Ryen W. White et al.

Viewpoint
Multi-Device Digital Assistance

Increased availability of cloud services and ownership of multiple digital devices create unique opportunities for digital assistants to provide guidance across a range of tasks and scenarios.

THE USE OF multiple digital devices to support people's daily activities has long been discussed.[11] The majority of U.S. residents own multiple electronic devices, such as smartphones, smart wearable devices, tablets, and desktop, or laptop computers. Multi-device experiences (MDXs) spanning multiple devices simultaneously are viable for many individuals. Each device has unique strengths in aspects such as display, compute, portability, sensing, communications, and input. Despite the potential to utilize the portfolio of devices at their disposal, people typically use just one device per task; meaning they may need to make compromises in the tasks they attempt or may underperform at the task at hand. It also means the support that digital assistants such as Amazon Alexa, Google Assistant, or Microsoft Cortana can offer is limited to what is possible on the current device. The rise of cloud services, coupled with increased ownership of multiple devices, creates opportunities for digital assistants to provide improved task completion guidance.

Case for Multi-Device Digital Assistance

Arguments in favor of multi-device support are not new. Cross-device experiences (CDXs) and MDXs have been discussed in the literature on interaction design, human factors, and per-

vasive and ubiquitous computing.[2,8] CDXs have focused on scenarios such as commanding (remote control), casting (displaying content from one device on another device), and task continuation (pausing and resuming tasks over time). In CDXs, devices are often used sequentially (that is, device A *then* device B) and are chosen based on their suitability and availability. Tools such as the Alexa Presentation Language (APL) enable developers to

create experiences for different device types (that is, A *or* B). In MDXs, different devices are used simultaneously for task completion (that is, A *and* B). Digital assistants can capitalize on the complementary input and output capabilities of multiple devices for new "better together" experiences.

Although the majority of sold smart speakers lack screens, their capabilities continue to expand, for example, through the introduction of "smart
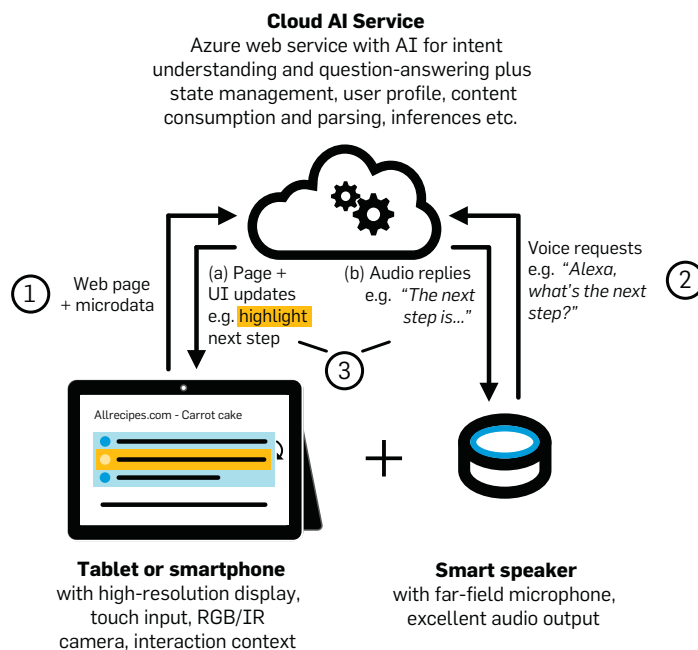
displays" that combine smart speakers and screens in a single device. Despite the value of screens housed on speaker hardware, other devices (such as tablets, smartphones) often have higher-resolution displays, have more powerful processors, and are interchangeable, making them more versatile across tasks and scenarios. These devices may also already be used for the current task, providing valuable contextual information. For example, a recent study of tablet use found that 40% of participants reviewed recipes on their tablet before cooking.[6] This contextual information also helps to address challenges such as deciding when to trigger recommendations about invoking MDXs or ground question-answering in a specific context (for example, a user inquiring "how many cups of sugar?" while on a recipe website). Integrating with existing usage and information access practices also means users need not learn a new device or system to capitalize on multi-device support.

