O'REILLY®

Design for Voice Interfaces

Building Products that Talk

Laura Klein

Short. Smart. Seriously useful.

Free ebooks and reports from O'Reilly at **oreil.ly/fr-design**

Free ebooks, reports and other articles on UX design, data-informed design, and design for the IoT. Get insights from industry experts and stay current with the latest developments from O'Reilly.

O'REILLY®

Design for Voice Interfaces

Laura Klein

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Design for Voice Interfaces
by Laura Klein

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Acquisitions Editor: Mary Treseler
Editor: Angela Rufino
Production Editor: Matthew Hacker
Copyeditor: Octal Publishing, Inc.

Interior Designer: David Futato
Cover Designer: Randy Comer
Illustrator: Rebecca Demarest

October 2015: First Edition

Revision History for the First Edition
2015-10-12 First Release

978-1-4919-3458-6

[LSI]

Table of Contents

CHAPTER 1

Design for Voice Interfaces

The way we interact with technology is changing dramatically again. As wearables, homes, and cars become smarter and more connected, we're beginning to create new interaction modes that no longer rely on keyboards or even screens. Meanwhile, significant improvements in voice input technology are making it possible for users to communicate with devices in a more natural, intuitive way.

Of course, for any of this to work, designers are going to need to learn a few things about creating useful, usable voice interfaces.

A (Very) Brief History of Talking to Computers

Voice input isn't really new, obviously. We've been talking to inanimate objects, and sometimes even expecting them to listen to us, for almost a hundred years. Possibly the first "voice-activated" product was a small toy called Radio Rex, produced in the 1920s (Figure 1-1). It was a spring-activated dog that popped out of a little dog house when it "heard" a sound in the 500 Hz range. It wasn't exactly Siri, but it was pretty impressive for the time.

The technology didn't begin to become even slightly useful to consumers until the late 1980s, when IBM created a computer that could kind of take dictation. It knew a few thousand words, and if you spoke them very slowly and clearly in unaccented English, it would show them to you on the screen. Unsurprisingly, it didn't really catch on.

Figure 1-1. Radio Rex.

And why would it? We've been dreaming about perfect voice interfaces since the 1960s, at least. The computer from Star Trek understood Captain Kirk perfectly and could answer any question he asked. HAL, the computer from *2001: A Space Odyssey*, although not without one or two fairly significant bugs, was flawless from a speech input and output perspective.

Unfortunately, reality never started to approach fiction until fairly recently, and even now there are quite a few technical challenges that we need to take into consideration when designing voice interfaces.

Quite a bit of progress was made in the 1990s, and voice recognition technology improved to the point that people could begin using it for a very limited number of everyday tasks. One of the first uses for the technology were voice dialers, which allowed people to dial up to ten different phone numbers on their touch-tone phones just by

speaking the person's name. By the 2000s, voice recognition had improved enough to enable Interactive Voice Response (IVR) systems, which automated phone support systems and let people confirm airplane reservations or check their bank balances without talking to a customer-support representative.

It's not surprising that when Siri first appeared on the iPhone 4S in 2011, many consumers were impressed. Despite her drawbacks, Siri was the closest we had come to asking the Star Trek computer for life-form readings from the surface of the planet. Then IBM's supercomputer, Watson, beat two former champions of the gameshow *Jeopardy* by using natural-language processing, and we moved one step closer to technology not just recognizing speech, but really understanding and responding to it.

Toys have also come a long way from Radio Rex. The maker of the iconic Barbie doll, Mattel, unveiled a prototype of Hello Barbie in February of 2015 (Figure 1-2). She comes with a WiFi connection and a microphone, and she can have limited conversations and play interactive, voice-enabled games.

Figure 1-2. Hello Barbie has a microphone, speaker, and WiFi connection.

From recognizing sounds to interpreting certain keywords to understanding speech to actually processing language, the history of designing for voice has been made possible by a series of amazing technological breakthroughs. The powerful combination of speech

recognition with natural-language processing is creating huge opportunities for new, more intuitive product interfaces.

Although few of us are worried about Skynet (or Barbie) becoming sentient (yet), the technology continues to improve rapidly, which creates a huge opportunity for designers who want to build easier-to-use products. But, it's not as simple as slapping a microphone on every smart device. Designers need to understand both the benefits and constraints of designing for voice. They need to learn when voice interactions make sense and when they will cause problems. They need to know what the technology is able to do and what is still impossible.

Most important, everybody who is building products today needs to know how humans interact with talking objects and how to make that conversation happen in the most natural and intuitive way possible.

A Bit About Voice and Audio Technology

Before we can understand how to design for voice it's useful to learn a little bit about the underlying technology and how it's evolved. Design is constrained by the limits of the technology, and the technology here has a few fairly significant limits.

First, when we design for voice, we're often designing for two very different things: voice inputs and audio outputs. It's helpful to think of voice interfaces as a conversation, and, as the designer, you're responsible for ensuring that both sides of that conversation work well.

Voice input technology is also divided into two separate technical challenges: recognition and understanding. It's not surprising that some of the very earliest voice technology was used only for taking dictation, given that it's far easier to recognize words than it is to understand the meaning.

All of these things—recognition, understanding, and audio output—have progressed significantly over the past 20 years, and they're still improving. In the 90s, engineers and speech scientists spent thousands of hours training systems to recognize a few specific words.

These are known as "finite state grammars" because the system is only capable of recognizing a finite set of words or phrases. You can

still see a lot of these in IVRs, which are sometimes known as "those annoying computers you have to talk to when you call to change your flight or check your bank balance."

As the technology improves, we're building more products with "statistical language models." Instead of a finite set of specific words or phrases, the system must make decisions about how likely it is that a particular set of phonemes resolves to a particular text string. In other words, nobody has to teach Siri the exact phrase "What's the weather going to be like in San Diego tomorrow?" Siri can probabilistically determine how likely it is that the sounds coming out of your mouth translate into this particular set of words and then map those words to meanings.

