Voice Toolkit

By

Lauren L. Dillard

A thesis submitted in partial
fulfillment of the requirements for the
Master of Science Media Management Program

THE NEW SCHOOL

(April 30, 2018)

CAPSTONE
MEDIA MANAGEMENT

Capstone Profile

Project Type:
Management Solutions Research

Personal Goals:
Lauren's goal is to develop expertise in designing voice-enabled experiences. This will allow her to position herself as a thought leader in the industry and a go-to resource at TandemSeven.

Capstone Project Goals:
The goal of Lauren's capstone is to develop training resources for designing voice-enabled experiences. This training material will prepare user experience designers at TandemSeven to design voice experiences for their clients.

Audience:
This toolkit has been developed for the user experience designers of TandemSeven.

Professional Profile:
Lauren Dillard is a problem solver working at the intersection of storytelling, entrepreneurship and user-centered design. She has over a decade of experience delivering engaging products. Lauren currently works as a senior user experience designer at TandemSeven.

Career Goals:
Shortly after beginning her degree program at The New School, Lauren was hired by TandemSeven. It is her goal to continue practicing user experience design, developing skill and knowledge in the field.

Table of Contents

Section 1

Introduction

1.1 Purpose

The purpose of this document is to serve as an early introduction and guide to the development of products and services that leverage voice interfaces. These tools evoke memories of the *Iron Man's* Jarvis, *Knight Rider's* HAL or the *Star Trek* computer, but — in practice — refer to smart speakers that can be controlled through spoken word. For the user experience designers at TandemSeven, it's probably something in between.

1.2 Methodology

Three stakeholder interviews were conducted with Partner, Business Lead for TandemSeven Consulting David Cowing; Vice President, User Experience Practice for TandemSeven Consulting Elizabeth Srail; and Partner, Business Lead for Cora Journey 360 David Clark. Cowing was interviewed in person in his office in New York on Feb. 20, 2018. Srail was interviewed remotely from her office in Chicago on Feb. 22, 2018. Clark was interviewed remotely from his home outside of Boston on March 3, 2018. All of the interviews were recorded and transcribed. The stakeholder interviews outlined key focus areas for this research and revealed outstanding questions around integrating voice into TandemSeven's customer experience design practice.

Eight interviews were conducted with subject matter experts from March 6 - 20, 2018. The interviews were conducted remotely; Six via Zoom Meetings and two over the telephone. The six meetings conducted online were recorded and transcribed. Detailed notes were captured by hand and transcribed for the two telephone meetings. A list of the participants can be found in the appendix.

Other research for this project spans multiple semesters. Content about the history of natural language processing and machine learning was collected in Fall 2017 for a Big Data and the Media course taught by Professor Robert Berkman. Other content was collected from recommended resources; industry posts and whitepapers; and relevant

news. These sources are referenced throughout this document and cited in the appendix.

1.3 Research Objectives

Through interviews with TandemSeven stakeholders, the following research objectives were identified for this work.

- Understand voice project level of effort, time and team
- Understand how to educate client on value and use cases
- Validate that user experience research practices apply to voice
- Understanding techniques, tools and process for design
- Understand what design means in the context of voice
- Understand how designers are thinking across modalities
- Understand deliverables for voice projects and skills
- Understand testing process / activities and pricing

Section 2

What is Voice?

2.1 The Backstory

Computers have never understood us. Clever humans have encoded machines to react to our actions. We communicate via a nudge of the mouse, a tap on a keyboard, or a swipe on a glass screen. But, still, our machines ascribe no meaning to our actions.

Enabled by a one-trillion-fold increase in available computing power (“Hard Disk Drive Morph”), the disciplines that enable conversational interfaces have begun to converge, creating a new interchange on the information superhighway.

Early Experiments

Guided by experience as a cryptanalyst during World War II, Andrew Booth of the University of London conducted the earliest experiments in natural language processing — in this case, machine translation of English to French. This 1948 experiment resulted in a computer-based dictionary lookup that aided human translators (Terras 123).

A few years later, while Bell Labs was developing the first spoken-number recognition tools (Pieraccini 3), IBM was working with Georgetown University to advance machine translation. Informed by the 1957 publication of “Syntactic Structures” by Noam Chomsky, these researchers treated the Russian language like a code that could be cracked using codified rules and language structures. Using typed input, the team successfully translated more than 60 russian sentences into English, but the work stalled (“701 Translator). It was too difficult to codify every possible sentence structure.

Israeli philosopher and mathematician Yehoshua Bar-Hillel “concluded that fully-automatic high-quality translation is impossible without knowledge” (Hancox). Though researchers had begun introducing machines to corpora of domain-specific data as early as 1966, that work had largely been conducted independently from work in natural language processing, including the work of machine translation.

Paradigm Shift

By introducing domain data, Stanford's Terry Winograd created a program that enabled the virtual manipulation of blocks via typed, natural language input (Winograd). G.G. Hendrix created the LIFER/LADDER system that "used semantic grammar to parse questions and query a distributed database," answering questions about U.S. Navy ships (Rao 220).

Roger Schank introduced "Conceptual Dependency Theory" in 1969. It was a turning point for natural language processing. This theory created a relationship between rote machine translation and a body of knowledge. Schank proposed that "there was a predetermined set of possible relationships that made up an interlingual meaning structure. These relationships could be used either to predict conceptual items that were implicit in a sentence or, coupled with syntactic rules, to inform a parser what was missing from a meaning and where it might be found in a sentence" (Schank 245).

By the 1980s, Moore's law — the prediction of year-over-year increases in computing power — enabled natural language processing using the probabilistic models described by Schank. This made both probabilistic modeling and machine learning possible.

"We thought it was wrong to ask a machine to emulate people. After all, if a machine has to move, it does it with wheels — not by walking. If a machine has to fly, it does so as an airplane does — not by flapping its wings. Rather than exhaustively studying how people listen to and understand speech, we wanted to find the natural way for the machine to do it," said Fred Jelinek, distinguished professor at Cornell in information theory and researcher at IBM ("Pioneering Speech Recognition").

Convergence

A number of products for voice transcription were released in the 1980s and 1990s that were expensive and difficult to use. They required the user to speak with a slow, spaced enunciation (Pinola). The first "large vocabulary" continuous speech recognition product

was called IBM MedSpeak (“Pioneering Speech Recognition”). By 2000, the price had come down and continuous speech recognition systems could operate with about 80 percent accuracy (Pinola).

Separately, machine learning was making advancements as well. A group of graduate students at Carnegie Mellon University developed “Deep Thought,” a computer that was able to defeat a grandmaster in a regular tournament game of chess in 1989. This work seeded the early development of IBM’s Deep Blue computer that defeated chess champion Garry Kasparov in 1997 (Greenemeier). Success with Deep Blue led to IBM’s development of Watson — the computer that would go on to defeat *Jeopardy!* grand champions Ken Jennings and Brad Rutter in 2011 (Markoff).

The industry had made impressive strides in both natural language processing and machine learning. Apple and Microsoft took note. Both the Mac OS X (2001) and Windows Vista (2006) operating systems contained native voice controlled technology (Pinola). As long as the mouse and keyboard were available, user adoption to voice-control remained low. The release if the iPhone in 2007 changed the paradigm. Voice control using either the iOS native client (Siri) or Google Voice Search (2008) eased the burden of tapping the tiny keyboard buttons shown on screen (Pinola).

