Introduction to the Voice Stack

Table of Contents

Who is this Whitepaper for?

This Whitepaper is intended for experienced CIOs, CTOs and CDOs who may not have exposure to voice services; senior technologists and technical leads who are now being asked to consider or integrate voice services into existing enterprise technologies; and university students of computer science, computational linguistics and information systems who need to have a foundational understanding of voice services. This Whitepaper does not go into significant technical detail and is easily consumable by non-technologists.

A Brief History of Human Computer Interfaces

Voice technology hasn't just eventuated overnight - it is the latest type of user interface in a long history of interaction evolution. When modern computers were first born in the 1950s, they were controlled - laboriously - through the use of manually-marked punch cards. Keyboards made controlling computers easier -although the instructions to control them were still typed using clunky, hard to remember commands. The advent of pointing devices such as the mouse and the graphics tablet, coupled with the graphical user interface, meant that computers became much more 'user-friendly'. Further advances with touchscreens saw the rise of us 'tapping' to control our computers - which were by now pocket-sized.

Voice user interaction is a natural continuation of this evolution - making it ever easier and more 'frictionless' for people to engage with computing devices.

THE TREND IS TOWARDS MORE NATURAL INTERACTION

What's Driving The Move to Voice?

There are several drivers behind the move to voice user interaction.

Mobility and ubiquitous computing

People are mobile. We have laptops, smartphones, wearables and even implantables with us. We don't want to carry a range of user input devices with us. We need the user interface to be ready to use wherever we're working—or playing from—right now. Voice assistants solve this by being 'always on' and 'always close'.

Always-on work habits and context-switching

We're leading busier lifestyles, with our time sliced between work, leisure, family and so on. Devices are 'always-on' and our work habits mean that the boundaries between work and personal time are blurring - and the Devices we use need to cater for both business and fun. This trend means that we're constantly 'context-switching' from one task to another; a practice which has been shown to impede concentration and productivity. Voice user interaction can assist here by reducing the friction involved with context-switching; it's more natural to ask for something than to type in, or tap, a request.