This sort of recognition, along with a host of other machine-learning advances, has made Natural-Language Processing (NLP) possible, although not yet perfect. As NLP improves, we get machines that not only understand the sounds we're making but also "understand" the meaning of the words and respond appropriately. It's the kind of thing that humans do naturally, but that seems borderline magical when you get a computer to do it.

VUI versus GUI: What's New and What's Not

These recent technological advances are incredibly important for voice user interface (VUI) designers simply because they are making it possible for us to interact with devices in ways that 10 or 20 years ago would have been the stuff of science fiction. However, to take full advantage of this amazing new technology, we're going to have to learn the best way to design for it. Luckily, a lot of the things that are core to user experience (UX) design are also necessary for VUI design. We don't need to start from scratch, but we do need to learn a few new patterns.

The most important part of UX design is the user—you know, that human being who should be at the center of all of our processes—and luckily that's no different when designing for voice and audio. Thomas Hebner, senior director of UX design practice and professional services product management at Nuance Communications, has been designing for voice interfaces for 16 years. He thinks that the worst mistakes in voice design happen when user goals and business goals don't line up.

Great products, regardless of the interaction model, are built to solve real user needs quickly, and they always fit well into the context in which they're being used. Hebner says, "We need to practice contextually aware design. If I say, 'Make it warmer' in my house, something should know if I mean the toast or the temperature. That has nothing to do with speech recognition or voice design. It's just good design where the input is voice."

This is important. Many things about designing for voice—understanding the user, knowing the context of use, and ensuring that products are both useful and usable—are all exactly the same as designing for screens, or services, or anything else. That's good news for designers who are used to building things for Graphical User Interfaces (GUIs) or for systems, because it means that all of the normal research and logic skills transfer very nicely when incorporating speech into designs. If you understand the basic User-Centered Design process and have applied it to apps, websites, systems, or physical products, many of your skills are completely transferrable.

Yet, there are several VUI-specific things that you won't have run into when designing for other sorts of interactions, and they're important to take into consideration.

Conversational Skills

Content and tone are important in all design, but when designing for speech output, it takes on an entirely new meaning. The best voice interface designs make the user feel like she's having a perfectly normal dialog, but doing that can be harder than it sounds. Products that talk don't just need to have good copy; they must have good conversations. And it's harder for a computer to have a good conversation than a human.

Tony Sheeder, senior manager of user experience design at Nuance Communications, has been with the company for more than 14 years and has been working in voice design for longer than that. As he explains it:

> Each voice interaction is a little narrative experience, with a beginning, middle and an end. Humans just get this and understand the rules naturally—some more than others. When you go to a party, you can tell within a very short time whether another person is easy

to talk to. Until recently, speech systems were that guy at the party doing everything wrong, and nobody wanted to talk to them.

While many early voice designers have a background in linguistics, Sheeder's background was originally writing scripts for interactive games, and it helped him write more natural conversations. But, designing for voice communication wasn't always successful. Early voice interfaces often made people uncomfortable because the designers felt as if people would need explicit instructions. They'd say things like, "Do you want to hear your bank balance? Please, say yes or no." This violates basic rules of conversation. Sheeder felt that these interfaces made people feel strange because "the IVR would talk to you like it was human, but would instruct you to talk to it like a dog. It was like talking to a really smart dog."

Designing for better conversational skills

Many designers argue that copywriting is an integral part of the user experience, and we should be better at it. That's absolutely the case for voice and speech design. If you want to incorporate voice interactions in your products, you're going to need to learn to make them sound right, and that means learning a few important rules.

Keep it short, but not too short

Marco Iacono, who designs products at Viv Labs,, explains, "When using text-to-speech, the experience can become frustrating if the system is too chatty. Especially in hands-free scenarios, the system must be concise and the user should control the pace of the interaction." In part, that can mean writing dialogs that are short, but not too short. Marco knows what he's talking about. Before his present position at Viv Labs, he spent several years as a Siri EPM at Apple where he worked on iOS, CarPlay and Apple Watch.

Written language is fundamentally different from spoken. When you first start writing dialogs, you might find that they sound stilted or just too long when spoken out loud by the product. That's normal. You want to keep all utterances much shorter than you'd expect. If you don't, people will become frustrated and begin cutting off the system, potentially missing important information.

On the other hand, you need to be careful not to omit anything really critical. Sheeder talked about the early days of voice

design for call-center automation, when the entire goal was to keep everything as short as possible. "There was a belief that shaving 750 milliseconds off a call would increase efficiency. But, by shaving off connector words and transitions, it actually increased the cognitive load on the user and lowered perceived efficiency." When the responses became too fast, it put more pressure on listeners, and they would grow frustrated or confused because they couldn't process the information. It ended up making the call centers less efficient.

Create a personality

People treat things that talk back to them as humans, and humans (most of them, anyway) have fairly consistent personalities. The same is true of VUIs. Siri has a different personality from Microsoft's Cortana, and they're both different from the Amazon Alexa.

Karen Kaushansky, director of experience at a stealth startup, has worked in voice technology since she began working at Nortel in 1996. She explains that successful voice interfaces have personas that are interesting, but also goal-based. "Are you looking to get through tasks quickly? To encourage repeat engagement? Different voice personas have different effects for the user."

Having a consistent personality will also help you to design better dialogs. It helps you make decisions about how your interface will talk to the user. In many ways, a voice persona is similar to a style guide for a visual product. It can help you decide what tone and words you should use. Will your interface be helpful? Optimistic? Pushy? Perky? Snarky? Fun? Again, it all depends on what the goals are for your product and your user. Whatever the choice, remember that both you and your users are going to have to live with this particular interface for a very long time, so make sure it's a personality that doesn't become grating over time.

One thing to consider when you're building a personality is how human you're going to make it. Marco Iacono warns that, "There's a sliding scale from purely functional to anthropomorphic. As you get closer to the anthropomorphic end of the scale, user expectations grow tremendously. Instantly, people expect it to understand and do more." The risk of making your product's

personality seem very human is that your users might be disappointed and frustrated as soon as they find the limitations of the system.