2.2 Adoption

After Web Services, Amazon may be best known for its machine learning capabilities — found in customized shopping experiences. Teams at Amazon were given the challenge to think of a not-so-distant future where cloud services were infinite — storage and capacity would no longer be tied to the capability of the user’s device — and machine learning could solve any problem (Fortune Editors). In 2011, the Alexa team pitched Amazon Senior Vice President Dave Limp on the Echo — a puck-like device with a lot of potential (Kim).

In a period of just two years, as of January 2017, Consumer Intelligence Research Partners estimated that 8.2 million people purchased an Amazon Alexa-enabled device

(Soper). As of December 2017, Amazon holds 69 percent of U.S. smart speaker market share. Google Assistant-enabled devices, while widely acknowledged to have a technical edge (Gebhart), trail behind Alexa-enabled devices with just 25 percent of smart speaker market share (“Amazon Echo & Alexa Stats”) (“Google Home & Assistant Stats”). This is likely due to Amazon’s three year head start on selling hardware (Gebhart).

According to AdWeek, “By 2021, an estimated 1.83 billion people worldwide will be using digital voice assistants, and in the next year, it’s projected that 30 percent of our interactions with technology will be through conversations with smart machines” (Hedges). Beyond adoption of smart speaker, it’s expected that the majority of voice assistant interaction will occur on smartphones. eMarketer predicts that over five billion assistants will be installed on smartphones worldwide by 2022. Including expansion to other platforms like PCs, TVs, tablets, cars and wearables, the total number of voice assistant-enabled devices is expected to reach 870 million in the United States by 2022 (Perez).

Former Amazon Tech Evangelist Liz Myers said, “I think the promise now of artificial intelligence is that it can help us as humans maximize our potential here on Earth. And I think that is really powerful. And I think we're building toward it. I'm seeing new aspects be released that certainly indicate that's the case.”

2.3 Observations

Smart speakers have made voice interactions accessible to the average consumer. They are inexpensive, purpose-driven devices that bring small moments of delight in a flooded consumer market — but they are by no means the end of voice.

Because of their popularity, virtual personal assistants like Alexa and Siri are top of mind, according to Bill Marshall, user interface designer at Nuance Communications. Marshall designs IVR, interactive voice response, system that are often accessed via telephone. “Alexa is all just one-shot for the most part,” Marshall says. “Whereas IVR is

much more of a conversation multi-step. It requires understanding and prompting at a level that makes a conversation possible.”

Kathryn Whitenton, digital strategy manager at Nielsen Norman Group, said that some designers group voice interaction into a broader category of conversational interfaces, which can include typed, tapped and spoken interaction. “And I'm kind of torn if I want to start shifting to using [‘conversational interfaces’] because I think there are a lot of similarities. Does it really matter if it's spoken aloud versus if it's a chat that you're typing into a computer?” Either way, Whitenton says, the interaction meets you on your terms.

For Elijah Vargas — senior manager design, voice user interface at Comcast — meeting the user on their terms increases usability and accessibility. The XR11 and XR15 Comcast remotes now come equipped with voice control (Xfinity). Vargas said, “Someone could ask for something or make a statement and the system or the device that this person may be talking to responds back. If they want to then continue in dialogue or in conversation, we are able to do that.”

Section 3

Designing Voice

In an analogy to the development of apps for iOS and Android operating systems on mobile devices, Amazon, Google and Apple are in a race to create a massive ecosystem of apps for their assistants — which are by-and-large contained within at-home smart speakers. A number of agencies have dedicated themselves to creating Skills for Amazon Alexa and Actions for Google Assistant-enabled devices. RAIN, SpokenLayer, VaynerMedia, WitLingo and XAPPMedia are just a few of the agencies that have been identified by either Amazon or Google for their expertise ("Actions on Google | Google Developers"). At the same time, voice has earned ubiquity outside of the smart speaker as well. Nuance Communications has collected decades of experience developing conversation interfaces for telephone systems. Comcast and Ford Motor Company are developing home-grown multimodal experiences.

Though this research will explore Skills and Actions in detail in a later section, the first step is to understand what is universal about designing for voice — regardless of the platform or underlying natural language processing and machine learning tools.

3.1 Project Scope

According to UI Designer Bill Marshall, when gearing up for a new large-scale IVR project, his team at Nuance Communications will staff the following roles: a program manager, project manager, one to three user interface designers, a development lead, a computational linguist (speech scientist), localization experts as needed, an audio team, and quality assurance and development resources as needed. Comcast's Senior Manager of Design, Voice User Interface Elijah Vargas stressed the need for a conversational designer — someone who specializes in expressing interaction through conversation.

Marshall says that enterprise IVR applications with complex integration with back-end systems will take more than a year to design, develop and implement. According to

Bradley Gregory, voice user experience designer at RAIN, the length of the project depends on the level of ask from clients. They can range from many months to a year or more. Gregory says, it depends on if the skills are meant for marketing or if they are strategic and integrate into the larger ecosystem. RAIN's Vice President of Emerging Experiences Greg Hedges added that it can take four to five weeks for certification if launching an Amazon Skill or Google Action.

Summary

- The design team should include at least one conversational designer or a user experience designer with strong writing skills.
- Simple smart speaker skills can be deployed in a few months, but complex voice experiences with backend integration will take a year or more to design, build, test and deploy.
- Certification for an Amazon Skill or Google Action takes four to five weeks.

3.2 Define

The define, discovery, or requirements phase of voice experience design comprises three phases. First, designers work understand their client, it's business drivers and existing practices. Next, designers work to better understand the user and their needs. Once a few product ideas or solutions begin to take shape, the feasibility of those ideas is validated.

The Client

Hedges said that his team uses stakeholder interviews to better understand the client. Beyond building understanding of the business drivers and strategy for the project, Hedges uses the interviews as an opportunity to assess excitement. "For us, channeling excitement is important, but we make sure it fits the broader strategy." RAIN and Nuance Communications both work to understand the client back-end and solutions. This knowledge is filed away for later to ensure that emerging ideas leverage existing client infrastructure.

Marshall added that early conversations with clients at Nuance Communications always include a survey on brand voice. This helps the IVR team develop a compelling, appropriate expression of the client's brand when design work begins. "We definitely go through that process in a fairly formal way. And then we will create a style guide for prompt writing that is based on that information." Marshall said that he doesn't usually get a chance to speak with users directly as part his discovery activities, but he has had the opportunity to interview call center agents. "Most of these folks have contact centers so we try to understand what experiences their agents have when they're talking to customers on the phone."

Lastly, Hedges said that designers at RAIN will read over anything else the client has out in the world. If they're conducting social listening or have other SEO and analytics data, they consider it an input to the research process.

The User

Because Shyamala Prayaga works as a voice interaction designer for Ford Motor Company, she can mostly skip the first step in this discovery process. With a deep understanding of the business drivers and the stakeholders for her projects at the company, she can focus on understanding users through user interviews and contextual inquiry. For Prayaga, follow-up interviews with existing users can help her understand where to look next. "We have different benchmarking teams and we have different research teams who keep doing lots of research back and forth on what worked for us and how people perceive an existing experience."

