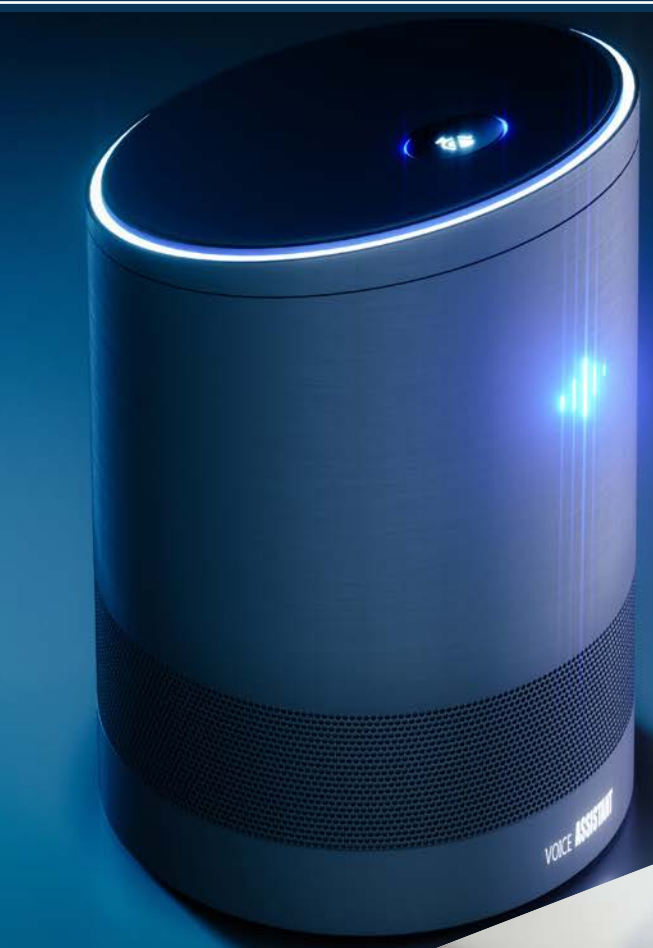
Embedded COMPUTING DESIGN

WINTER 2018
VOLUME 16 | 4

EMBEDDED-COMPUTING.COM



VOICE ASSISTANT BATTLES

BEGINS ON PAGE 22

DEVELOPMENT KIT SELECTOR

Digi-Key ELECTRONICS

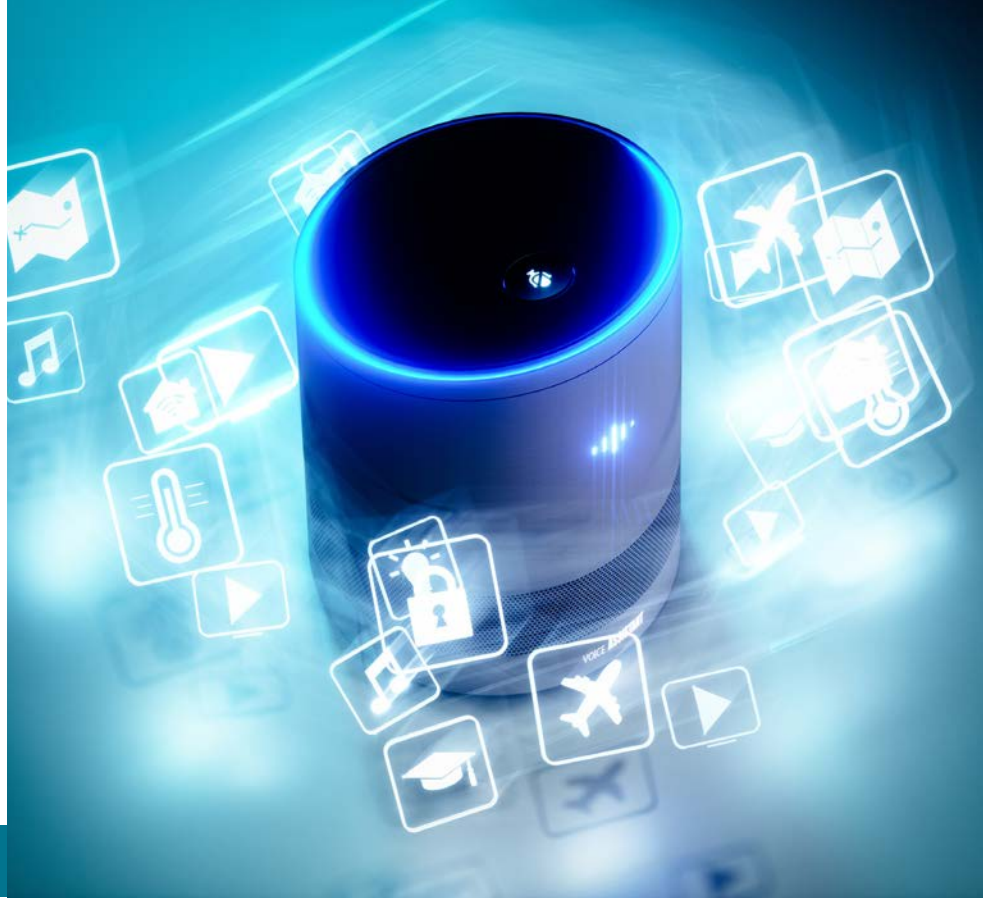
www.embedded-computing.com/designs/iot_dev_kits

IS EMBEDDED TECH UP TO THE AUTONOMOUS DRIVING CHALLENGE?

BEGINS ON PAGE 10

GETTING A VOICE USER EXPERIENCE RIGHT IS HARDER THAN YOU THINK

By Jeff LeBlanc, ICS and Boston UX



Voice interaction is one of the most disruptive technologies of the 21st century. Every day, more devices are hitting the market with a voice user interface (VUI) component. While many of the technical challenges to voice-enabling a device have already been addressed, for designers, making the experience of using the device a pleasing one for the end user is still an open question. This article addresses some of the challenges and best practices around designing a VUI that is effective, natural, and engaging for the user, including designing for confidence thresholds, accommodating barge in, using N-best lists, and how to talk with (instead of at) the user in a real conversation.

"J.A.R.V.I.S., are you up?"
"For you sir, always."
– "Iron Man" (2008)

While voice user interfaces (VUI) have been on the periphery of the public mindset since 1968, when HAL and Dave Bowman had their disagreements, it wasn't until Tony Stark started bantering with J.A.R.V.I.S. in 2008 that the notion of a helpful voice-controlled "smart home" started to come into focus.

The hugely successful Amazon Echo device, released in 2014, combined the latest in voice-recognition technology with powerful cloud-based computing to provide an in-home experience that nearly rivaled that depicted in the movies. Turning on the lights or your sound system was never so easy. Since then, Google, Apple, and other technology companies have jumped into the fray and are tripping over each other to provide the finest interactive voice experiences for your home, workplace, and car.

Voice-rec background

This technology has been a long time coming: Bell Labs and IBM worked on speech systems as far back as the 1950s. But, it wasn't until the late 1990s that Dragon's NaturallySpeaking software gained enough traction to bring speech recognition to consumers' collective consciousness. While it was revolutionary at the time, NaturallySpeaking required a fair amount of "training" by the end user to achieve the 90 percent accuracy level that makes speech recognition viable as a form of human-computer interaction. So the technology was not nearly as natural as it could be.

In the years since then, developers, designers, and technologists have toiled away, trying to "solve voice." Yet, we've only gained an additional 5 percent in recognition accuracy.

Why is designing more human-like voice interfaces so difficult?