Listen to yourself
: To ensure that your conversations sound natural and efficient (not irritating), you're going to need to do a lot of testing. Of course, you should be usability testing your designs, but before you even get there, you can begin to improve your ability to write for voice interfaces. Abi Jones, an interaction designer at Google who does experimental work with voice interfaces and the Internet of Things (IoT), suggests role playing the voice UI with someone else in order to turn it into a real dialog and listen to how it sounds. She then uses accessibility tools to listen to her computer reading the dialog.

Of course, none of these rules are entirely different from things we encounter in designing for screens or services. When we're writing for any product, we should maintain a constant tone and keep it short and usability test everything, too. These are all skills we need as UX designers in any context. However, it does take a few adjustments to apply these patterns when speech is the primary method of input and output.

Discoverability and Predictability

Discoverability and predictability are definitely concerns when you're designing for interfaces for which the primary input method is voice, especially if you're taking advantage of NLP. This makes a lot of sense when you consider the difference between a visual interface and a voice interface.

Natural-language interfaces put the entire burden of deciding what to ask for on the user, while visual interfaces can give the user context clues such as interrogatory prompts or even explicit selection choices. When you go to your bank's website, you're often presented with several options; for example, whether you want to log in or learn more about opening an account or find a branch.

Imagine if your bank was more like Google (Figure 1-3). You just went to the site and were given a prompt to ask a question. Sometimes that would work fine. If you wanted to check your balance or order checks, it might be much easier to do as a conversation. "I need new checks." "Great, what's your account number?" and so on.

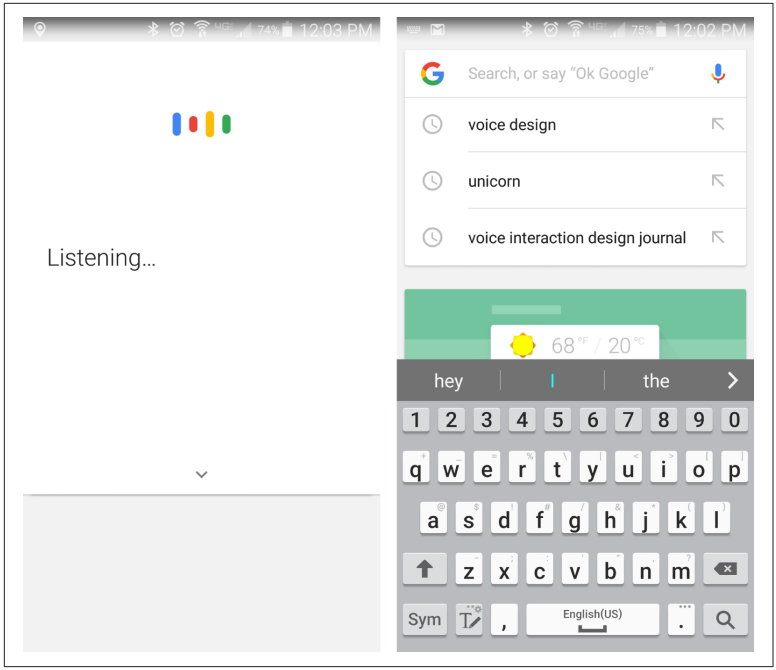

Figure 1-3. Ok Google, tell me about unicorns.

But, what if you thought you wanted to open a new business account that was tied to your old savings account, and there were several options to choose from, each with different fee structures and options? That's a much harder conversation to start, because you might not even know exactly what to ask for. You might never even realize that the business plans existed if you didn't know to ask for it.

This sort of discoverability is a serious problem when designing for open prompt voice interfaces. When Abi Jones first began designing for voice, she carried around a phony voice recorder and treated it like a magic device that could do whatever she wanted it to do. "It made me realize how hard it was to say what I wanted in the world," she says.

Even in voice interfaces that limit inputs and make functionality extremely discoverable—like IVRs that prompt the user to say specific words—designers still must deal with a level of unpredictability in response that is somewhat unusual when designing for screens. Most of our selections within a visual product are constrained by the UI. There are buttons or links to click, options to select, sliders to

slide. Of course, there is occasional open-text input, but that's almost always in a context for which it makes sense. When you type anything into the search box on Google, you're doing something predictable with that information, even if the input itself is unpredictable.

Siri, on the other hand, must decide what to do with your input based on the type of input. Does she open an app? Search the web? Text someone in your contacts list? The unpredictability of the input can be a tricky thing for designers to deal with, because we need to anticipate far more scenarios than we would if we constrained the user's input or even let the user know what he could do.

Designing for better discoverability and predictability

If you want to make features within your voice interface more discoverable, one option is to make your interface more proactive. Instead of forcing users to come up with what they want all on their own, start the conversation.

Karen Kaushansky thinks that Cortana does this especially well. "If you're in the car with headphones on and you get a text message, Cortana knows you're driving and announces the text message and asks if you want it read. It won't do that if your headphones aren't in, because it might not be private. It knows the context, and it starts the dialog with you rather than making you request the conversation be started."

By triggering user prompts based on context, like Cortana does, you can help users discover features of your interface that they might not otherwise know existed. In this case, the user learns that text messages can be read aloud.

The other option is simply to explain to users what they should say. Many IVRs that tried NLP have now gone back to giving users prompts. For example, instead of asking, "What do you need help with today?" your bank's telephone system might say something like, "What do you need help with? You can say Bank Balance, Order New Checks, Transfer Money, etc." Kaushansky points out that in some cases, even though the technology is more primitive, it's easier for users. "Using 'You can say..." can be better. Otherwise people don't know what to say."

Privacy and Accessibility

One of the most troubling aspects of voice interfaces, especially voice-only, is the obvious fact that everything might be audible. Now, that's probably fine when asking Alexa to play you some show tunes (Figure 1-4), but it's less fine when you're at work in an open plan office trying to access your health records. Again, context is everything.