Digital Strategy Manager Kathryn Whitenton at Nielsen Norman Group describes contextual inquiries as an opportunity to understand key tasks and determine user goals or motivations. For Gregory, he has an opportunity to learn how tech savvy his users are. This will shape how he writes prompts and helper text. Lastly, for Marshall, interviews help him understand which problems he's trying to solve. Once a few key problems have been identified, the scope and scale of those problems can be validated

with a survey. Both Vargas and Prayaga circle back with surveys to recruit candidates for user testing and validate interview findings.

Validation

Though solution ideas begin to emerge as problems become clear, Prayaga, Vargas and Hedges turn to competitive analysis or benchmarking. This serves two purposes; It scopes the client's opportunity and validates the technical feasibility of that idea. According to Bradley, he assesses the technical feasibility of new skills in two ways. First, he'll look at examples of SSML (speech synthesis markup language) to understand what's possible. If that doesn't resolve the question, he'll move on to similar skills to understand how they're completing the same tasks. Lastly, Prayaga layers in market research and predictions. "We also do lot of competitive benchmarking, market research and then predictions with certain kinds of things. Where is the entire industry heading and what do we anticipate in the next five years?"

Deliverables

After conducting stakeholder interviews, contextual inquiries and other bench research, designers deliver requirements, user stories or use cases, system personas, and user personas.

For Prayaga, the requirements documentation includes feedback on feasibility and prioritization. "We try to see the feasibility. Whether it is feasible technically, financially, every way. And then based on that we prioritize them into our different development timelines. Like 'When do we want it? Which model do we want it to work with?' and things like that."

There's a subtle distinction between user stories and use cases. User stories are simple sentences designed to describe the actions of the main character. For example, "As a user, I want to post a picture so that other users can see it." Designers can deliver one or the other or both. User stories can be used as a jumping off point for a more detailed use case (Seque Quality Control Team).

Both Prayaga and Marshall deliver system personas. For Marshall, these system personas are often accompanied by a list of names. “We have voice talent that we try to suggest because we know that voice talent is good.” According to Prayaga, “We don't have to always do [user] persona creation since that's a one-time thing depending on what kind of vehicle you're working with. So, we do create personas for users. And we do create personas for system, that’s because the system does have a personality based on which we design the prompts.

Summary

- Stakeholder interviews, current-state reviews, and analytics are still good tools to better understand the client, their business needs and their understanding of user.
- User interviews and contextual inquiry are still good tools to better understand users, use cases and existing problems. Surveys can validate findings.
- Market research, competitive benchmarking, and industry predictions can help validate solution ideas and asses feasibility. This is uniquely important for voice projects because lead times are so long.
- Deliverables include requirements with an assessment of priority and feasibility; user stories and / or use cases; user personas; and, of note, system personas. These deliverables can be used to socialize solutions and develop alignment.

3.3 Design

While the types and depth of deliverables in the design phase are determined by how development will proceed, this research has revealed that there is really only one best practice for designing a voice interface. Each designer interviewed, in turn, recommended that anyone attempting this task start with a basic script. Then, and only then, should they proceed to write the extensive variation and logic needed to make a voice experience smooth and seamless.

Activities

“I really strongly advocate that you do this away from the computer at first,” said former Amazon Tech Evangelist Liz Myers. The goal of this activity is to produce a concise script that achieves the user’s objective and sounds like a real conversation. “There's a happy path — or a red thread — and I think it's really important to do that red path well. And that means not embellishing, not dealing with all the possibilities right off the bat. What is what is the simplest, shortest, crispest way to get a cab?

Vinita Atre, former WitLingo intern and HCI graduate student at the University of Maryland College Park, agrees. “In designing for voice, we should probably first consider the challenges and limitations of voice,” Atre says. “It's a linear medium. It’s only one kind of information flow. It flows in only one direction.” As soon as the one-way scripts are ready, Gregory and his colleagues at RAIN will begin rapid prototyping by speaking the parts out loud. Myers leverages one of the many web-based tools available to speak text in addition to role playing with her colleagues. Additional information on available prototyping tools can be found in the appendix.

When designing for multi-modal experiences — experiences that cut across many media — Prayaga recommends designing for one domain and a time and then working to understand the overlap. “There are parts of it that are contingent on the other,” Vargas said about his work a Comcast. “We think about it in the concept of building escalators over elevators. So, we're purposely bringing everything together. But knowing, for whatever reason — like if one thing wasn't happening — it doesn't break the entire experience. There are still parts of it that work.”

Once the happy path has been finalized, Hedges directs his team to begin adding detail to the experience. According to Myers, this includes asking questions like, “What happens if the skill closes prematurely?” Hedges and his team write at least a few alts of the conversation before writing spoken variations for each step in the conversation — from both sides. It’s not quite as important for the UI to utter a new response every time, but the user may express their intent in any number of ways. Each time the user

speaks, the UI needs to recognize what they're saying as a given step in the conversation. This means Gregory has become a prolific writer.

Best Practices

1. Solve a Problem

Whitenton said that engagement with voice skills is quite low. "If I was to evaluate a voice project, I would want to look at how closely it relates to solving problems for the audience." Hedges added that skills can either be sustainably useful or fun and frivolous, but only one of those types of skills will bring the user back for sustained use. If the goal is to create a skill that users engagement, solve a problem for them. Additionally, many voice experience suffer from low discoverability. It's nearly impossible to understand what the experience can do. Be sure to make the problem-solving capabilities of the skill clear from the onset.

2. Use Vocal Variety

In the process of writing her book, Myers has coined the phrase "vocal variety." Vocal variety can be expressed in two ways. First, designers can vary both what and how words are spoken. By pairing a list of interjections — sometimes called "speechcons" — with a few variances on more detailed responses, the designer can build a robust response. Inflections can be built into the TTS (text-to-speech) system by using SSML (speech synthesis markup language). Second, different types of sound can be integrated into the experience including voice, music and sound effects. Sound effects are sometimes called "earcons," similar to "icons." "Randomization is now built into standard templates at Amazon because we discovered that you really need this vocal variety to keep the skills interesting and more natural."

3. Understand Context

Voice experiences should recognize context clues to better serve users. This includes remembering context created within the conversation as well as recognizing and responding to the context of the user. If a user asks "Who is the lead actor in Arrested Development?" After receiving a response, when the user asks, "What else is he in?"

Alexa should remember who “he” is. If a user is using headphones, the experience should not rely on a screen to proceed.

4. Respect Privacy

Vargas recommends ensuring that users feel safe and not violated by either the device or the experience they are using. Part of this, as explained by Prayaga, is to make sure users understand how their information is stored. She recommends that designers always get consent. Of course, there are legal implications here too. Additionally, the experience should be aware of whether it’s safe for the user to speak personal information out loud. If a credit card number is needed to complete a purchase, users should have an alternative to reading it out loud in front of strangers.

5. Be Efficient

It can be extremely efficient to navigate directly to a given function or task with your voice. Vargas and his colleagues at Comcast call it “direct tuning.” However, direct tuning requires the user to remember what to say. Voice experiences should also enable users to accomplish tasks with relative ease. In Prayaga’s words, “‘Is that task easy via voice or via touch?' 'How many steps would it take to complete it if I have to via touch or via voice?' We have our own best practices around some of these and based on that we decide what we should do.” Additionally, the voice experience should limit the length of spoken prompts using the one-breath test (can the designer speak the prompt with just one breath?) and avoid presenting too many options at one time.