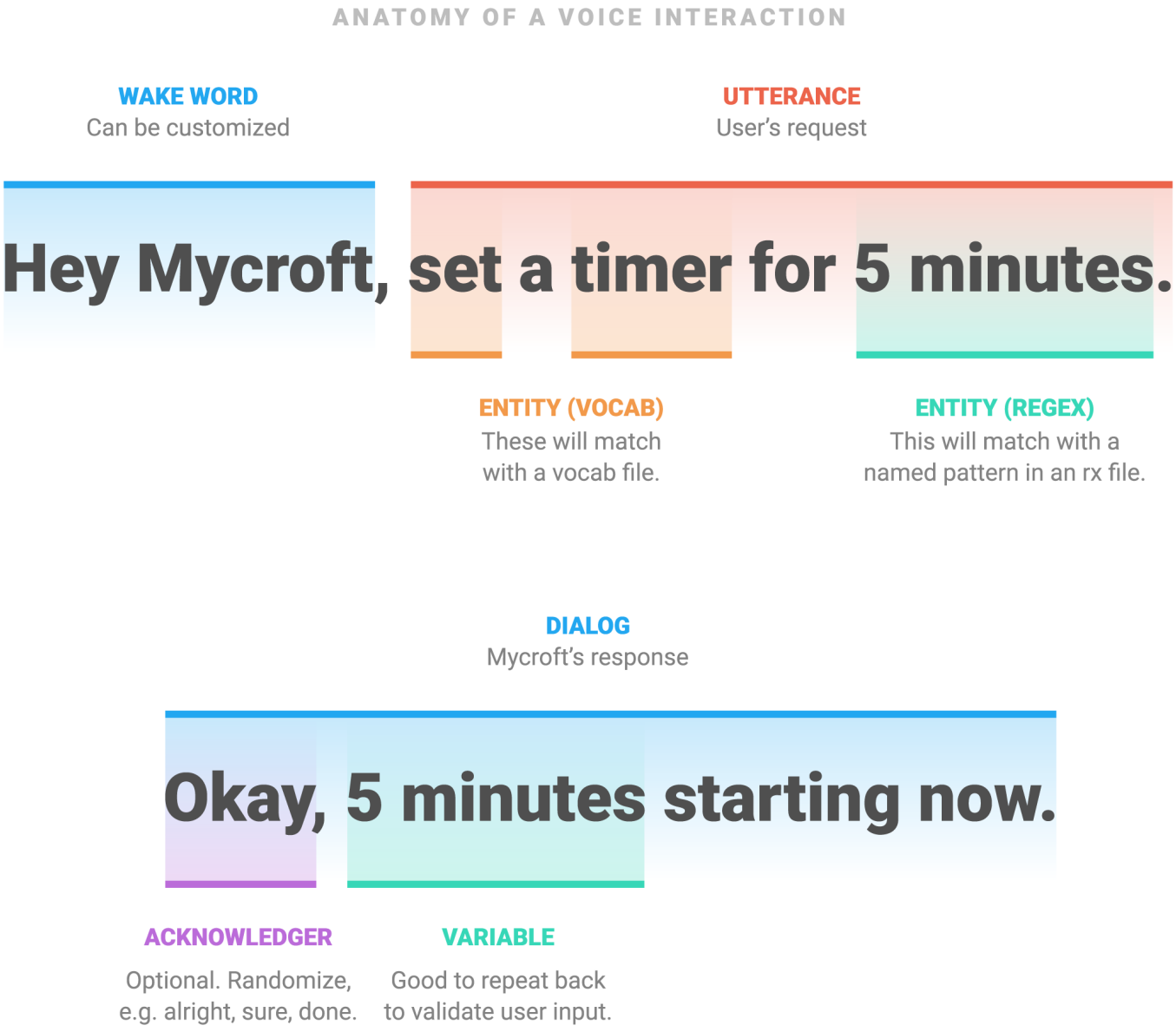
Technology improvements

The third driver spurring a move to voice interaction is the evolution of the technology itself. Several improvements to processor speed, machine learning algorithms to detect speech to text, and voice models for text to speech, have all combined to make implementing voice technology much more feasible. In the past, voice technology worked by recognizing a very small subset of keywords, and being able to use a very small set of commands in response - think of an interactive voice response (IVR) telephone system. Now that processing power and machine learning has advanced, we're better able approximate 'natural language understanding'.

My language, my way

There are over 7 billion people on the planet, and nearly 7000 active languages. While some technologies can cater for different languages, being able to interact naturally with a computer in your own native tongue is a strong attractor for billions of people who don't natively speak major languages.

What Does the Voice Stack Look Like?



The Voice Stack is like a layer cake - each section providing part of the overall experience.

Wake Word

Sometimes called a 'hot word', the ***Wake Word*** is what a voice assistant is trained to hear when you're about to issue a command. You've probably already heard Wake Words before - such as 'Hey Siri', 'OK Google', 'Alexa' and so on. A voice assistant is continually 'listening' for the Wake Word. This gives rise to one of the chief ethical concerns of voice technology - what can a voice assistant 'hear' when you're not issuing commands? For this reason privacy, and the safe storage and transmission of voice data are becoming more important.

Utterance

Once the Device is 'listening' for commands, the user will speak an ***Utterance***. This might be something like 'What time is it in London?' or 'What's the price of Bitcoin today?' or 'What's the weather like in Istanbul on Friday?'. The voice assistant will make an audio recording of the Utterance.

Speech to Text (STT)

Once the voice assistant has a recording of the Utterance, it then runs the Utterance through a Speech to Text processor, to turn the audio into words. This is one of the most technically challenging and labor-intensive parts of a voice stack. There are billions of people in the world, with different accents, intonation, prosody and so on. Different voice assistants use different STT engines; their accuracy varies.

Intent Parsing

Once the audio recording has been converted to text, an ***Intent Parser*** then combs through the text to identify the user's ***Intent*** - that is, what command they want the voice assistant to perform. Again, different voice assistants use different approaches to Intent Parsing. For example, with Mycroft, we have two different types of Intent Parsers - one which works on keywords identified in the Utterance, and another which is based on a neural network to take a 'best guess' at what the user intended.