Rebecca Nowlin Green, principal business consultant at Nuance Communications, helps Nuance's clients define their customer services experiences by incorporating speech recognition and other self-service technologies. She explains that well-designed voice interfaces should always have a fall back input method for any sensitive information.

Accessibility can also be an issue. Although voice recognition is quite good, it can be significantly reduced when dealing with non-native speakers, background noise, or even a bad phone connection in the case of IVRs. Abi Jones pointed out that you need to shout louder than the music playing on the Amazon Alexa to turn the volume down. The environment in which you're interacting with a product can have a huge impact on accessibility and ease of use.

Conversely, better voice UIs and audio output can increase the accessibility of products for people with poor vision or who have trouble typing or tapping on mobile screens. Smart homes can make everyday tasks easier for people with limited mobility by allowing access to devices without having to physically access them.

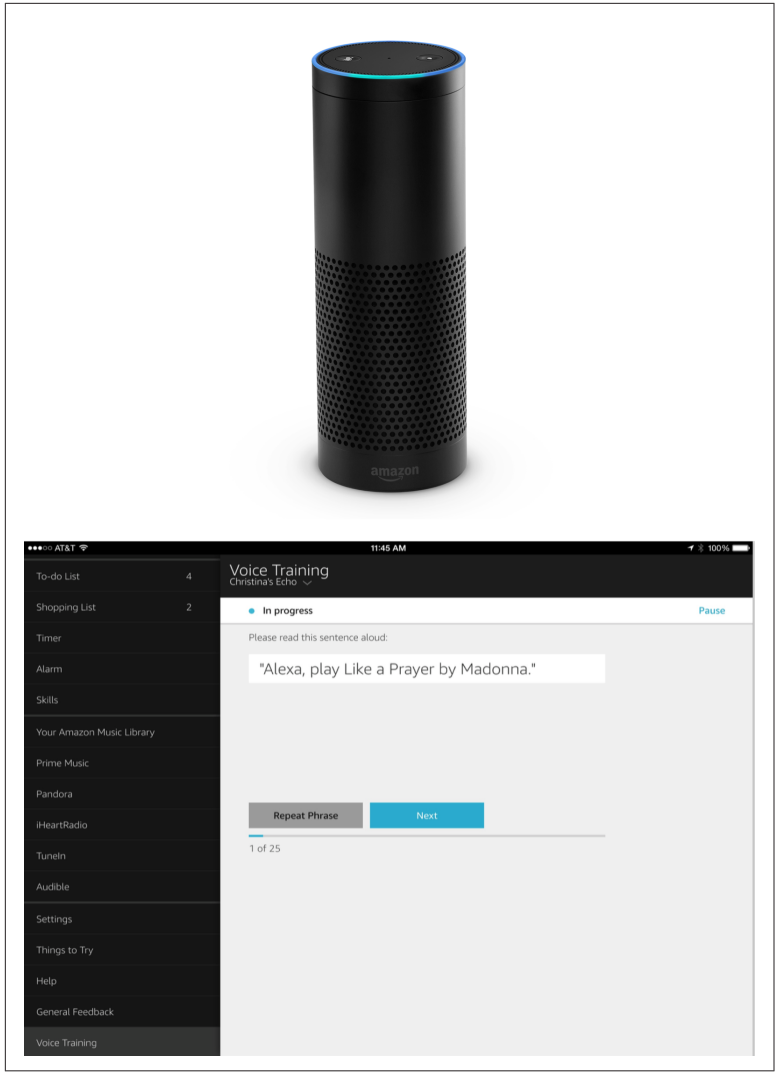

Figure 1-4. Amazon Alexa and companion iPad app.

Designing for better privacy and accessibility

The critical pieces of designing for better privacy are understanding what your users consider private information and predicting the context in which that information will be used. Sure, nobody wants to yell a social security number out loud on the train. But even email subject lines or the names of people who are sending text messages can be embarrassing in some situations.

Giving the user the ability to easily take a device out of voice mode and accommodating alternate methods of input and output are crucial to improving privacy. A more responsive audio output system would also help. Abi Jones points out that, "When you whisper to a human, they'll often whisper back. We adjust instantly to cues from the people around us. Our devices don't get this, so a mobile phone always talks back at the same volume as the ringtone. That's not necessarily what you want." Creating voice interfaces that take into account things such as ambient noise, location, and the volume of the input could dramatically improve privacy outcomes.

Better understanding of context can also help with accessibility issues. Nazmul Idris, an ex-Googler and founder of the startup TRNQL, is working on an SDK to make it easier for designers to take context into account when creating input and output systems. In one of the early sample apps being developed with TRNQL's technology, he showed how when the user is sitting still, the default input method is keyboard, but when the device senses that the user is walking, the default input method switches automatically to voice. In other words, if your phone knows you're seated at your desk, it knows you'd probably rather type. But if you're on the move, voice is safer because you won't accidentally walk into traffic while looking at your phone.

As our personal devices know more and more about us—where we live, where we work, when we're at the movies, when we're listening to music—they can make better decisions about how we might like to interact with them. Of course, this can mean an uncomfortable tradeoff for some users between privacy and accessibility. We might appreciate the phone knowing when we're walking, but we might not want it storing information about where we're walking. Making good choices about these trade-offs is part of designing for any mobile or personal device, and the additional privacy concerns associated with voice and audio technology can complicate the decision-making process.

Just remember, privacy and accessibility concerns are serious and never easy to handle, and you could potentially cause a lot of harm by having a device announce something when it shouldn't. Understanding your users' concerns and contexts can help you to make better decisions that will protect your users' data while also making things easier for them.

Pure VUI versus Multimodal Interfaces

While early VUIs were created to automate phone calls people were making to companies, VUIs are now showing up in things like connected home devices, mobile phones, and wearables. The interesting thing about these products is that most of them have multiple forms of input and output. Instead of relying entirely on voice in and audio out, devices might have small screens, flashing lights, companion apps, or small keyboards. These multimodal interfaces can create some fascinating design challenges.