When designing VUIs, there are two key aspects that must be addressed. The first is ensuring the interface can capably recognize sound as human speech. Known as Automated Speech Recognition (ASR), this is the core of speech-to-text software engines. ASR can be performed on modern consumer hardware with reasonable processing speed. But ASR is more typically done in the cloud. Devices like the Amazon Echo only do enough local processing to find their "wake word," while the rest of the

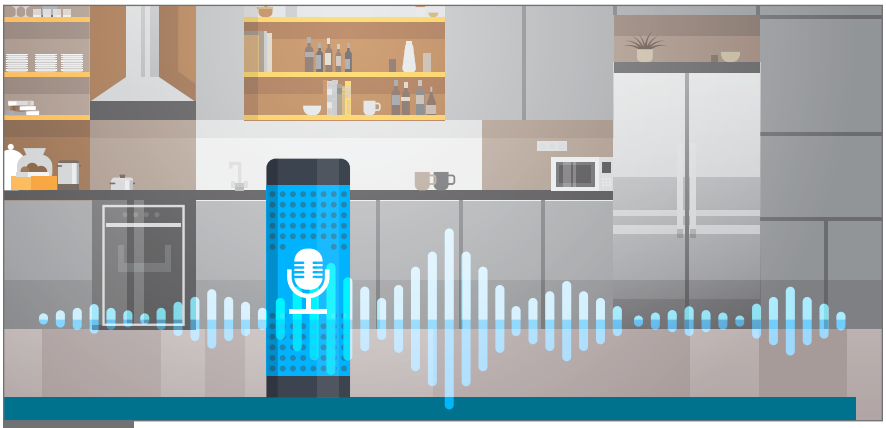


FIGURE 1 Speech-recognition systems must move beyond following commands to actually engaging the user in a dialog.

work is done by remote computing resources. So yes, Alexa is listening to everything you say. But she only cares when you say her name.

The second, and more difficult, aspect of the voice experience is ensuring that the device knows what to do with the speech once it's recognized. Natural Language Understanding (NLU) – which combines a variety of disciplines including linguistics, cognitive science, and artificial intelligence – has challenged computer scientists for years. Although some experts view ASR as the "hard part" of developing VUIs, I disagree. We've been holding steady at about 95 percent accuracy for many years, comparable to human-to-human communication. Yup, even human-to-human communication isn't 100 percent accurate. Think about how many times you say "Huh?" or "What?" when speaking with another person. Yet those conversations are easily understood.

Our challenge as UX designers is figuring out how to create an exceptional interactive voice experience, getting as close as possible to mimicking the person-to-person interface experience.

Say what?

This phenomenon is known as a Natural User Interface or NUI. Getting simple commands to work correctly is straightforward; it's mostly a matter of pulling the correct keywords out of the utterance. For instance, getting your smart home to respond correctly to "turn on the dining room lights" isn't too complex. It just involves creating an interface that can recognize the desired action ("turn on") and what to perform that action on ("dining room lights").

But there are still challenges: Since we have slightly less than 100 percent speech-recognition accuracy, the device might not understand your exact utterance. Perhaps the voice assistant heard you say "turn on the dine room lights." While a human can easily make the leap from dine room to dining room, that's not the case in the binary world of computers. "Dine" does not equal "dining," so your voice assistant doesn't understand what you're asking. You end up frustrated, eating in the dark. Fortunately, we can design around this. The solution lies in moving beyond simply taking utterances and commands to actually engaging our user in a dialog. (Figure 1.)

In our example, the smart home understands your intent – you want to turn on the dining room lights – but it didn't get quite enough information to carry out the task. So we program the VUI to do something typical in human-to-human interaction: ask for clarification. Our smart home could respond "Sorry, I didn't quite catch that. What do you want to turn on?"

This interaction is built on the concept of confidence level: How sure is your smart home that it really understood you? If the smart home is pretty sure it understands your request – say greater than 75 percent accuracy – it can just execute it. If it's only somewhat sure, the device can ask for clarification. By leveraging confidence level and engaging in dialogue, you can clarify your request without having to restart the whole command interaction from the wake word.

