Supporting Complex Tasks Using Multiple Devices

Elnaz Nouri
Microsoft Research AI, Redmond, USA
elnouri@microsoft.com

Adam Fourney
Microsoft Research AI, Redmond, USA
adamfo@microsoft.com

Robert Sim
Microsoft Research AI, Redmond, USA
rsim@microsoft.com

Ryen W. White
Microsoft Research AI, Redmond, USA
ryenw@microsoft.com

ABSTRACT

Intelligent digital assistants help people perform simple tasks such as question answering, event scheduling, etc. Complex task assistance requires monitoring goals, intentions, and task state. Additionally, complex tasks may exceed the capabilities of one single device or application, giving rise to the need to support people as they engage with two or more devices (e.g., smartphones, smart wearable devices, tablets, and desktop or laptop computers) simultaneously. Using multiple connected devices to support tasks has now become a viable design option for digital assistance since many users already own multiple devices. In this paper, we introduce a framework for developing multiple device experiences (MDX), then demonstrate some of the early capabilities of the framework by means of the *AskChef* cooking assistant. AskChef allows people to use their existing mobile devices (e.g., tablets), together with existing headless smart speakers (e.g., the Amazon Echo) for rich step-by-step guidance and contextual question answering over the culinary recipe loaded in their mobile device's web browser. We conducted a pilot study in which people engaged in cooking recipes with the AskChef MDX system. All participants reported finding the support to be helpful, and indicated a willingness to use the system in the future. Nevertheless, AskChef erred on occasion, affording an opportunity to study how people respond to errors in multi-device scenarios. We examine participants' perceptions and actions, and outline some directions for future research.

KEYWORDS

Task Support, Multi Device Experiences, Intelligent Assistants, Conversational Systems, Voice Assistants, Speaker and Screen, Complex Task Support, Task Assistance, Multiple Device Assistance

1 INTRODUCTION

In recent years, the usage of digital voice assistant devices at home has grown steadily [1, 14] and is forecast to continue to show a compound annual growth rate exceeding 50% by 2020 [18, 24]. Leading device manufacturers (e.g., Amazon, Google) and third-party service developers (e.g., CapitalOne, Dominos, Honeywell) create skills for assistants to allow users to accomplish a range of tasks such as playing music, ordering pizzas, shopping online, scheduling an appointment, checking the weather, making a payment, and controlling other smart devices (e.g., appliances, lighting, thermostats) [11]. 64% percent of users have reported that they purchased the speakers in order to control other smart home devices [4]. More generally, consumer studies in different markets for personal digital assistants show that currently the top use cases for them fall into the following categories: playing music and listening to weather and news, trailed by requesting updates on travel, playing audio books, and setting reminders [14, 15]. Relevant to this work, these devices are often placed in the kitchen [17], and are used to manage cooking processes such as setting timers [17] when boiling eggs [15]. Each of these examples is a simple task, commonly accomplished in one turn or exchange with the digital assistant. The majority of users perceive their assistants to be simple task-based systems that can help them while they are engaged in doing other primary tasks. Users do not attempt to use them for complex tasks, mainly because most assistants cannot provide visual confirmation, and users perceive the cost of the failure of the task to be too high [14, 19]. This lack of trust in the ability of the assistants to help in achieving complex goals is partly due to the failure of the natural language understanding capabilities of assistants when dealing with complex and multi-part requests and the framing of natural language commands [10, 16].

These observations lead us to the important question: *How can digital assistants be deployed for supporting more complex multi-turn tasks?*

Assisting users to accomplish more complex tasks requires the assistant to understand the nature of the complex task, users' utterances and multi-modal behavior, goals and intents, and the current state or progress of the task. In many cases, it also requires providing assistance as people interact with multiple devices in their environment (e.g., engaging with a smart phone, smart speaker, and IoT appliance). The ability to use multiple devices to support complex tasks is important and challenging [7]. Leveraging the availability and capabilities of multiple smart devices, and the increasing trend in dual usage, prompts the design of scenarios in which multiple devices are employed for completing multi-step

tasks. Two thirds of Americans own multiple electronic devices, such as smart phones, smart wearable devices, tablets, and desktop or laptop computers [5], making multi-device experiences (MDXs) viable for many individuals. Each device has unique strengths in aspects such as display size and resolution, processing power, portability, sensing and communications capabilities, and audio input and output. For example, pairing an Amazon Echo smart speaker with an Apple iPad tablet to support a complex task can add visual feedback, contextual understanding (e.g., through content being viewed on the iPad), and touch input to the smart speaker, while also adding far-field speech recognition, and Zigbee personal area network radios to the tablet.