6. Fail Gracefully

The voice experience should never blame the user for an error. It can be helpful to fail with a little bit of humor. “I don't know if you've discovered them yet, but there's some SSML — they're called interjections. Emotional, spontaneous, emotional pithy sayings,” Myers said. “So when something goes wrong in this phone booth skill, I go, ‘Oh man, would you like to try again?’ And I give them another chance. And if they make another mistake, I'm randomizing on that too. She goes, ‘Bummer, I think I misheard. Would you try again?’”

7. Build Cul-de-sacs

After giving the user another chance to speak, it may be necessary to prevent a dead-end. “How can we potentially provide them the opportunity to easily get out of that situation?” Vargas recommends building a cul-de-sac around the dead end. That will help the user find another avenue to explore. “Now, they can just kind of continue and move forward.” Planning for eventual failure enables designers to keep the user in the experience and give them a greater chance of reaching their goal.

8. Think Multi-modal

Imagine you ask your Alexa-enabled smart speaker to order an Uber for you. As soon as you shut and lock the door, you’ll need to know how far away the driver is. Today, the Uber skill does provide a slick handoff to your mobile device. This is one of many possible paradigms that multi-modality will take. According to Gregory, “There are only so many use cases were voice alone is useful.” Designers working on voice experiences should think of themselves not as voice-user-interface designers but experience designers that have access to any of the user’s senses.

Iterative and Usability Testing

Through the design process, there are two points at which user testing can greatly improve the usability of the final product. First, Hedges says that iterative testing occurs throughout the design phase to ensure that responses sound right. Vargas adds that Comcast conducts small studies through the design phase to ensure that prompts look and sound right for Comcast’s X1 platform.

For other designers like Marshall, “We generally do not test until we have a functioning application, which means we just do usability testing before launch.” These later tests are not used for iterative work on voice prompts. Instead, they’re used to catch show-stopping problems before development begins. For Prayaga, usability tests are conducted via the Wizard of Oz method. Someone behind the curtain controls a TTS (text-to-speech) device to make it seem like the user is interacting with a live

experience. Whitenton recommends giving users a practice session or time to explore before the test begins. “Normally, we would never do that with something like a website,” Whitenton said. “We want to get people's realistic first impression, because that's how they're going to use it in the real world. But because it's an unusual situation and because we would expect — potentially — failure rates to be so high that it's not even worth measuring ... we may have to modify that protocol to give people a walkthrough, explain what it is and what it can do, let them try it a few times and then do a test after that.”

Brand Development

According to Myers, some designers argue that brand goes away as soon as an experience is voiced by an assistant like Alexa. However, plenty of designers are using platform capabilities that allow them to substitute their own voices and leverage music or sound effects. For Marshall, the discussion around brand begins with stakeholder interviews. “We have some worksheets that will give to our customers for them to fill out, so that we can learn what they're trying to project with the system.” From these worksheets, designers at Nuance create a system persona.

Marshall, Myers, Gregory, Hedges and Prayaga all leverage pre-recorded voices to express brand. The best case scenario is to leverage a well-know voice associated with the brand. For example, the Oprah Alexa Skill uses recordings of Oprah herself to read passages from her book. The *Jeopardy!* Alexa Skill uses recordings of Alex Trebek. In the case of *Jeopardy!,* Trebek opens the skill but the questions added each day are voiced by Alexa.

Hedges, while working on a Starbucks skill, spent time in the coffee shop hoping to hear something relevant. He discovered that many of the baristas say something like, “What can I get started for you?” This turn-of-phrase is uniquely Starbucks. Hedges integrated that phrasing into the skill. Vocabulary and phrases appropriate to the community can be a good way to signify brand.

Music and sound effects, especially those unified across advertising channels, can be used as well. Because of the unique nature Skills and Actions developed for smart-speaker platforms, Myers has begun using an opening and closing sound or chime to indicate when her skills are running. That unique chime or any associated music create a strong brand association — just like a commercial jingle.

Lastly, Skills and Actions developed for smart-speaker platforms will have an icon and description associated with them. These platforms and their associated assets will be discussed in another section, but these asset can help improve discoverability. Using the appropriate keywords can help a user find the right skill.

Gregory works to incorporate as much branding as possible without being over the top. And, of course, the client always has final say.

Regulatory Concerns

There are a number of regulatory issues that can impact voice experience design.

If any user information is stored, the user should be informed of how that information is stored and how it will be used. According to Myers, this is particularly relevant in Germany and other European countries. “If you are not upfront about what your skill does — whether it saves data from session to session, when it's listening or when it's not — if you don't have a privacy policy, you have lost in this market.” And, the best practice for Nuance Communications is to never store personally identifiable information in unsecure ways.

It is possible for voice experiences to be HIPPA compliant. For Marshall’s IVR systems that requires an authentication step called “caller gatekeeping” that asks the user for some piece of personally identifiable information that can be validated against a client back-end system.

By their very nature, voice experiences increase accessibility for those who are visually impaired. However, designers have created another problem that may need to be reckoned with. Suddenly, we've created an interface that may not be accessible for those who are hearing impaired. Hearing impairment affects 48 million or 14.7 percent of Americans ("Statistics and Facts about Hearing Loss | CHC"). According to Prayaga, certain chimes and tones in audible ranges can be used in lieu of voice for those who are hearing impaired.

Lastly, Marshal, Hedges and Gregory each work with their client or the client's legal department to understand what legal requirements may impact their design.

Deliverables

For voice experiences, there are two stages of design. The first stage is creative and iterative. The second stage is disciplined work that requires the designer to document, in painstaking detail, all of the exceptions a user may encounter inside of the experience. Deliverables include happy-path scripts, demos, user or task flows diagrams, storyboards, alternate phrasings, state machine diagrams, wireframes, final visual assets, and audio files.

1. Creative and Iterative

Armed with the list of requirements from the discovery phase and a script or demo of the "happy path," RAIN schedules a workshop with clients to ensure they're on board with the solution. After scripts, Gregory moves into user or task flows, "just like any user flow or task flow you've seen before." After drafting a "skeletal" design or rough outline of the experience, Marshall quickly moves into detailed design.

2. Painstaking Detail

Designers work toward fleshing out their skeletal flow diagrams with alternate paths and contingency plans. RAIN has developed their own visual language for these user flows that helps them keep track of the happy path. Additional modalities are layered in and, for Prayaga, this requires additional specifications for each modality. At Ford, they call

them domains. At this point, Vargas said the Comcast team would develop a storyboard to help designers working on difference understand how their piece fits.

For Gregory and Prayaga, detailed design involves writing voice scripts for all possible paths through the experience. According to Hedges, each utterance (something a user could say) must have 40-100 alternate phrasings to ensure that the natural language processing system can accurately map the user's intent.

Part of accurately mapping the intent is understanding which parts of what the user says should be tracked or acted on — the variables. Hedges and his team create a state machine diagram that helps the team understand how variables are used throughout the experience.

Finally, wireframes and final visual assets must be created for any visual modalities including desktop, mobile, wearable interfaces or the app store itself. And any audio files include voice, sound effects and music must be recorded, trimmed and compressed.

Summary

- The most important design activity is writing and workshopping a script for all of the core interactions.
- Modalities should be handled by tackling one first and then layering in the other modalities one at a time.
- Best practices include: Solve a problem, use vocal variety, understand context, respect privacy, be efficient, fail gracefully, build cul-de-sacs, and think multi-modal.
- Brand can be expressed through recorded voice; turn-of-phrase; and unified music and sound effects. Additionally, each skill has an icon and a series of descriptions that should be used to represent the brand.