On the N-best list

This next design technique builds atop this conversational approach to try to predict what you might say based on expected responses from prior conversations. It's not unreasonable for your smart home to hear "dine" instead of "dining." Or even other similar-sounding words like "diving."

By collecting these near misses in something called an N-best list, your smart home can capture likely possibilities. Now your home's VUI can either ask you for confirmation of a word on the list or simply go ahead and execute that command. Having your home respond with "I think you asked me to turn on the dining room lights. Is that right?" shows that your home is smart enough to (most likely) figure out what you said but is courteous enough to double check just in case it didn't quite understand the request 100 percent.

Flowcharts and maps

Flowcharts allow VUI designers to map out the possible branches found in even simple interactions. Continuing the conversation about the dining room lights, to ensure a smooth, natural dialog the VUI designer has to think about what your likely response would be. You might answer the request for clarification about turning on the lights with a simple "yes." In that case, the smart home should turn on the lights.

But if you listen to recordings of human-to-human conversations, they're often not as precise. What if you responded with "yup" instead of "yes"? Or "that's

right" or "make it so" or any number of affirmations? What if you responded in the negative? "No. Nope. Uh-uh." Would your smart house know what to do?

This scenario is precisely why checking lists instead of simple keyword matching is critical. It's the best way to achieve the most natural interaction.

Barging in

Another aspect of human-to-human communication that bears mentioning is that of interruption. Sometimes we're impolite and don't wait for the other person in the conversation to finish speaking before we start talking. Other times, interrupting is the only way to move a conversation forward in a timely manner. In both cases, the ability to interrupt makes a conversation more natural.

Here's an example. You got into a fender bender and called your insurance company to file a claim. Listening to a long list of options on the company's automated phone system, you interrupt as soon as you hear "press 3 to reach the claims department." You eagerly tap the "3" key and don't bother to listen to the rest of the list.

This ability to barge in and interrupt the conversation is something VUI designers need to incorporate in order to create a human-like voice interaction. (If your waiter was reading off the list of salad dressings and you said "Stop, I want that one, the vinaigrette" and he kept on listing dressings, things would get a bit awkward.) The Amazon Echo does a great job of supporting barge-in, letting a user say "Alexa, cancel" at any time.

The takeaway

Designing a compelling, human-sounding voice assistant is certainly possible. Google's new Duplex phone bot, for instance, comes complete with conversational tics common to most humans, including "ahs" and "ums" peppered throughout the dialog. Some people have even expressed concerns about just how human it sounds as the line between AI and human speech becomes increasingly blurred.

Still, this is the future. So how do we deliver? By paying attention to basics like those I've outlined, designers can create the natural, effortless voice-powered interactions today's consumers expect. ECD

Jeff LeBlanc *is Director of User Experience for both Boston UX and ICS. He heads the creative team. With an engineering degree from Worcester Polytechnic Institute (where he's also an adjunct professor), he's an expert at bridging the gap between design and development. What makes his day? Applying human factors principles to UX design. And 3D-printing a wearable Iron Man suit.*

OpenSystems Media works with industry leaders to develop and publish content that educates our readers.

Qt or HTML5? A Million Dollar Question

By Burkhard Stubert, Chief Engineer, EmbeddedUse

With a five times smaller footprint, four to eight times lower RAM requirements, and a more efficient rendering flow than HTML, using Qt for user interfaces provides faster start-up times and maintains the cherished 60 fps and 100 ms response time, where HTML would struggle. Learn how the author says he could save one of the world's largest home appliance manufacturers millions of euros by choosing Qt over HTML. The secret? Qt scales down to lower-end hardware a lot better, without sacrificing user experience.

http://www.embedded-computing.com/white-paper-library/qt-or-html5

Embedded COMPUTING DESIGN

Check out our white papers at www.embedded-computing.com/white-paper-library