Unfortunately, as is the case with all exciting new technologies, companies often decide to use voice input for the wrong reasons. People add voice because it's cool. They switch to NLP because they think people would prefer to just ask questions rather than be given instructions. They remove screens or buttons because they want to reduce clutter in the interface.

The only good reason to add voice as an input method or audio as an output method is because it makes the product better for the user. As with any other part of design, voice design should serve the needs of the user and solve a specific problem. Using voice and audio can be a powerful design tool, but used badly, it will only make your product worse.

"You need to determine how a conversational or voice interaction will improve the base experience for your use cases," Marco Iacono explains. "Are you simplifying something that would otherwise take six or eight taps? Are you accelerating the user deep into a task?"

When you think about it, some pieces of information are really easy to say, but they're hard to type, and vice versa. The same goes for output.

It might be convenient to say, "Give me all of the restaurants South of Market in San Francisco that serve brunch and are open now," but you probably wouldn't want Siri to read the names of all 760 out loud to you. In that case, voice input and screen output makes the most sense as an interaction.

On the other hand, finding a parking structure near where you're driving might be easier by looking at Google Maps, but you'll probably want to be read the directions as you go, so you don't miss a turn. In that case, screen input coupled with voice output is very useful. If you also wear a smart watch, it may buzz on your wrist when a turn is coming up, adding haptic feedback to the mix.

Abi Jones had a very specific moment in time when she realized that she wanted alternate forms of input. "I was changing a diaper and really needed to know if what I was looking at was a diaper rash, and I didn't want to get my phone out. I realized that having an interface I didn't have to touch right then would be really helpful."

So, as a designer, how do you tackle this? How do you decide on when to use different types of input or output. As Tony Sheeder explains it:

> Making decisions about using voice input and audio output is very device, task, and context specific. For example, you shouldn't look at a screen in a car, so you might use a voice interface to control navigation. But you don't want to change the steering wheel to voice control, either. You want to take advantage of standards that already exist.

Consider Using Voice and Audio for...

Shared interfaces in smart homes
: The most obvious use for voice input is for products that don't have screens or at least don't have screens nearby. Home automation devices such as the Nest thermostat (Figure 1-5), which has a small screen, and Philips Hue lightbulbs, which don't have any screens, have companion apps that users can install on phones or tablets. However, voice interfaces to products like these are significantly more useful than companion apps, in many ways. That's one of the reasons why the Philips Hue lightbulb already has an integration with the Amazon Alexa.

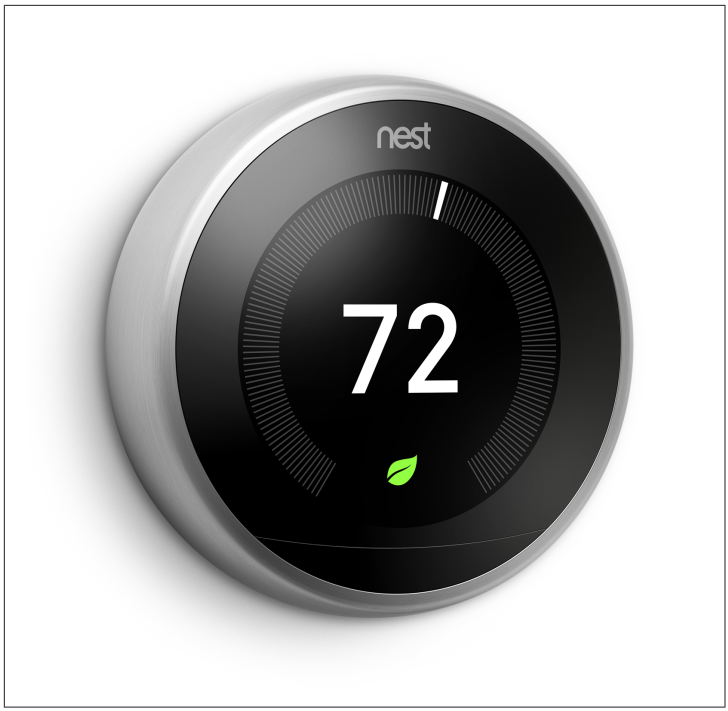

Figure 1-5. Nest, I'm cold!

Abi Jones explains, "Voice interfaces aren't good at distinguishing speakers, which means that, by default, voice UIs become usable for everyone in a space. Anyone can control the Amazon Alexa in the room." Having voice interfaces for smart home devices makes them more accessible to everybody, including guests, children, or people who just don't want to have to pull out their phone every time they want to turn on the lights.

Voice interfaces could also give people a way to control devices without being tethered to them. You can change or monitor the oven temperature without leaving the living room or make sure that all the lights are off in the house from bed.

Languages that are hard to type

This one might seem perfectly obvious, but some languages are easier to type than others. Providing people with the means to speak those words rather than type them can save a tremendous amount of time.

But, even if you're designing in English or any other fairly standard language, don't forget that there are certain types of words or concepts that people need to input that aren't necessarily easy to represent. Mathematical formulas, musical notes, or chemical bonds are very easy to say, but require very specific notation that can be difficult to input.

Complicated things that people can articulate

In fact, voice interfaces can be used well whenever we require complicated input from users that is easier to speak than to type. Suppose that you want to watch a movie tonight. You could go to your television listings or Netflix or Amazon Video and start to flip through the seven thousand choices in your queue. Or you could try to filter it based on their preset categories.

Or you could say, "Give me comedy movies without Adam Sandler, available tonight after 8 pm, for free. Only show the ones with four or more stars." Wouldn't that be better?

Complex queries with multiple inputs and filters are often easier for users to describe in words, which is what makes Google Search so powerful, whether the user is speaking or typing. Combining that power with voice input really shines, though, when you're searching on a device like a television that doesn't have a handy keyboard.

Safety in hands- or eyes-free environments

One place that voice interfaces have already begun to be used is in cars. Whether we like it or not, people will interact with devices in cars—whether it's navigation or making a phone call or texting—and most of us would really prefer that these drivers kept their hands on the wheel and eyes on the road.