- Regulatory issues do impact voice experiences, especially regulation around stored user data. Work with the client or the client’s legal team to understand how this impacts the experience.
- Deliverables include happy-path scripts, demos, user or task flows diagrams, storyboards, alternate phrasings, state machine diagrams, wireframes, final visual assets, and audio files.

Section 4

Developing Voice

Though this research has revealed that smart-speaker platforms are only the beginning of voice interaction, this next section will discuss in detail the two most popular of them. Amazon and Google both offer stand-alone natural language processing tools, Lex and Cloud Natural Language respectively, that can be used to develop any type of conversational interface.

Understanding the voice interaction paradigm of today sets a baseline for developing the voice experiences of the future.

4.1 The Assistants

In an analogy to the race to develop mobile apps, more than 26,000 apps have already been developed for Amazon Alexa-enabled devices and the Google Assistant, which can be found on Google Home and the Google Assistant mobile app for iOS and Android (“Amazon Alexa & Echo Stats”) (“Google Home & Assistant Stats”). Amazon and Google have developed the most popular voice assistance in the industry.

The platforms vary in both form and function. Amazon calls their apps “Skills.” Google calls them “Actions.” According to Former Amazon Tech Evangelist Liz Myers, Skills and Actions prepare assistants to understand a given context, rather than “boiling the ocean” for a response. Another major distinction between the platforms reveals itself when designers begin detailed design work. While it is imperative for designers to write at least 30 variations on what a user might say to Alexa, the Google Assistant says best practice is ten. Where Amazon is matching the user’s speech word-for-word to the designer’s plan, Google is applying machine learning. Additionally, The Google Assistant can invoke Actions through implicit direction. This means that the user does not need to know the name of a give skill to invoke it. Amazon Alexa-enabled devices may have market share, but the Google Assistant is much closer to theoretically understanding users. Though he prefers Alexa-enabled devices for their smart home

experience, Comcast Senior Manager Design, Voice User Interface Elijah Vargas said, “Alexa is more of a dictionary and a Google is your encyclopedia set.”

Amazon Alexa

As of December 2017, 24,385 skills have been developed for the Alexa platform (“Amazon Echo & Alexa Stats”). The team at Amazon recommends a three-step design approach: Define the experience, design what the user says, then design how Alexa responds.

In-depth, up-to-date resources on the best practices for developing skills can be found at https://developer.amazon.com/designing-for-voice/.

Defining the experience is as simple as defining the purpose and capabilities of the skill. Amazon recommends answering the question, “What does the user want to do?” and “What will they be doing before, during and after?” Then, designers should write user stories that define what the user can do, the ways they can invoke the skill and where any back-end information will come from.

As recommended by the industry experts interviewed for this research, Amazon recommends designing the happy path, the shortest route from initiating the skill to solving the problem. After scripting the happy path, designers can move into alternate paths and decision trees. Will the user need to make a decision? Could that change their path through the experience? Any behind-the-scenes decisions and logic should be captured at this point. Next, Amazon recommends defining intents, utterances and slots. This jargon is specific to Amazon Alexa-enabled experiences.

- **Intents** are the unique things that your skill is able to do. It’s a derivation of “What is the user’s intent?” Intents are invoked by utterances.
- **Utterances** are words, phrases or sentences spoken by the user to engage and fulfill an intent. Utterances are made up of keywords, natural speech sounds, and slots for any information that can vary. Amazon encourages designing at least 30

utterances, but Hedges recommends 40-100 for the best experience. Amazon has built in a number of utterances such as “cancel,” “stop,” and “help”. There are tools available to help designers write and expand their list of utterances. See the appendix for details.

- **Slots** are the variable parts of an utterance. Amazon has built in some slot values such as city or state. Otherwise, the designer will need to define the possible values of the slot.

Next, Amazon recommends designing how Alexa will respond. The highlights of the recommendations from Amazon include be brief, speak and write naturally, prompt with guidance, use conversation markers, add variety and pay special attention to lists.

All responses should pass the one-breath test; a designer should be able to speak the response in a single breath. Amazon recommends against prompting with a menu of options. Instead, the utterances should be robust enough to handle anything they might say naturally. In the spirit of speaking and writing naturally, designers should use contractions and make sure to use a TTS simulator to test how the response will sound.

At times, it will be necessarily to include a list of either items or options. Amazon recommends paying special attention to lists, ensuring their brevity, that they are organized in parallel structure. It is also possible to fine-tune the pacing of the list using SSML. It’s also helpful to tell the user where they are in a list, for example using an interjection that says, “Last one!”. The action or question directed at the user should always come at the end of a response. The user won’t remember your question or know how to respond if it precedes any length of text.

Alexa allows designers to include adaptive prompts. These are responses that change with the user’s experience level. For example, as the user learns how to use the skill, you can prompt them without suggestion or hint text. While adaptive prompts are recommended for the voice responses, Amazon writes that it’s ok to be predictable for visuals. No adaptive prompts are needed here. Both the Echo Show and the Echo Spot

have screens that designers can leverage if they choose. All graphics should pass the distance test. For the Echo Show, the designer should be able to see them from seven feet. For the Echo Spot, the designer should be able to see them from five feet.

Amazon has provided templates for these graphics: https://developer.amazon.com/docs/custom-skills/display-interface-reference.html#display-template-reference.

Any pre-recorded audio, including sound effects and music, must be prepared for skill development. The audio should be hosted encoded as MPEG version 2 with a 48 kbps bit rate and 16,000 Hz sample rate to be compatible with Alexa-enabled devices. The audio files must be hosted via an internet-accessible, secure domain. According to SaySpring, the domain must have a valid, trusted SSL certificate ("How to Include MP3 Audio Files in Your Voice Design"). Amazon allows short-form and long-form audio of less than 90 and greater than 90 seconds respectively.

According to RAIN Vice President of Emerging Experiences Greg Hedges, certification can take four to five weeks. Submission to the Skill Store requires the following assets:

- Public Name (2-50 characters)
- One Sentence Description (up to 160 characters
- Detailed Description (up to 4,000 characters)
- Example Phrases (sample utterances to appear on the skill detail card on the Alexa app)
- Small Skill Icon (108x108 px png or jpg)
- Large Skill Icon (512x512 px png or jpg)
- Category
- Keywords (up to 30)
- Privacy Policy URL
- Terms of Use URL

Actions on Google

As of January 2018, 1,719 apps have been developed for the Actions on Google platform (“Google Home & Assistant Stats”). Designing a Google Action can be broken down into two distinct phases. First, designers should define the action and its attributes. Next, the designer should define the dialog.

In-depth, up-to-date resources on the best practices for developing Actions can be found at https://console.actions.google.com/.

According to Google, some actions are inherently better suited for voice. These actions should be quick but compellingly useful. The first step toward developing a Google Action is to pick at least one but no more than a few related use cases. Google cautions designers to spend time picking a name for their Action. Each name can only be used once and, because users must say the name of the app to explicitly invoke it, it should be easy to remember and easy to say, even for users with accents. With this, users can also begin thinking about the phrases that can be used to implicitly invoke the action. Based on these phrases, the Google Assistant asks Actions on Google to invoke the best app to fulfill the intent.