MDXs primarily address capability shortfalls in what individual devices can do, rather than there being anything inherent in the type of tasks that requires physical separation of devices. While a tablet device with a sufficiently powerful speaker and microphone would be perfectly capable of serving the exact same purpose as a speaker plus tablet with just one device, such devices could still be a long way off and require that people purchase a new device. A significant advantage of MDXs is that people can get support now, by pulling together devices they already own. Even new devices will have weaknesses that could well be addressed in combination with existing hardware.

CDX scenarios, such as reviewing background material while engaging with another device illustrate the benefits of multi-device support. Microsoft SmartGlass, a companion application for Xbox that supplements the core gaming console experience, acts as a remote control, and provides additional functionality such as recommendations, profile access, and leaderboards. Companion applications for Netflix (netflix.com) and Hulu (hulu.com) are similar. Conversely, in MDXs, all devices are integral, and contribute in important and complementary ways to the user experience and to task completion.

Figure 1a. Illustrative schematic of the *Ask Chef* MDX, combining tablet and smart speaker, mediated by cloud services, to help with recipe preparation.

In ①, the Web page URL is passed to the service (by an instrumented Website in this case) where the microdata (JSON) is parsed to extract ingredients and preparation steps. The preparation instructions are chunked for audio output and an enhanced version of the Website is displayed, with the steps enumerated and the current step highlighted. In ②, the user requests the next step via voice. The system responds in ③ with two actions *simultaneously*: (a) highlighting the next step on the tablet, and (b) vocalizing the step through the speaker.

Cloud AI Service
Azure web service with AI for intent understanding and question-answering plus state management, user profile, content consumption and parsing, inferences etc.

① Web page + microdata

(a) Page + UI updates e.g. highlight next step

(b) Audio replies e.g. "The next step is…"

③

Voice requests e.g. "Alexa, what's the next step?" ②

Allrecipes.com - Carrot cake

Tablet or smartphone
with high-resolution display, touch input, RGB/IR camera, interaction context

Smart speaker
with far-field microphone, excellent audio output

Figure 1b. Still image of the *Ask Chef* MDX, using an Apple iPad and an Amazon Echo smart speaker. Interactions are coordinated via a Website communicating with the cloud as in Figure 1a.
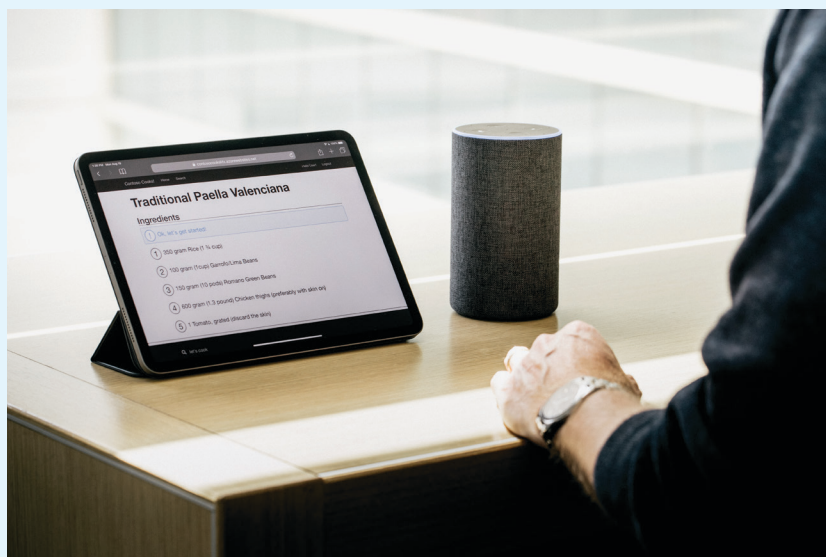
Potential for Multi-Device Digital Assistance

Usage of digital assistants on multiple devices can be low, primarily due to a lack of compelling MDXs. Support for multiple devices in digital assistants is typically limited to CDXs, mostly to assist with task continuation, including quickly resuming edits on a tablet for a document started on a desktop computer (for example, the "Pick Up Where I Left Off" feature in Cortana in Windows 10) or system-initiated pointers from smart speakers to screen devices to view lists or tables.