The goal of our work is to introduce a multi-device framework that can be used to create experiences to support users in accomplishing complex tasks. Through rich multi-modal, stateful interaction, we want to help people meet their goals through awareness of their attention, the task, and the information at hand.

We have developed *MDX*, an extensible cloud-based framework that, among other things, manages multiple devices to guide users through completion of complex tasks. In order to demonstrate the capabilities of our framework, we have developed an example application system for assisting users in cooking recipes. An instance of this system is an application called AskChef which assists users in preparing a recipe, and allows a smart speaker and screen to be leveraged simultaneously towards this goal. Cooking has also been the focus of related task completion systems [21]. A cloud-based intelligent agent, acting as a broker between speaker and the user's browser, maintains the state of the task, which includes an awareness of the recipe the user is browsing. In this setup, the user can interact with the recipe, and receive guidance from the agent using either or both devices.

In this paper we address the user behavior immediately after they receive a response from the digital assistant. We study usage of the AskChef system in terms of its visual, vocal, and touch capabilities, as well as its accuracy at determining the intent of, and responding to users' requests.

2 BACKGROUND

Digital companions are generally designed to not only assist users by accomplishing central tasks, but they also are intended to have long term discourses and personal knowledge of the user, and are able to lay the foundations of a relationship with the user [26]. Earlier work [6, 13] had brought about visionary insight in to how digital assistants with conversational capabilities are to be designed. More recent studies have focused on the development of conversational assistants for specific contexts or settings with a preference towards multi-modality, or embodiment, to more effectively simulate the nuances of human communication [9].

As one might expect, a major use-case for having a digital assistant equipped with natural language understanding capabilities is the ability to interact with it 'hands free.' This increases convenience for the user and facilitates time saving [26]. However, the majority of digital assistants use only one device (a smart speaker) for engaging with users. This has limitations in areas such as user understanding and skill discovery [25].

Multi-device experiences have previously been studied from the perspective of synchronizing state across form factors, in order to engage in sequential access to the underlying state [22]. For instance, a user may desire to edit a document on their work computer, and read it later on their phone. In this setting, users engage with device A *then* device B. Experiences may also be developed that render seamlessly on multiple devices [12] (i.e., device A *or* device B). Others have discussed the simultaneous use of multiple devices to complete a task [7] (i.e., device A *and* device B); our focus in this paper. Furthermore, the growing range of intelligent devices on the market, each enabled with some combination of screens, cameras, far-field microphones and other sensors, enables the possibility of corralling available sensing and output resources and bringing them all to bear in assisting the user with their current task. We posit that users of digital assistants will benefit from the simultaneous use of multiple form factors and input/output modalities, and that these benefits will be observed across a range of daily activities.

3 MDX FRAMEWORK

The use of voice commands can be of significant benefit when executing tasks such as cooking, where keeping your hands free is of paramount importance. Furthermore, the far-field microphone and noise localization features of consumer smart speakers yield a suitable sensing modality for interaction in noisy and spacious environments such as a kitchen or garage. However, it is often difficult, if not impossible, to accurately convey the context (i.e., a sense of where you are within a given task) via only voice interactions. Multi-device experiences can help address these issues by leveraging the inherent strengths of each device at the same time.

In the next section we describe a system (*AskChef*) built using MDX which pairs a speaker with a screen-capable device such as PC, tablet, or phone to add a visual interface to the experience, where users can quickly and easily recognize where within a given task they currently are. To support such a system, a broker component that sits between the page content and the back-end services (i.e. smart speaker skill, Q&A models, etc) is required. The broker exposes an application programming interface (API) which provides synchronization of actions between the voice and visual interfaces, as well as enabling visual cues to the user indicating where they are within a task.