Next, Google encourages users to begin thinking about “surfaces” immediately. The Google Assistant is available for Google Home, but it’s also available as a mobile app for both iOS and Android. As recommended by both the subject matter experts interviewed for this research and the Amazon Alexa team, designers should focus on the happy path. Once satisfied with the happy path, designers can flesh out additional paths that lead to the happy path. Google encourages designers at this stage to consider conversation repair scenarios that will facilitate progression when the user does something unexpected.

Google recommends creating actions, creating intents, defining their entities and scripting an appropriate response. The Google Actions team offers a number of templates that allow anyone interested to develop trivia skills, personality quizzes and

flashcards. Because Google Actions are heavily integrated with DialogFlow, any Action that does not require a call to a separate database can be created by a reasonably tech-savvy person without the help of a developer. However, they do require understanding of a couple of key concepts.

- **Actions** let users accomplish tasks with your app. Actions can be invoked using an explicit invocation (e.g. “Talk to <app name>”) or an implicit invocation (e.g. the user can invoke a skill without knowing its name).
- **Intents** are words, phrases or sentences that define the grammar of what a user must say to trigger the action — also called fulfillment. Google recommends that designers write 5-10 intents. Google will expand those intents with natural variations. A good experience will also include a fallback intent or an intent that triggers if no match is found. Intents trigger responses.
- **Parameters** are variables. The values of those variables are called entities. Entities that are extracted from user intents must be mapped back to the associated parameter so that the system will know which parameter it is dealing with.
- **Entities** are words, phrases or sentences said by the user that are extracted as parameter values. Any important data you want to get from a user’s request will have a corresponding entity. There are system-, developer- and user-defined entities (“Entities | DialogFlow”).
- **Responses** are the words that are spoken by Google Assistant to the user as a reaction to their intent. Simple responses can be composed in DialogFlow as long as they do not require fulfillment.
- **Fulfillment** often requires back-end services. Parameters captured from intents spoken by the user are passed along to do their thing with existing back-end services and return a relevant result.

After designing the experience, Google offers a short checklist to help ensure that your Action will meet their approval standards. First, all Actions should include greetings and goodbyes. Greetings, especially in for implicitly invoked skills, should be descriptive,

giving the user enough information to understand what they've stumbled into. Goodbyes help the understand that the Action has closed and should leave the user with a good impression. When it comes to conversational dialog, Google recommends that the experience take turns with the user and sound natural. Lastly, Google has stressed the importance of conversation repair. Errors can be prevented by expecting variation in user responses — and communicating that to Google through a robust set of intents. When errors do occur, attempt to assist by providing helpful reprompts or by pivoting to another question. Users can get frustrated if they get stuck in a loop of the same question. Designers can be prepared to help at any time by adding fallback and help intents. Lastly, Google recommends letting users replay information they may have missed.

More information about response types, including using surfaces, can be found here: https://developers.google.com/actions/assistant/responses.

Submission to the Google Actions platform requires the following assets:

- "You can use this app to …" Phrases
- Choice of Assistant Voice
- Short Description: 100 characters
- Full Description: 4,000 characters
- Sample Invocations
- Large Banner (1920 x 1080)
- Small Logo (192x192)

4.2 Testing

A number of tools and full-service offerings have sprung up around the need for testing both Skills and Actions. There are two types of testing.

As mentioned in the design section of this document, there is a need for usability testing before development begins. Similar to the paradigm of https://www.usertesting.com/, Pulse Labs promises to match available testers to Skills or Actions in need of testing.

Pulse Labs, founded by Abhishek Suthan — a former Vice President at Goldman Sachs, is still in its nascent stages, formulating it's offering and pricing. More information about Pulse Labs can be found at https://pulselabs.ai.

Both the Alexa and Actions on Google Developer Console offer simulators that allow testing at any stage of the development process. Designers can type raw text or SSML to hear that responses will sound like spoke by the assistant. This can be used for fairly convincing Wizard of Oz testing, especially if the designer is willing to put in a little work to get a few simple interactions strung together. These simulators also enable the development team can use build-in debugging tools to see how the code is responded to voiced commands. The simulators can be found at https://developer.amazon.com/alexa/console/ask or https://console.actions.google.com/ respectively.

If moving into final stage usability testing, draft Skills and Actions can be accessed via a "developer" device — a device linked to either your Amazon or Google accounts, specifically for testing. Or, if texting an Alexa skill, Echosim has rolled out a slick-looking device simulator that can be accessed via a web browser. Echosim was born at a 2015 hackathon, according to Alexa developer Glenn Cameron. "Its simplicity makes it easy for anyone to understand what an Echo is and what it does without having to explain Alexa's unique UX" ("Introducing Echosim.io"). More information can about Echosim can be found at https://echosim.io/.

As of publication of this research, no full-service quality assurance agencies have sprung up the handle the demand for intensive pre-release testing.

Section 5

Appendix

5.1 Glossary

Action (GOOGLE): These Let users accomplish tasks with your app. Actions can be invoked using an explicit invocation (e.g. “Talk to <app name>”) or an implicit invocation (e.g. the user can invoke a skill without knowing its name).

Adaptive Prompts (AMAZON): As a user gains experience using a given voice interface, the prompts should become more concise. For example, on the tenth time someone asks their real-life assistant for something, they begin to develop a short-hand. They don’t have to describe every detail the same way they did the first time. The Amazon Alexa ecosystem supports adaptive prompts.

“Ask” (AMAZON): Paired with the name of a Skill, the “ask” command spoken by the user allows the user to access a specific Skill on an Amazon Alexa-enabled device.

Confirmation (AMAZON): At times, it can be essential to confirm with the user that what was spoken was interpreted correctly. This can be done in two ways. First, implicit confirmation or landmarking can be used when the stakes are low. This repeats what the user said, in part, to confirm and orient them to the next step or response. Explicit confirmation can be used when the stakes are high. Explicit confirmation would ask something like, “Did you say … ?”

Conversational Interfaces: Similar to a voice user interface, or VUI, conversational interfaces either comprise solely conversational interactions or they include conversational interactions as part of a multimodal experience. Unlike VUI, conversational interfaces can include typed, tapped and motion interaction in the form of a conversation.

Earcon: Similar to an icon (or eye-con), an earcon is a sound that is used to represent a program or function. The program or function can be represented with a short clip of music, a sound effect or voice.

Entities (GOOGLE): Entities are words, phrases or sentences said by the user that are extracted as parameter values. Any important data you want to get from a user's request will have a corresponding entity. There are system-, developer- and user-defined entities ("Entities | DialogFlow").

Fallback Intent (GOOGLE): An intent that is invoked if what the user said does not match any of the regular intents of the Action.

Fulfillment (GOOGLE): Refers to the act of passing a parameter to a service outside of the Google Assistant environment to return a result for the user, this fulfilling their request.

Happy Path: Also referred to as the "red thread," this is the shortest, most efficient path for a user to achieve their goal with a given application.

Intent (AMAZON): The unique things that your skill is able to do. It's a derivation of "What is the user's intent?" Intents are invoked by utterances.

Interjection: Also known as a speechcon, are special words and phrases that TTS systems can pronounces more expressively. these exclamations can be included using SSML ("Speechcon Reference (Interjections)").