More than 60 million smart speakers have been sold in the U.S. alone and co-ownership of screen devices by smart speaker users is high. The benefits of building digital-assistant experiences combining screen and non-screen devices are bidirectional: screens offer visual output to augment audio plus (often) support for touch/gestural interactions, smart speakers offer hands-free capabilities through far-field microphones plus high-quality audio output. Smart speakers can also serve as the hub for Internet-of-Things (IoT) smart home devices (for example, Amazon Echo Plus has a Zigbee[a] radio to support device connectivity) and are often situated where they can easily be accessed and are proximal to task hubs (for example, in the kitchen for cooking support or in the family room to control entertainment devices). Many of these IoT devices already use conversational interfaces tailored to the device. Although engaging with devices might have once required immediate physical proximity, conversational interfaces and improved technical capabilities in areas such as far-field speech recognition have alleviated this need. A successful multi-device digital assistant would manage and leverage multi-modal input/output and conversational interfaces to create seamless simultaneous interactions with multiple smart devices, irrespective of the device manufacturer.

Guided Task Completion

At Microsoft, we have developed an MDX application called *Ask Chef* as part of a larger effort to build an exten-

a   See https://bit.ly/1KfBrJS

Digital assistants can capitalize on the complementary input and output capabilities of multiple devices for new "better together" experiences.

sible MDX platform for any application or website. *Ask Chef* focuses on using screen and non-screen devices for recipe preparation assistance. Cooking is a common, time-consuming task that requires state and ingredient tracking, and involves multiple steps. People routinely bring smartphones or tablets into their kitchens to help manage these processes and to access step-by-step instructions via Web browsers. Here, there is often a need for hands-free, far-field interaction.[4] Likewise, smart speaker devices such as the Amazon Echo or Google Home are frequently placed in the kitchen and are used at mealtime to set timers or manage short cooking-related processes.[b] There is an opportunity for digital assistants to help people prepare recipes more effectively by providing support including coaching,[6] status tracking, and coordinating multiple courses of a meal.

Figure 1 illustrates multi-device digital assistance in *Ask Chef*, spanning two devices: a tablet (for example, Microsoft Surface, Apple iPad) and a smart speaker (such as Amazon Echo, Google Home), and mediated by a cloud service, which orchestrates the experience to establish and maintain session state, apply artificial intelligence (for example, for intent understanding and contextual question answering), and handle events and interface updates across different devices.

Powering MDXs via cloud services means system designers do not need to rely on over-the-air updates to client-

b   See https://bit.ly/2KZdA5q

side code to make experience and/or algorithmic modifications, that usage can be easily aggregated and analyzed to improve assistance offered, and that assistants can run on third-party hardware, enabling scaling via Amazon[c] and Google[d] skills kits. Skill availability is only part of the challenge for digital assistants; *discoverability* of skills remains an issue on devices that lack displays.[12] Screen devices can surface skill recommendations and support recall of prior skills on headless devices.

The current implementation of *Ask Chef* relies on schema.org microdata for the Web page being accessed. This markup is used to extract ingredients and preparation instructions. Important extensions include generalizing to content that is not enriched with such data and integrating additional content to augment the recipe (for example, refinements from user comments[3]). Recommending assistance for the current step in the task (including instructional content: videos, Web pages, and so forth), while also considering the previous steps, assistance already offered, and future steps. Determining how to utilize wait times between steps in recipe preparation (for example, "bake for 20 minutes") can be challenging, and users may elect to spend that time in different ways (from food preparation to unrelated tasks [such as email, social media, and other activities]).[6] Beyond task guidance, digital assistants could also provide pre-preparation support (for example, add items to a shared grocery list) and post-preparation support (for example, help revisit favorite recipes).

Looking Ahead

MDXs enable many new scenarios. Support for guided task completion can extend beyond cooking to include procedural tasks such as home improvement, shopping, and travel planning. Other scenarios such as homework assistance, games, puzzles, or calendar management could be enhanced via MDXs. Support for these scenarios could be authored by third parties. For example, educators could compose or flag visual/multimedia content to accompany tutorial/quiz materials, to

c See https://amzn.to/2cDSN3K
d See https://bit.ly/2NC7VnF

> **A significant advantage of MDXs is that people can get support now, by pulling together devices they already own.**

help students learn more effectively than with text or audio only.[1]

As experience with devices such as the Amazon Echo Show has demonstrated, augmenting voice-based digital assistants with a screen can also enable new scenarios (for example, "drop ins"—impromptu video calls). This adds value even though the screen is small; more would be possible with a larger, higher-resolution display that could be located as far from the smart speaker as needed. The user-facing camera (webcam, infrared camera) on many laptops and tablets can add vision-based skills such as emotion detection and face recognition to smart speakers. Powerful processors in tablets and laptops enable on-device computation to help address privacy concerns associated with handling sensitive image and video data.