Figure 1 shows the architecture of the MDX framework. The architecture consists of three main layers- System, artificial intelligence (AI), and user experience (UX). The System layer provides the necessary infrastructure to manage authentication, state management (both dialog context and navigational context), parsing of web page content (in this case focusing on pages that use schema.org semantic markup), and synchronization across devices via an event hub. The AI layer provides the models necessary for language understanding (recognizing the intent of, and responding to, user utterances), document understanding, question understanding, and answer generation. Domain models apply to the specific domain supported in the experience (cooking in the case of AskChef). Horizontal models span domains and offer support for more general activities such as navigation (scrolling, pagination, etc.). Finally, the UX layer provides the visual presentation logic for visualizing the
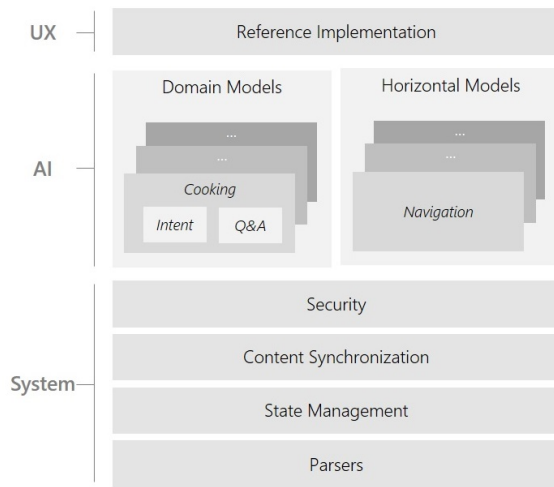
Figure 1: High-level MDX architecture.

Figure 2: Screenshot of the AskChef MDX application visual interface for third-party content (in this case running as a web browser extension).

current state. In our implementation we have explored enabling a visual UX layer on top of third-party web content, as well as a UX layer integrated with first-party content.

4 ASKCHEF COOKING APPLICATION

In the case of cooking, the UX layer described in the previous section presents the recipe, highlights the current step, and provide a visual readout of spoken responses.

Interactions with AskChef are available in two forms:

(1) **Voice**: Voice commands can be used not only to navigate within a task, but can also be used for open-ended question-and-answer exchanges specific to the task at hand.

(2) **Visual+Touch**: Touch, either via mouse or finger(s), can be used on the paired visual display to navigate within a specific task.

The implementation of AskChef in our study uses a browser plugin to unobtrusively overlay presentation on top of recipe web sites, as seen in Figure 2 - both in the right rail (showing the text of system and user engagement) and in the web page (to highlight current steps and/or answers). In practice, the UX layer could also be an application or website running any screened device.

4.1 Language and Document Understanding

AskChef employs a call and response interaction style- users issue verbal queries and the system responds by updating state or responding with an appropriate answer. Text-to-speech (TTS) and utterance intents are processed by Microsoft's TTS Cognitive Services and Language Understanding services, respectively [2, 3]. The language model was developed using queries issued by users in an initial "Wizard of Oz" setup [23].

To enable the system to intelligently respond to user queries, it requires an understanding of the navigational context. For instance, commands such as "Go to the third step," require that the system can identify what part of the document consists of the corresponding step. In the case of recipes, we take advantage of the schema.org
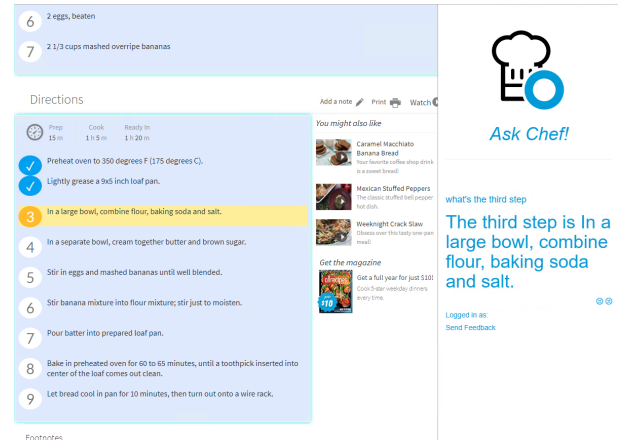
structured markup available on most recipe pages to parse out steps, ingredients, and nutritional information [8]. In other domains, more sophisticated parsing and document representation may be required. Table 1 illustrates an example series of interactions between a user and the system.