Intent (GOOGLE): Words, phrases or sentences that define the grammar of what a user must say to trigger the action. Google recommends that designers write 5-10 intents. Google will expand those intents with natural variations. A good experience will also include a fallback intent or an intent that triggers if no match is found. Intents trigger responses.

Invocation (AMAZON): Skills on Amazon Alexa-enabled devices can be explicitly invoked by the user by stating "Ask <Skill name> …" or "Tell <Skill name> …". By pairing these invocation phrases with a command or intent, the

user can achieve their objective with one shot — this is referred to as full intent. Additionally, a user may express either a partial intent or no intent at all.

Invocation (GOOGLE): Actions can be implicitly or explicitly invoked by the user. If the user says “Talk to <action name>,” they are explicitly invoking a specific Action. If the user says, “Do you have any good spaghetti recipes?”, they may be implicitly invoking a cooking Action that has defined spaghetti recipes as one of their implicit Actions.

IVR: Interactive Voice Response is a technology that enables telephone callers to interact with computers via their voice or a touch-tone keypad. IVR systems that enabled touch-tone calling emerged in the 1960s. By the 1980s, IVR systems recognized simple voice commands (Luu).

Landmarking: This is a type of prompt, or response spoken to the user by the voice user interface. Landmarking indicates to the user that the assistant heard them correctly, orients them to the interaction and helps instill trust (“Voice Design Best Practices (Legacy)”). For example, “Here’s a recipe for spaghetti. Start by boiling water ...” rather than “Start by boiling water …”

Multimodal: Voice experiences can be multimodal, meaning they can present responses via many modalities or surfaces. This enables the user to interact using the medium that is the best fit for their goals and context.

Modality: Refers to one medium, surface or interface for a given experience.

One-breath test: This test is meant to be used by designers while writing the responses that will be spoken by a text-to-speech application. Responses should not require more than one full breath to be spoken. This indicates that responses should be concise.

One-Shot Model: A model of interaction with a voice interface wherein the user’s objective is achieved with only one utterance, without conversation or back-and-forth.

Parameters (GOOGLE): Parameters are variables. The values of those variables are called entities. Entities that are extracted from user intents (spoken commands) must be mapped back to the associated parameter so that the system will know which variable (parameter) it is dealing with.

Phonemes: The smallest phonetic unit in a language that is capable of conveying a distinction in meaning, as the m of mat and the b of bat in English (“Phoneme”).

Prompt: A prompt is a type of response. The intention of a prompt is to engage the user to utilize an existing utterance or parameter. Prompts can be open-ended, menu style, use implicit confirmation (also called landmarking), or re-prompt the user.

Red Thread: Also referred to as the “happy path,” this the shortest, most efficient path for a user to achieve their goal with a given application.

Response: A response refers to any feedback given to the user based on a prompt. Responses can be presented to the user via any or all modalities.

Skill (AMAZON): The formal name for an app on an Amazon Alexa-enabled device.

Slots (AMAZON): The variable parts of an utterance. Amazon has built in some slot values such as city or state. Otherwise, the designer will need to define the possible values of the slot.

Speechcon: Also known as interjections, are special words and phrases that TTS systems can pronounces more expressively. these exclamations can be included using SSML (“Speechcon Reference (Interjections)”).

SSML: Speech synthesis markup language is an XML-based markup language for speech synthesis tools.

Surface (GOOGLE): A Google Home or an Android phone can each be a surface for the Google Assistant. Surface refers to the modalities and devices for which your Action is capable of providing interaction.

System Persona: Many voice user interface designers create system personas. This is a tool meant to personify the brand or interface, rather than the user. This helps the designer understand whether their voice interface should be professional, stern, funny or sassy in a given conversation.

“Talk To” (GOOGLE): Paired with the name of an Action, the “talk to” command spoken by the user allows the user to access a specific Action on a Google Assistant enabled device.

“Tell” (AMAZON): Paired with the name of a Skill, the “tell” command spoken by the user allows the user to access a specific Skill on an Amazon Alexa-enabled device.

Trigger phrase (GOOGLE): The word, phrase or sentence that is used to trigger a smart speaker device. The default trigger phrase for Google Assistant-enabled devices is “Ok Google.” The default trigger phrase for Amazon Alexa-enabled devices is “Alexa”.

TTS: Text to speech tools convert text into a synthesized speech output.

Use Case: A use-case is a detailed description of how a product or service will be used. This could include a description of the actors, preconditions, triggers, success scenarios (happy paths) or alternative flows (“Use Cases”).

User Story: These are short, descriptive statements that describe functionality from the perspective of a user. For example: “As a user, I can order a car to pick me up at my house.”

Utterances (AMAZON): Words, phrases or sentences spoken by the user to engage and fulfill an intent. Utterances are made up of keywords, natural speech sounds, and slots for any information that can vary. Amazon encourages designing at least 30 utterances, but Hedges recommends 40-100 for the best experience. Amazon has built in a number of utterances such as “cancel,” “stop,” and “help”. There are tools available to help designers write and expand their list of utterances.

Voiceprint: This usually refers to a visual representation of a word, phrase or sentence spoken by a user. With recent advances in technology, this voiceprint can be used to authenticate or identify a specific user. This is especially useful for devices in multi-user or high-traffic areas.

VUI: A voice user interface, often abbreviated VUI, is an interface that either comprises solely voice interactions or includes voice interactions as part of a multimodal experience.

Wake word (AMAZON): For Amazon Alexa-enabled devices, the wake words include “Alexa,” “Computer,” and “Amazon”. According to former Amazon Tech Evangelist Liz Myers, the wake work prompts Amazon smart speakers to turn on the web-connected microphone. “You see it in the blue ring and then she's ready to listen up to the cloud.”

Wizard of Oz Testing: This type of testing is intended to enable designers to test a voice interface before development activities have begun. Recordings of synthesized speech can be saved on a sound board or pinned to a slide in a Apple Keynote deck and triggered when appropriate by a someone “behind the curtain.”

5.2 Resources

Design

Happy Path

The happy path is as simple as writing a back-and-forth script. All that is required is a tool that captures text input. Scriptwriting tools were also mentioned as an option, but no specific tool was named.

- Microsoft Word
- TextEdit

Diagramming

Any of the following tools can be used to create a robust user flow or state-machine diagram of the conversation flow.

- Adobe Illustrator
- LucidChart
- Microsoft Powerpoint
- Microsoft Excel
- Omnigraffle
- Sketch
- Vizio

Managing Team

When tracking deliverables across many modalities, Comcast's Senior Manager of Design, Voice User Interface Elijah Vargas said that his team uses Confluence. The ability to upload images or other files within the project management tool is essential.

- Atlassian Confluence

Prototyping

While testing can be as simple as staging a team member behind a curtain with a script, it's helpful to test out how responses will sound using a text-to-speech simulator. These simulators are built into Skills and Actions workflows and are readily available on the

web. Additionally, it can be useful to queue a recorded TTS clip using either a soundboard (high-tech option) or Apple Keynote (low-tech).

- Apple Keynote
- Soundboard
- TTS Simulator

Development

The links below reference tools, frameworks or specifications that can be useful in the technical development of a new Action or Skill. These are specific to smart-speaker applications.