Voice input and audio output make it possible for people to interact more safely with devices. It won't prevent distracted driving, but at least it keeps people from staring at a screen, and that might be the best we'll get until the cars can drive themselves.

It's not just in cars, though. Consider an operating room. Obviously we don't want our doctor texting a friend while cutting into us, but more and more devices are making their way into hospitals, and a voice interface with which a surgeon can

quickly access information from a medical record while operating could save lives.

Whenever we're in a situation in which we might need to call up information or respond quickly to something without using our hands or eyes, voice interfaces and audio outputs can increase safety and efficiency.

Don't Use Voice and Audio for...

Anything requiring negotiation or a lot of variables

Although speech recognition and interpretation technology is rapidly improving, we're still a long way from being able to have a real conversation with a device. Most successful interactions are one, or at most two, sentences long. Abi Jones points out that there is a slight conversational nature to the interactions with the Amazon Alexa, but even that is limited. She says:

> If you ask for a radio station that doesn't exist, it will ask you if you want to create it. But if you want to skip a song and turn up the volume, those are two completely separate actions, and they have to be performed separately.

This means that longer conversations that might involve multiple questions or negotiations are not great candidates for voice input. Scheduling a single doctor's appointment might be possible, if annoying, with a voice interface. However, scheduling a series of meetings where later ones relied on previous decisions would likely be a disaster. Yet, doing that visually, by selecting dates on a calendar, could be done quite easily.

Huge amounts of input or output

Voice input and audio output can be significantly slower than text for large amounts of data, especially when most of that data will be ignored. Remember the brunch restaurant in SOMA example? Although it might be easier to verbally describe the kind of place you're looking for, having all of those results read back to you would be awful and slow. Similarly, quickly selecting several items from a list is a much faster input process when you can scan the list visually and select by tapping rather than explaining to the device which items you'd like to select.

When you're dealing with large amounts of input or output, visual interfaces are almost always superior to audio or voice. This is true for most people when dealing with text input, as well.

Even though some people are outstanding at dictating their thoughts, it does take quite a bit of practice to be able to accurately speak an entire email without going back and editing it later. Short texts and email responses are great for voice input, but there's a reason this report was written with a keyboard. It's just easier for most people.

Hard-to-describe input
: In the previous section, we looked at languages or jargon that's easier to describe than to type. There are also concepts that can be difficult to describe out loud even though they're simple to represent visually.

  Although there are many ways in which a voice interface might be easier for interacting with a smart television, you almost certainly wouldn't want to change the color balance by describing it to the TV. You probably wouldn't want to change the side mirror settings on your car by talking to it either. Just because voice input works for some interactions in a context, it doesn't mean that it's right for all of the interactions.

Comparing lists of complicated things
: Another area for which voice interfaces fail miserably is in comparing complicated lists of items. Picture four different computers or smartphones arranged in a well-designed grid of features and prices. That's a very simple, understandable interface. Now, instead, imagine all of those items being read to you by a computer.

  Even though there isn't an enormous amount of data, it still requires the user to hear and remember several different options at once, which becomes difficult very quickly, especially with no visual cues. There's a reason that stores have floor models and labels and product sheets as well as sales associates. Some information is simply easier to take in visually and shouldn't be forced into a conversation, especially one with a computer.

Successfully Combining Modes

Tony Sheeder was around when the first Dragon Mobile Assistant was being developed. The Dragon Mobile Assistant is an app designed by Nuance Communications that makes it possible for users to be more hands-free with their phones. It uses outstanding

voice recognition technology to do things like set appointments, send texts, post to social media, or check the weather. In other words, it's exactly what the name implies. But despite having an eerily good understanding of what the user is asking for, the very first version had an interesting design issue.

Sheeder explains that the voice interaction design and the visual design of the phone were, by necessity, initially done separately by different groups of designers. Both groups felt that they had to handle all the input and output by themselves. "It ended up having a lot of redundant information. It would show you something on the screen and read it to you," Sheeder said. When the designers learned to work together and began relying on one another, they figured out which things made the most sense visually and which should be handled with audio. Consequently, the next versions dealt with input and output more naturally.

To successfully combine modes, it's important to understand not just when you should use voice and when you shouldn't, but how you can effectively combine voice interfaces with other methods of input and output.

There are several different types of multimodal experiences, including the most common models, which you can see in Table 1-1.

Table 1-1. Common multimodal interfaces

Input methods	Output methods	Examples
Voice, touchscreen	Small screen, haptic	Smart watches
Voice, keyboard, gestures, touchscreen	Screen, audio, haptic	Smartphones
Voice, companion app	Audio, companion app	Amazon Alexa
Voice, camera/gesture	Screen, audio	Xbox Kinect
Voice, scroll knob, hard buttons, touchscreen	Audio, screen	Some car interfaces

There will, undoubtedly, be more input and output combinations over the next few years, and that means designers are going to need to pay a lot of attention to understanding which input and output methods to employ for the best usability. Rebecca Nowlin Green, says that, when you add more than one input or output method "complexity just skyrockets."

There are some useful tips for figuring out which combination of input and output methods is right for your product.

Pure voice—finite state

Finite-state, pure voice interfaces are things like classic IVRs. These are the systems where you call in and hear a voice say, "What can we help you with? You can say, 'Check my balance, Open an account, Order new checks, or Representative.'" At each point in the flow, the system only understands those specific commands.

Although they're not sexy, they're still very commonly used by companies who want to reduce call-center costs by handling common tasks and routing callers correctly for more complicated tasks.

When to use it? Finite-state, pure voice systems are still useful for certain systems. Because the only input and output methods are voice and audio, they're going to be handy for products that don't have a screen. This obviously includes IVR phone systems, but it could also be a physical device like a screenless wearable device (Figure 1-6).

