.org/10.1145/3311956

[4] Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language* 68, 3 (2013), 255–278.

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[6] Allan Bell. 1984. Language style as audience design. *Language in Society* 13, 02 (June 1984), 145. https://doi.org/10.1017/S004740450001037X

[7] Linda Bell and Joakim Gustafson. 1999. Repetition and its phonetic realizations investigating a Swedish database of spontaneous computer directed speech. In *Proceedings of the XIVth International Congress of Phonetic Sciences*, Vol. 99. 1221–1224.

[8] 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

[9] 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* 31, 4 (09 2019), 349–371. https://doi.org/10.1093/iwc/iwz016

[10] Leigh Clark, Nadia Pantidi, Orla Cooney, Philip Doyle, Diego Garaialde, Justin Edwards, Brendan Spillane, Emer Gilmartin, Christine Murad, Cosmin Munteanu, and et al. 2019. What Makes a Good Conversation? Challenges in Designing Truly Conversational Agents. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, Article Paper 475, 12 pages. https://doi.org/10.1145/3290605.3300705

[11] Benjamin R Cowan, Holly P Branigan, Habiba Begum, Lucy McKenna, and Eva Szekely. 2017. They Know as Much as We Do: Knowledge Estimation and Partner Modelling of Artificial Partners.. In *CogSci 2017 - 39th Annual Meeting of the Cognitive Science Society.*

[12] 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* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 23, 8 pages. https://doi.org/10.1145/3342775.3342786

[13] 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, New York, NY, USA, 43.

[14] S Dornic. 1979. Information processing in bilinguals: Some selected issues. *Psychological Research* 40, 4 (1979), 329–348.

[15] 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.

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

[17] 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

[18] Andy P Field, Jeremy Miles, and Zoë Field. 2012. Discovering statistics using R.

[19] Emer Gilmartin, Francesca Bonin, Loredana Cerrato, Carl Vogel, and Nick Campbell. 2015. What's the game and who's got the ball? genre in spoken interaction. In *2015 AAAI Spring Symposium Series*.

[20] 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

[21] 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.

[22] Pierre Guiraud. 1954. *Les Charactères Statistiques du Vocabulaire. Essai de méthodologie.* Presses Universitaires de France, Paris, France.

[23] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage Publications Sage CA: Los Angeles, CA, 904–908.

[24] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[25] Abhinav Jain, Minali Upreti, and Preethi Jyothi. 2018. Improved Accented Speech Recognition Using Accent Embeddings and Multi-task Learning.. In *Interspeech*. 2454–2458.

[26] 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

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

[28] David R. Large, Leigh Clark, Gary Burnett, Kyle Harrington, Jacob Luton, Peter Thomas, and Pete Bennett. 2019. "It's Small Talk, Jim, but Not as We Know It.": Engendering Trust through Human-Agent Conversation in an Autonomous, Self-Driving Car. In *Proceedings of the 1st International Conference on Conversational User Interfaces* (Dublin, Ireland) *(CUI '19)*. Association for Computing Machinery, New York, NY, USA, Article 22, 7 pages. https://doi.org/10.1145/3342775.3342789

[29] David R. Large, Leigh Clark, Annie Quandt, Gary Burnett, and Lee Skrypchuk. 2017. Steering the conversation: A linguistic exploration of natural language interactions with a digital assistant during simulated driving. *Applied Ergonomics* 63 (Sept. 2017), 53–61. https://doi.org/10.1016/j.apergo.2017.04.003

[30] Patrick Mair and Rand Wilcox. 2019. Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods* (2019). Forthcoming.

[31] Lotte Meteyard and Robert A.I. Davies. 2020. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language* 112 (June 2020), 104092. https://doi.org/10.1016/j.jml.2020.104092

[32] Roger K Moore. 2017. Is spoken language all-or-nothing? Implications for future speech-based human-machine interaction. In *Dialogues with Social Robots*. Springer, 281–291.

[33] 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

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

[35] Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied linguistics* 24, 4 (2003), 492–518.

[36] 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

[37] Andrew Pawley and Frances Hodgetts Syder. 1983. Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics* 7, 5 (1983), 551–579.

[38] 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.

[39] 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* (Melbourne, Australia) *(OzCHI '18)*. Association for Computing Machinery, New York, NY, USA, 548–553. https://doi.org/10.1145/3292147.3292236

[40] 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* (Glasgow, Scotland Uk) *(CHI EA '19)*. Association for Computing Machinery, New York, NY, USA, 6. https://doi.org/10.1145/3290607.3313038

[41] R Core Team. 2018. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[42] 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.

[43] Norman Segalowitz and Jan Hulstijn. [n.d.]. Automaticity in bilingualism and second language learning. *Handbook of bilingualism: Psycholinguistic approaches* ([n. d.]), 371–388.

[44] Richard Towell, Roger Hawkins, and Nives Bazergui. 1996. The development of fluency in advanced learners of French. *Applied linguistics* 17, 1 (1996), 84–119.

[45] Roeland Van Hout and Anne Vermeer. 2007. Comparing measures of lexical richness. *Modelling and assessing vocabulary knowledge* (2007), 93–115.

[46] Catherine Watson, Wei Liu, and Bruce MacDonald. 2013. *The effect of age and native speaker status on synthetic speech intelligibility.*

[47] Richard Wiese. 1984. Language Production in Foreign and Native Languages: Same or different? In *Second Language Productions.* Narr Verlag, Tübingen, Germany, 11–25.

[48] 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. (2020). Submitted.