Skill invocation

If the Intent Parser recognizes an Intent, the voice assistant then invokes a ***Skill*** to handle that Intent. For example, there might be a Time Skill to handle time questions, a Cryptocurrency Skill to handle questions about Bitcoin pricing and a Weather Skill to handle inquiries about weather. The Skill then uses parts of the Utterance to respond to the user - for instance in the Utterance

'What's the weather like in Istanbul on Friday', the Skill would take both the city ***keyword*** and the day ***keyword***, perform a weather query and then respond to the user.

Text to Speech (TTS)

Finally, the voice assistant responds to the user using a voice. This layer is called Text to Speech (TTS) and is essentially the reverse of the STT layer - turning text into speech. Again, different voice assistants use different TTS engines; their pitch, gender representation and how 'natural' they sound varies between providers.

Central platform

Most voice assistants will also have some sort of central platform with which they register. The central platform is responsible for registering the voice assistant, and will usually gather metrics from the voice assistant device itself.

Considerations for your Organization's Voice Stack

Privacy

Voice assistants are located in homes, in intimate spaces such as lounge rooms and bedrooms. As your organization considers your approach to adopting voice, consider how you will protect the privacy of your users and clients. Who can hear the voice interaction when it occurs? Who is able to see details of what was asked, and when? One of the strongest criticisms of the 'big' voice providers is that their technology is positioned to gather data in order to better target advertising to you. Is that something that resonates with your organization's values, or with your corporate objectives? If not, you might want to consider a privacy-first voice stack, like that provided by Mycroft AI. In many ways, data is the new oil - and 'big' voice companies are positioning to gather as much data as possible so they can then turn this data into new products and service offerings - often without the informed consent of end users.

Security

If the voice assistant is deployed on a network, how likely is it that someone unauthorized could 'listen in' on a voice interaction? This may not present a problem in some contexts, but in others such as medical or financial interactions, you will want to ensure the security of the device - and the data it's carrying - on the network. Once the data goes 'off-network' - say back to a cloud server - how can you ensure that the data will be safe? Most voice services operate in the cloud - the vendor's own cloud - meaning you have little control over the data or its safe transit. A better, safer option is being able to host voice services on-prem - within the security of your own networks - where you can better control what data is sent and held where.

Openness and interoperability

Another major consideration for many organizations - particularly those with multiple enterprise systems - is how the voice stack interoperates with other technology. Most voice assistant products commercially available currently are proprietary - they may have APIs and so on but the inside of the product is a 'black box'. A related fear for many CTOs and CIOs is 'vendor lock-in' - committing to a product selection also commits you to a vendor relationship that may not be competitive over time.

Voice Technology Key Terms

Device - the specific voice assistant Device, such as a Mycroft Mark 1.
Fallback - a Skill that is designated to be a ‘catch-all’ when the voice assistant cannot interpret the Intent from an Utterance.
Intent - when a user speaks an Utterance to a voice assistant, the voice assistant tries to interpret the Intent of the Utterance using an Intent Parser, and match the Intent with a Skill.
Skill - when a voice assistant hears a Wake Word, then an Utterance, the Intent Parser will try to find a Skill that is designed to handle the Utterance. The Skill might fetch some data, or play some audio, or speak, or display some information. If the Intent Parser can’t find a Skill that matches the Utterance, then the voice assistant will usually invoke a Fallback.
Speech to Text - the process of converting an audio Utterance into text that can be inspected by an Intent Parser.
Text to Speech - the process of converting text based information into voice-like audio, delivered through the voice assistant.
Utterance - an Utterance is how you interact with a voice assistant. An Utterance is a command or question - like “What’s the weather like in Kansas City?” or “Tell me about the Pembroke Welsh Corgi”.
Wake Word - the Wake Word is the phrase you use to tell the voice assistant you’re about to issue a command

Where do I go next?

To learn more about Mycroft AI's open source voice stack, visit https://mycroft.ai