In general, you'll use a finite-state system when your product is simple enough that it's not worth going for NLP. They're useful for products for which users can be trained to do a very small number of tasks. For example, a bedside clock that lets you set alarms doesn't necessarily need a full NLP system. It just needs to understand preset commands such as, "Set alarm," that users could memorize. The same is true for the autodialer on a corporate phone system. It's not handling open-ended queries. It's just recognizing a specific list of names and directing calls.

One of the main problems with finite-state systems—and the reason so many people hate most IVRs—is that they often require users to go through a labyrinth of prompts to get to the one thing they want. If the system tries to handle too much, it can require a huge amount of investment on the user's part, only to end with having to talk to a representative or being disconnected.

Figure 1-6. When can I start talking to my Fitbit?

Simple systems that handle just a few predictable tasks that users might not know how to ask for naturally are good candidates for a pure voice, finite-state interface. For example, car's audio system might be fine for one. There are a limited number of things you might want from it: play a song, turn up the volume, and so on. The user interacts with it daily, so they're more likely to use the same vocabulary for the commands every time. Each command is simple and discrete, so users won't get trapped. And finally, it's very easy to recognize and recover from a mistake.

Pure voice—NLP

As soon as the technology improved, many IVR systems moved to NLP. This means that, when you call a company for help, you might get a computer asking, "What can I help you with today?" after which there is a very good possibility that it will recognize what

you've asked for and give it to you as long as you use exactly the words it recognizes.

Whether or not you choose a pure voice system is, just like in the previous section, probably determined by whether or not your product has a screen. Whether you want NLP is a different question.

Rebecca Nowlin Green helps companies decide what sort of voice interface is right for them. She usually recommends natural-language understanding IVRs for high call volumes with a lot of routing complexity. In other words, if there are a huge number of different things that a user might be calling about, designing a flow to quickly get users to the right destination can help skip a few layers of questions and avoid errors.

Of course, when customers call about things such as health insurance or banking, they often don't know exactly the language they need to use to get what they want. They might have difficulty explaining to a computer, "I want to know if you will pay me back for this thing that my doctor says I need to have done." But, if you give them the option to make a selection like, "Get preapproved for a medical procedure," they can recognize that as something that sounds right. In this case, you might want to offer some directed dialog to give the user clues as to what they can ask for. Just because you opt to let the user to say anything doesn't mean that you have to leave them with no suggestions about what they might want to say.

Voice input/visual output

These days, of course, we're integrating voice input into more and more products that also have accompanying screens. In many cases, we're allowing the user to give voice input, but providing visual output rather than audio.

Texting on smart watches is an excellent example of when to use this combination. The watch faces are small enough to make any other form of inputting text next to impossible, but the screens can easily show the user the result of voice input, giving them a way to easily check the recognition and recover from any errors.

Smart televisions would also be good candidates for this sort of interface. They don't currently have an input method that lends itself to complicated input, but they certainly have enough functionality to make natural-language queries useful. Being able to say, "Show me all the times that this week's *So You Think You Can Dance* is

being shown," would be significantly easier than searching for it by using the arrow keys on the remote control.

Thomas Hebner says one of his favorite uses of a voice interface is a popular pizza ordering app, with which the user can simply say an order out loud. It's easier to say, "Two small pizzas, one just cheese, and one with pepperoni and mushrooms," than it is to make all of those selections by tapping. The app confirms the order on the screen so that the user can verify that everything is correct, which is obviously faster and more pleasant than having the entire order read aloud.

In fact, any interface that includes open-ended, complex input that is easy for the user to speak out loud but that produces results that would be unpleasant to have read back is the perfect candidate for an interface that accommodates voice input and provides visual output.

Physical input/audio output

This particular interface combination seems unusual, until you realize that it's actually the traditional input for every single stereo you've ever used. Physical input could be tapping on a mobile phone screen, but it could also be physical buttons or even using gestures in front of a camera.

The other versions of this are a little bit less common, although things like self-checkout systems in grocery stores often have spoken instructions that are separate from anything shown on the screen. Audio output can be useful for any product that will be used infrequently by large numbers of untrained people. A reassuring human voice giving instructions can help someone who otherwise might struggle with a process such as purchasing groceries, for which several tasks must be performed in a specific order.

A little of everything

Many products are moving to multimodal interfaces that combine voice and physical inputs with screen and audio output. Navigation apps might be the perfect example of a category of product that combines all of these elements well.

Users can touch places on the map, scroll around to see what's nearby, or type in an address using physical input. When driving, they can simply say the name of a destination; this way, they don't

have to take their eyes off the road or hands off the wheel to change the destination. Audio output makes it possible for the mobile device to give clear navigating instructions, while the map shows turns and other information, like traffic, that would be hard to express verbally.

It's an outstanding combination of input and output methods, each one contributing to the user experience in the way best suited to it. Each input and output affordance in a well-designed navigation system takes into account the context and needs of the user. You can be hands- and eyes-free when you have to be, but still have access to the screen when it's useful.

Designs like this don't just happen. They're carefully crafted based on a deep understanding of how and when users interact with products. Navigation systems are used in cars, so voice and audio become obvious choices. Of course, not all products have such clear-cut uses, so it can be challenging to decide when voice and audio interfaces will improve the experience. Having a design framework for making those decisions can be helpful.

How Long Until Star Trek???

As Abi Jones says, "The more you do research on voice UIs, the more it makes it utterly extraordinary how easy and fluid it is to communicate with humans."

Unfortunately, this is true, and it means that it doesn't look good for the Star Trek computer interface in the very near future. Despite Watson's win on Jeopardy, we're still a long way from having real conversations with our smart toasters, and honestly, that's probably for the best.

There are a few things holding us back from this future. Some of them can be solved by technology, but others might only be solved by humans growing more comfortable with the technology over time.

The Problems We Still Face

"None of this is easy. There are still some fundamental challenges with even just the basics, such as getting yes/no recognition performance," Rebecca Nowlin Green says. But it's getting better.

Some products in 1999 had around a 65 percent recognition rate, whereas today's rates are closer to 92 percent. Nonetheless, that still means that we're talking to systems that don't understand us eight percent of the time, which can be frustrating when we're trying to accomplish a task. After all, if you have not made their problem go away, people won't use your technology.