4.2 Context Based Question Answering

As can be seen in Table 1, AskChef recognizes utterances from the user that pose a question regarding the content of the loaded webpage and provides the best appropriate answer from the content of the page in response to user's question. In some cases, the system is able to recognize the user's intent ("Could you repeat that?", "How hot do I heat the oven?") but is missing the capability to handle the request, and so responds accordingly ("I don't yet support your requested action."). Other cases that are not depicted include responding to speech recognition failures, which are detected as requests that seem out of context.

5 EXPERIMENT

In this section we explain the details of our experiment.

5.1 Goal

This experiment reflects our initial testing of the AskChef system after its development. The goal was to evaluate several aspects of the user experience. Specifically, we examined ease of use, error rates in speech or intent recognition, users' preferences for interaction with voice versus touch, and overall user impressions.

5.2 Procedure

We invited eight participants (four male, four female) with an average age of 34, to use the AskChef system with a smart speaker (Harman Kardon Invoke, comparable to Amazon's Echo) and device with a screen (laptop PC) with the goal of preparing a culinary recipe. All participants reported having used an intelligent digital assistant before participating in our study and that they do cook. Participants had not previously used a digital assistant recipe

User	System Response
what are we cooking?	
	we are cooking Banana Bread
what's the 1st ingredient?	
	1 cup + 2 Tablespoons all-purpose flour
how much flour	
	1 cup + 2 Tablespoons all-purpose flour
could you repeat that	
	I'm sorry, I don't yet support your requested action but I am learning all the time.
what's the next ingredient	
	2 Tablespoons cornstarch
next	
	1 teaspoon ground cinnamon
what's next	
	2 teaspoons baking powder
what's the 1st step	
	Preheat the oven to 325°F. Lightly butter the bottom and sides of a 9-by-5-inch loaf pan.
what's next	
	Mix flour, salt, baking powder, baking soda, corn starch, cinnamon and sugar in a large bowl
how hot do I heat the oven	
	I'm sorry, I don't yet support your requested action but I am learning all the time.
how much sugar do I need	
	2 Tablespoons sugar

Table 1: Example interaction session between a user and the AskChef system.

preparation. The experiment was run in a public kitchen area at our research facility. Participants were briefed upon arrival about the logistics of their interaction and the task, its potential hazards, and their rights. They were told that they can get guidance from the system for cooking the recipe but that they can freely and naturally choose to use the speaker or screen. They could choose one from a small selection of recipes, all selected based on the possibility of preparing them using a microwave (primarily for simplicity and risk mitigation). The session was recorded using an overhead camera and a webcam attached to the screen. The laptop with the screen and smart speaker were situated about 1m apart and were both set on the table. Because we want to use our system in real kitchen environments at home, and intended to investigate user behavior in real and possibly naturally noisy conditions in the kitchen environment, we did not make any effort to block access by other people to the kitchen area while our participants were cooking the recipe. Figure 3 shows a participant using AskChef system and engaging with the speaker and the screen while preparing a meal in our study.

The average time taken to prepare a recipe was 30 minutes. Two participants made Chicken Taco, three made Cinnamon Muffin, one made Rissoto and two made Tiramisu. Total number of turns exchanged between the system was 288 turns. Average number of turns per interaction was 36 turns.

Figure 3: Experimental participant preparing a recipe with assistance from the AskChef system. The smart speaker and the laptop are clearly visible on the right of the image.

6 INITIAL FINDINGS

In this section we review the initial findings from our experiment.

6.1 General Impressions

All participants reported that they liked the idea of using the AskChef system throughout the steps in preparation of the recipe and that they liked using the system in practice. They felt that the

Intent	Engagement Rate
ingredient_quantity	20.25
step_ordinal	13.50
intent recognition error	12.88
ingredient_list_all	8.59
step_next	6.75
recipe_name	6.13
ingredient_about	4.29
ingredient_next	3.07
quantity_convert	3.07
Other	21.45

Table 2: Engagement by intent (% of utterances).

support provided by the AskChef system for cooking the recipe was useful and that they would consider using the system again.

6.1.1 Qualitative Feedback on Conversational and Voice Aspects. This early version of the AskChef system only successfully understood user utterances 40% of the time. All participants reported that they felt that the system's failure in responding to their questions affected the quality of their experience in a negative way. Five participants reported that they did not like using the activation phrase needed to address the conversational assistant at every turn of the conversation (which was a requirement in the version of AskChef that has been corrected in later iterations). This finding is consistent with user demands for major digital assistant manufacturers to make it easier to engage a skill and keep it engaged without requiring re-invocation [4].