Multi-device digital assistance is not limited to a single, static device pairing. For example, it includes scenarios such as dynamically connecting a smartphone and any one of many low-cost smart speakers as users move around a physical space; imbuing, say, any Amazon Echo Dot with the capabilities of an Echo Show. Although we targeted MDXs comprising two devices, there are situations where three or more could be used (for example, adding a smartphone to *Ask Chef* for timer tracking and alerting); these experiences must be carefully designed to avoid overwhelming users. Multi-device interactions can also help correct errors in speech recognition and yield useful data to improve voice interfaces.[9]

In sum, MDXs unlock a broad range of more sophisticated digital assistant scenarios than are possible with a single device or via CDXs. Utilizing complementary devices *simultaneously* could lead to more efficient task completion on current tasks, cost savings for device consumers, and unlock new classes of digital assistant skills to help people better perform a broader range of activities. [C]

References
1. Carney, R.N. and Levin, J.R. Pictorial illustrations still improve students' learning from text. *Educational Psychology Review 14*, 1 (Jan. 2002), 5–26.
2. Dong, T., Churchill, E.F., and Nichols, J. Understanding the challenges of designing and developing multi-device experiences. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems* (2016), 62–72.
3. Druck, G. and Pang, B. Spice it up? Mining refinements to online instructions from user generated content. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, (2012), 545–553.
4. Jokela, T., Ojala, J. and Olsson, T. A diary study on combining multiple information devices in everyday activities and tasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, (2015), 3903–3912.
5. Kiddon, C. et al. Mise en Place: Unsupervised interpretation of instructional recipes. In *Proceedings of Empirical Methods on Natural Language Processing*, (2015), 982–992.
6. Müller, H., Gove, J., and Webb, J. Understanding tablet use: A multi-method exploration. In *Proceedings of the 14th International Conference on Human-Computer Interaction with Mobile Devices and Services* (2012), 1–10.
7. Pardal J.P. and Mamede N.J. Starting to cook a coaching dialogue system in the Olympus framework. In *Proceedings of the Paralinguistic Information and Its Integration in Spoken Dialogue Systems Workshop* (2011).
8. Segerståhl, K. Crossmedia systems constructed around human activities: A field study and implications for design. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (2009), 354–367.
9. Springer, A. and Cramer, H. Play PRBLMS: Identifying and correcting less accessible content in voice interfaces. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (2018), 296–305.
10. Sørensen, H. et al. The 4C framework: Principles of interaction in digital ecosystems. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, (2014), 87–97.
11. Weiser, M. The computer for the 21st century. *Scientific American* Special Issue on Communications, Computers and Networks, (1991), 94–104.
12. White, R.W. Skill discovery in virtual assistants. *Commun. ACM 61*, 11 (Nov. 2018), 106–113.

Ryen W. White (ryenw@microsoft.com) is Partner Research Manager at Microsoft Research AI, Redmond, WA, USA.

Adam Fourney (adamfo@microsoft.com) is Senior Researcher at Microsoft Research AI, Redmond, WA, USA.

Allen Herring (allenh@microsoft.com) is Principal Research Engineer at Microsoft Research AI, Redmond, WA, USA.

Paul N. Bennett (pauben@microsoft.com) is Senior Principal Research Manager at Microsoft Research AI, Redmond, WA, USA.

Nirupama Chandrasekaran (niruc@microsoft.com) is Principal Research Engineer at Microsoft Research AI, Redmond, WA, USA.

Robert Sim (rsim@microsoft.com) is Principal Applied Science Manager at Microsoft Research AI, Redmond, WA, USA.

Elnaz Nouri (elnouri@microsoft.com) is Senior Applied Scientist at Microsoft Research AI, Redmond, WA, USA.

Mark J. Encarnación (markenc@microsoft.com) is Principal Development Manager at Microsoft Research AI, Redmond, WA, USA.