Both systems and humans can have trouble with things like "barge-ins," which are the times when the computer is saying something and the user talks over it in order to skip the directions. People have trouble with interrupting because it can feel unnatural to cut off somebody who is speaking. Devices have trouble with it because they don't always catch the first part of the command, which leaves shouting the same command over and over at an inanimate object that keeps saying, "Hmmm...I didn't get that."

Of course, as voice interfaces become more common, we'll see more conflicts and confusion. It's generally pretty clear when you're typing on a phone or pushing a button on your oven which device you're interacting with. That's less true of voice activated devices. Tony Sheeder explains, "If I say 'raise the temperature to 350 degrees,' the system should know if you're talking about the oven or the iGrill and not the home thermostat. We shouldn't have a dozen different ways of interacting with all the different systems, like we do now."

The same is true of smart watches and wearables. When I say, "OK Google" I want my watch to respond, but not the four other Google watches in the room. These aren't problems that can be solved entirely by better technology. At some point, we might all have some sort of smart hub in our homes that route our instructions to the various voice-controlled systems. We might have voice recognition for our phones, wearables, and cars. But for the immediate future, these problems will have to be solved by good, thoughtful, context-aware design and by designers who realize the potential for conflicts.

The biggest problem we still face may be the human one. All the experts agree that people still kind of hate voice interfaces. Abi Jones says:

> When humans talk to each other, we create a shared understanding of the world through dialog. We're willing to forgive a lot. When watching user studies and interactions with computers, someone interacting with a voice interface might come into it being forgiving, but when it reveals a lack of humanity, people start treating it like a computer again.

The Future of Voice

So, what should the future look like? There are a huge number of products that could be improved in the near term. Tony Sheeder says:

> TV interfaces are all awful. Making selections with a grid of letters is awful. Speech just cuts through that whole thing. Also, cars can be made a lot safer. Auto interfaces just aren't doing what they need to do.

Also, voice recognition is no longer only available to a few companies that specialize in it. Now that it's on phones and wearables, we're seeing more and more companies incorporate it into apps as a feature. The other day I ordered something from Amazon by shouting at my watch. Was it necessary? No. Was it kind of fun? Yes. Will it lead to my bankruptcy? Probably.

The distant future is harder to predict, because that's how time works. The combination of big data and improved voice recognition might get us closer to real conversations with computers sooner than we imagine. But, the real opportunities lie in allowing people to interact more naturally with devices where screens and keyboards don't make sense.

Tony Sheeder thinks that there's a lot to be done with Virtual Reality and Augmented Reality:

> People in gaming environments shouldn't be tied to clicking on buttons to interact with elements. Speech offers nuance and fine-grained possibilities you don't get from other interfaces.

Rebecca Nowlin Green predicts more virtual assistants and avatars and more one-on-one interactions with entities that represent companies:

> Big data will get more and more relevant by tracking user behaviors over time and using that information to influence future experiences. Your favorite coffee shop will wake you up with a coffee assistant.

Abi Jones thinks that we'll get more ubiquitous interfaces that are only there when you need them. She also thinks that using voice for accessibility purposes will improve things overall:

> A lot of things we had originally for accessibility, like curb cuts or elevators, are good for all of us.

But, as with any new technology, we're going to see a lot of bad interfaces in the beginning as companies learn when to use voice and designers learn how to use it to make things better for customers. As Marco Iacono points out:

> When the Apple Watch launched, it took awhile for developers to translate their service into this new product. Initially, some approached it as a shrinking down of their iPhone app functionality. But it didn't take long to identify the key features that are relevant for the short, snackable interactions on your wrist.

Most of the VUI designers I talked to for this report had 15 or 20 years' of experience in this sort of design, so saying we're at the beginning might seem laughable to them, but it's clear that we're still very early in the process of this technology becoming mainstream.

As Thomas Hebner says, "This is an incredibly exciting time in voice design. Voice designers are mostly in IVR, but with the APIs opening up and with more consumer electronics, we're on the edge of a boom. The world hasn't woken up to the idea that it needs voice design. Early on, there were some really bad IVR apps, and then it got better. We're at the beginning of that pain now. In a couple of years, people will be clamoring for voice design."

Becoming a VUI Designer

If Hebner is correct, now is an outstanding time to become a VUI designer. If you're transitioning from more traditional UX or product design, you should find it a fairly natural transition. In fact, if you're currently designing apps or working with wearables or smart home devices, you might not have much of a choice about learning the fundamentals of VUI design.

Some companies, such as Nuance Communications, have been building voice technologies for more than 25 years. Others, like

Google and Amazon, are just at the beginning of incorporating voice into their products. Voice input is completely integrated into all of the smartphones and watches currently on the market, and there is still enormous room for growth.

Resources

The hardest part about becoming a VUI designer right now might be the lack of classes and resources available to new designers. If you're serious about voice design, your best bet is to get a solid grounding in User-Centered Design, good user research techniques, and information architecture, and then to begin working with a team with some voice design experience.

Books

Nass, Clifford. *Wired for Speech: How Voice Activates and Advances the Human-Computer Relationship*.

Cohen, Michael H., James P. Giangola, and Jennifer Balogh. *Voice User Interface Design*.

Conferences and Talks

"Fundamentals of voice interface design," Tanya Kraljic (Nuance Communications) at O'Reilly Design Conference, January 19–20, 2016.

"Evangelizing and Designing Voice User Interface: Adopting VUI in a GUI world", Stephen Gay and Susan Hura.

Organization

The Association for Voice Interaction Design

About the Author

Laura Klein fell in love with technology when she saw her first user research session in 1995. Since then, she's worked as an engineer, UX designer, and product manager at both startups and large companies in Silicon Valley. Her book, *UX for Lean Startups* (O'Reilly), and her popular design blog, Users Know, both help teams learn more about their users and apply that knowledge in order to build better products. She currently consults with companies that want to improve their research, user experience, and product development processes.