6.1.2 Qualitative Feedback on Visual+Touch aspect. Two participants reported that they preferred to see the instructions visually before hearing them. They also preferred to read instructions from the screen even though they could interact via voice.

6.1.3 Qualitative Feedback on MDX experience. Participants used the screen to gauge their progress, to quickly confirm the needed ingredients, or read information when the voice assistant failed to adequately answer their questions. They used voice queries to get basic information about the recipe (e.g., preparation time, servings, etc.), to navigate the recipe steps (next, previous), and to ask contextual questions about the current step (e.g. "How long do I stir?"). The pilot study was a successful demonstration of how capabilities of the devices can complement one another for complex task support.

6.2 Intent Engagement

AskChef recognizes the intent of the users at every turn. To do this, it uses a machine-model trained on utterances from a separate, but related user study [23]. The intents are manually defined through a review of those utterances and our requirements for system functionality. We measured the rate at which users engaged the various language understanding intents covered by the system. The top ten intents are reported in Table 2. Requesting the quantity of a user-specified ingredient and navigating to the next step in the task were the top two most common user intents observed in the study.

Component	Error Rate
Speech Recognition Error	21.3
Intent Error	35.9

Table 3: Error rates for key system components (% of utterances).

6.3 System Performance

The main measurement criteria for the performance of this version of AskChef is based on the performance of the conversational component of the system. We used human judgments to evaluate the performance of the conversational capabilities of our system. A judge (one of the authors) reviewed the ⟨user utterance, system response⟩ pairs for correctness. If human evaluation found the system's response to be valid, then that response was considered a success. The error rate per utterance is summarized in Table 3. For speech recognition, AskChef depends on the hardware used and the digital assistant back-end. As such, we have much less control over those errors. We note that of the errors in intent detection 74% were failures to identify any appropriate intent (as opposed to choosing an incorrect intent), suggesting that the system can be readily improved by adding support for additional intents.

7 CONCLUSIONS AND FUTURE WORK

The ability to use multiple devices simultaneously unlocks a range of new interactive experiences to help users complete complex tasks. Our study explored the experience of employing multiple devices, consisting of a screen and voice-based digital assistant, to assist with preparing a recipe. We found that users valued aspects of the system that enabled them to attend to the task of cooking by using voice capabilities and control over the screen. Going forward, our goals include broadening the range of application scenarios for MDX beyond cooking (e.g., education, gaming, navigation, accessibility, etc.), improving intent understanding and question answering (the latter was not covered in this paper), as well as developing mitigation strategies to handle any shortcomings in speech recognition performance. Future work could include having the system more actively involved and prompting the user to complete specific steps or check if an action is completed. In addition, our scenario could be easily extended to use more devices to enhance task state estimation or user understanding. For instance, the overhead camera that was recording the sessions, and the laptop camera and microphone could be integrated. One major goal for multi-step tasks such as cooking is to provide step-wise recommendations (e.g., instructional content) to users to help drive task completion. Beyond tasks where activities are listed and known, activity detection methods could help recognize ongoing activities (e.g., [20]) and use that to generate suggestions.

REFERENCES

[1] 2017. Alexa, Say What?! Voice-Enabled Speaker Usage to Grow Nearly 130% This Year. Retrieved Sep 2, 2018 from https://www.emarketer.com/Article/Alexa-Say-What-Voice-Enabled-Speaker-Usage-Grow-Nearly-130-This-Year/1015812

[2] 2018. Bing Speech API - Speech Recognition Software. Retrieved Sept 21, 2018 from https://azure.microsoft.com/en-us/services/cognitive-services/speech/

[3] 2018. Language Understanding (LUIS) Documentation. Retrieved Sept 21, 2018 from https://docs.microsoft.com/en-us/azure/cognitive-services/luis/

[4] 2018. The Smart Audio Report. Retrieved Sep 2, 2018 from https://www.nationalpublicmedia.com/wp-content/uploads/2018/07/Smart-Audio-Report-from-NPR-and-Edison-Research-Spring-2018ᴅownloadable-PDF.pdf

[5] M. Anderson. 2015. *Device Ownership*. Pew Research Center.