- Actions on Google Developer Console (https://developers.google.com/actions/)
- Alexa Developer Console (https://developer.amazon.com/alexa/)
- DialogFlow (https://dialogflow.com/)
- Jovo Framework (https://github.com/jovotech/jovo-framework-nodejs)
- SaySpring (https://www.sayspring.com/)
- SSML Markup Specification (https://www.w3.org/TR/speech-synthesis/)
- Utterance Expander (http://www.makermusings.com/amazon-echo-utterance-expander/)
- Voxa Framework (http://voxa.readthedocs.io/en/stable/)

5.3 Sources

Stakeholder Interviews

Date	Name	Title
Feb. 20, 2018	Dave Cowing	Partner, Business Lead for TandemSeven Consulting
Feb. 22, 2018	Elizabeth Srail	Vice President, User Experience Practice for TandemSeven Consulting
March 3, 2018	David Clark	Partner, Business Lead for Cora Journey 360

Interviews with Subject Matter Experts

Date	Name	Title	Organization
March 6, 2018	Bill Marshall	User Interface Designer	Nuance Communications
March 9, 2018	Kathryn Whitenton	Digital Strategy Manager	Nielsen Norman Group
March 13, 2018	Liz Myers	Tech Evangelist	Amazon (Formerly)
March 14, 2018	Vinita Atre	HCI Graduate Student	University of Maryland, College Park
March 15, 2018	Elijah Vargas	Senior Manager, Voice User Interface	Comcast
March 20, 2018	Shyamala Prayaga	Voice Interaction Designer	Ford Motor Company
March 20, 2018	Bradley Gregory	Voice User Experience Designer	RAIN Agency
March 22, 2018	Greg Hedges	Vice President of Emerging Experiences	RAIN Agency

Works Cited

"701 Translator." IBM Archives, www-03.ibm.com/ibm/history/exhibits/701/701_translator.html.

"Actions on Google | Google Developers." Google, developers.google.com/actions/tools/.

"Amazon Echo & Alexa Stats." Voicebot, 16 Mar. 2017, www.voicebot.ai/amazon-echo-alexa-stats/.

"Introducing Echosim.io – A New Online Tool Built by the Community, for the Community." Amazon Developer Services, developer.amazon.com/post/Tx3BB1JHNS1TDTS/Introducing-Echosim-io-A-New-Online-Tool-Built-by-the-Community-for-the-Communit.

"Entities | Dialogflow." Dialogflow, dialogflow.com/docs/entities.

Fortune Editors. "The Exec Behind Amazon's Alexa: Full Transcript of Fortune's Interview." Fortune, Time, Inc., 14 July 2016, fortune.com/2016/07/14/amazon-alexa-david-limp-transcript/.

"Google Home & Assistant Stats." Voicebot, 7 Apr. 2017, www.voicebot.ai/google-home-google-assistant-stats/.

Greenemeier, Larry. "20 Years after Deep Blue: How AI Has Advanced Since Conquering Chess." Scientific American, 2 June 2017, www.scientificamerican.com/article/ 20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/.

Hancox, P.J. "A brief history of Natural Language Processing." University of Birmingham School of Computer Science, www.cs.bham.ac.uk/~pjh/sem1a5/pt1/pt1_history.html.

"Hard Disk Drive Morph." Processing Power Compared, Experts Exchange, 2015, http://www.pages.experts-exchange.com/processing-power-compared.

Hedges, Greg. “Forget Burger King's Google Home Gag. Here's the Real Way to Win Voice.” Adweek, 18 Apr. 2017, www.adweek.com/creativity/forget-burger-kings-google-home-gag-heres-the-real-way-to-win-voice/

“How to Include MP3 Audio Files in Your Voice Design.” Sayspring, 23 Jan. 2018, www.sayspring.com/blog/include-mp3-audio-files-voice-design/.

Kim, Eugene. “The inside Story of How Amazon Created Echo, the next Billion-Dollar Business No One Saw Coming.” Business Insider, 2 Apr. 2016, www.businessinsider.com/the-inside-story-of-how-amazon-created-echo-2016-4.

Luu, Anh. “The History of Interactive Voice Response (IVR).” Call Tracking Service, Phonexa, 22 Sept. 2016, phonexa.com/the-history-of-interactive-voice-response-ivr/.

Markoff, John. “On 'Jeopardy!' Watson Win Is All but Trivial.” The New York Times, 16 Feb. 2011, www.nytimes.com/2011/02/17/science/17jeopardy-watson.html.

Perez, Sarah. “Voice-Enabled Smart Speakers to Reach 55% of U.S. Households by 2022, Says Report.” TechCrunch, 8 Nov. 2017, techcrunch.com/2017/11/08/voice-enabled-smart-speakers-to-reach-55-of-u-s-households-by-2022-says-report/.

"Phoneme". Dictionary.com Unabridged. Random House, Inc. 15 Apr. 2018. <Dictionary.com http://www.dictionary.com/browse/phoneme>.

Pieraccini, Roberto. “From AUDREY to Siri: Is speech recognition a solved problem?” International Computer Science Institute at Berkeley, http://www.icsi.berkeley.edu/pubs/speech/audreytosiri12.pdf.

Pinola, Melanie. “Speech Recognition Through the Decades: How We Ended Up With Siri.” PCWorld, PCWorld, 2 Nov. 2011, www.pcworld.com/article/243060/speech_recognition_through_the_decades_how_we_ended_up_with_siri.html.

“Pioneering Speech Recognition.” IBM100 - Pioneering Speech Recognition, www-03.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/.

Rao, Gauri, et al. “Natural Language Query Processing on Dynamic Databases Using Semantic Grammar.” International Journal on Computer Science and Engineering, vol. 2, no. 2, 2010, pp. 219–223, www.enggjournals.com/ijcse/doc/IJCSE10-02- 02-20.pdf.

Schank, Roger C. “Language and Memory.” Cognitive Science, vol. 1, no. 4, 1980, pp. 243–284, www.web.stanford.edu/class/linguist289/schank80.pdf.

Seque Quality Control Team. “User Stories vs. Use Cases: Pros and Cons for Agile Development.” Segue Technologies, 24 June 2016, www.seguetech.com/user-stories-vs-use-cases-pros-cons-agile-development/.

Soper, Taylor. “More than 8M People Own an Amazon Echo as Customer Awareness Increases 'Dramatically'.” GeekWire, 5 July 2017, www.geekwire.com/2017/8-million-people-amazon-echo-customer-awareness-increases-dramatically/.

“Speechcon Reference (Interjections): English (US).” Speechcon Reference (Interjections): English (US) | Custom Skills, developer.amazon.com/docs/custom-skills/speechcon-reference-interjections-english-us.html.

“Statistics and Facts about Hearing Loss | CHC.” Center for Hearing and Communication, chchearing.org/facts-about-hearing-loss/.

Terras, Melissa M., et al. Defining Digital Humanities: A Reader. Routledge, 2016, https://books.google.com/books?id=xAYpDAAAQBAJ&pg.

“Use Cases.” Usability.gov, Department of Health and Human Services, 9 Oct. 2013, www.usability.gov/how-to-and-tools/methods/use-cases.html.

“Voice Design Best Practices (Legacy).” Custom Skills, developer.amazon.com/docs/custom-skills/voice-design-best-practices-legacy.html.

Winograd, Terry. “SHRDLU.” Stanford HCI Group, www.hci.stanford.edu/winograd/shrdlu/.

Xfinity. “The X1 Voice Remote Overview.” Xfinity Help & Support, 1 Mar. 2018, www.xfinity.com/support/articles/get-to-know-xr11-remote.