[6] Richard A Bolt. 1980. *"Put-that-there": Voice and gesture at the graphics interface*. Vol. 14. ACM.

[7] Tao Dong, Elizabeth F. Churchill, and Jeffrey Nichols. 2016. Understanding the Challenges of Designing and Developing Multi-Device Experiences. In *DIS '16 Proceedings of the 2016 Conference on Designing Interactive Systems*. ACM, 62–72. http://dl.acm.org/citation.cfm?id=2901851

[8] R. V. Guha, Dan Brickley, and Steve MacBeth. 2015. Schema.Org: Evolution of Structured Data on the Web. *Queue* 13, 9, Article 10 (Nov. 2015), 28 pages. https://doi.org/10.1145/2857274.2857276

[9] Stefan Kopp, Lars Gesellensetter, Nicole C Krämer, and Ipke Wachsmuth. 2005. A conversational agent as museum guide–design and evaluation of a real-world application. In *Intelligent virtual agents*. Springer, 329–343.

[10] Page Laubheimer and Raluca Budiu. 2018. Intelligent Assistants: Creepy, Childish, or a Tool? Users' Attitudes Toward Alexa, Google Assistant, and Siri. Retrieved Sept 21, 2018 from https://www.nngroup.com/articles/voice-assistant-attitudes/

[11] Xinyu Lei, Guan-Hua Tu, Alex X Liu, Chi-Yu Li, and Tian Xie. 2017. The insecurity of home digital voice assistants-amazon alexa as a case study. *arXiv preprint arXiv:1712.03327* (2017).

[12] Michal Levin. 2014. *Designing Multi-Device Experiences: An Ecosystem Approach to User Experiences across Devices*. O'Reilly Media.

[13] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.

[14] Ewa Luger and Abigail Sellen. 2016. Like having a really bad PA: the gulf between user expectation and experience of conversational agents. (2016), 5286–5297.

[15] Alli Montag. 2018. Here is what people actually ask Amazon's Alexa. Retrieved Sept 21, 2018 from https://www.cnbc.com/2018/06/26/study-what-people-actually-ask-amazons-alexa.html

[16] Chelsea Myers, Anushay Furqan, Jessica Nebolsky, Karina Caro, and Jichen Zhu. 2018. Patterns for How Users Overcome Obstacles in Voice User Interfaces. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 6.

[17] Sheryl Ong and Aaron Suplizio. 2016. Unpacking the Breakout Success of the Amazon Echo. Retrieved December 12 2018 from https://www.experian.com/innovation/thought-leadership/amazon-echo-consumer-survey.jsp

[18] Sarah Perez. 2018. 39 million Americans now own a smart speaker, report claims. Retrieved Sep 2, 2018 from https://techcrunch.com/2018/01/12/39-million-americans-now-own-a-smart-speaker-report-claims/

[19] 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.

[20] Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 1194–1201.

[21] Ulrich Schäfer, Frederik Arnold, Simon Ostermann, and Saskia Reifers. 2013. Ingredients and recipe for a robust mobile speech-enabled cooking assistant for german. In *Annual Conference on Artificial Intelligence*. Springer, 212–223.

[22] Katarina Segerståhl. 2009. Crossmedia systems constructed around human activities: a field study and implications for design. In *IFIP Conference on Human-Computer Interaction*. Springer, 354–367.

[23] Alexandra Vtyurina and Adam Fourney. 2018. Exploring the Role of Conversational Cues in Guided Task Support with Virtual Assistants. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 208.

[24] D. Watkins. 2016. Strategy analytics: Amazon, google to ship nearly 3 million digital voice assistant devices in 2017. Retrieved Sept 21, 2018 from https://www.strategyanalytics.com/strategyanalytics/news/strategy-analytics-press-releases/strategy-analytics-pressrelease/2016/10/05/strategy-analytics-amazon-google-to-ship-nearly3-million-digital-voice-assistant-devices-in-2017#

[25] Ryen W White. 2018. Skill discovery in virtual assistants. *Commun. ACM* 61, 11 (2018), 106–113.

[26] Yorick Wilks. 2010. Is a Companion a distinctive kind of relationship with a machine?. In *Proceedings of the 2010 Workshop on Companionable Dialogue Systems*. Association for Computational Linguistics, 13–18.