Just Talk! Voice Controllers for the Internet of Things

IN THIS DOCUMENT

In less than a decade, tens of billions of electronic devices will be connected to the Internet-of-Things (IoT), a vast infrastructure of *smart* products that represent tens of thousands of different applications. Each of us will typically interact with hundreds of devices every day in order to control and monitor systems in our home/car/office, or communicate with friends and business colleagues. Many of these devices will be integrated into products that we're already familiar with – white goods, entertainment systems, security systems; others, such as smart speakers and domestic robots, will be new.

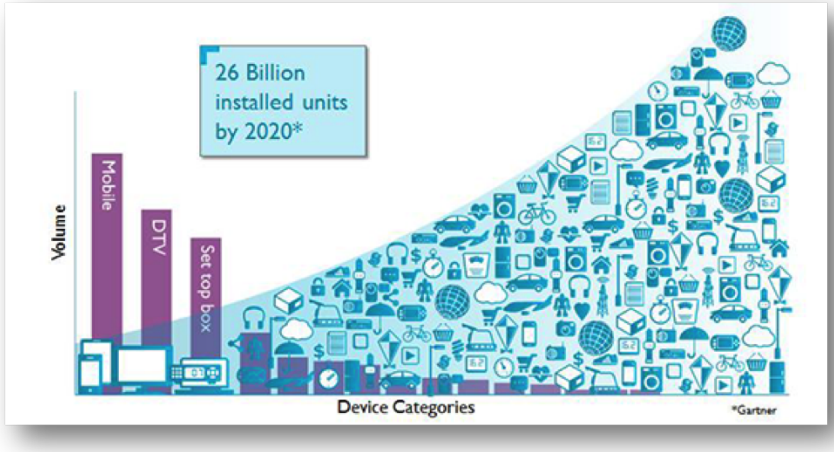

Figure 1: 26Bn connected devices by 20120. Source: Gartner

How will we control all of these smart devices? The contemporary user interface panacea – "There's an app for that" – does not scale to meet the challenge of controlling hundreds of devices in an efficient way. The most natural way for us will be to talk to the devices.

Publication Date: 2016/3/9
Document Number: XM010276A

1 The rise of Natural Language Interfaces

During the last half century, the electronics industry has progressed through a sequence of dominant applications from defense, through enterprise to the consumer-centric landscape we see today.

As electronic devices have proliferated through our professional and personal lives, user interfaces have evolved to enable the technology to be more easily accessible. Initially, we were expected to adapt to a machine's codified procedures – through special words and sequences on command line interfaces (CLI). Later, vendors attempted to enable a new class of user by inventing metaphors that resonated with our everyday life – windows, folders, trash cans etc – the Graphical User Interface (GUI, pronounced 'gooey') was born. More recently still, GUI technology has evolved to include touch by capturing gestures in two dimensions.

Until now, it has been sufficient for application designers to carefully craft user interfaces to facilitate the *user's learning process*. This trend is about to change – the Internet-of-Things represents a seismic shift from the dominance of individual applications to massive diversity of many types of device. It is not possible or desirable for us to learn how to interface with these machines individually, a revolution in UI is required - *the machines must learn their required behavior from their users*. The obvious interface for this learning is a Natural Language Interface (NLI) where the user uses words and phrases to interact with the machine and the machine can ask questions to determine the users intent.

As elegant and efficient as it is, the GUI still requires humans to learn a computer's language. Now computers are finally learning how to speak ours. [David Pierce, WIRED]

Innovators in Natural Language Interfaces started developing speech recognition software in the 1970s. More recently companies have been looking for ways to integrate speech recognition with voice capture and artificial intelligence, into consumer products. Smart phones were one of the first successful examples, with the introduction of voice based search in Apple's Siri and Microsoft's Cortana, both based on Natural Language Processing (NLP) and Text-to-Speech (TTS) technologies that run remotely on internet connected cloud servers. However, the smartphone as the universal remote controller is as impractical as it is unpopular. Hundreds of apps would be required to manage all the different devices, users would be obliged to carry their phones in the home, and IoT product vendors would be obliged to compromise the consistency, control and quality of their user experience.

With the majority of the 'heavy-lifting' of NLI done by the Cloud, economical Natural Language Interfaces will proliferate in all manner of equipment in the home. Initially in smart TVs and wireless speakers, followed by home automation and security, the technology will then move on to domestic white goods, as well as driving new product categories like digital assistants and domestic medical robots. A tipping point will emerge after which any machine that does not implement such technology will be regarded as primitive and undesirable.

2 Challenges for Natural Language Interfaces

Consumers are often suspicious of new technologies, and unprepared for change. Every advance in technology introduces new questions about security and protection of user personal data. It takes time to become confident with the integrity of a new technology before it is trusted.

Will consumers overcome their fear and inability to master technology? Google research in 2014 showed that 55% of teens and 41% of adults in North America already use voice search on a daily basis.

Just as our teenage children instinctively use gestures to interact with touch screen devices, so younger children use language before they learn to control devices with their hands. These younger generations are already learning to expect NLI.

Unlike previous evolutions of user interface, the concept of language is hardly new. Late adopters will naturally adapt existing language skills to control their devices, just as those devices adapt to the learned expectations of the user.

3 Marching forwards

Technology for *understanding* humans is advancing rapidly. Where responses to customer questions were very simplistic and patchy when Siri was first released, major advances have been made to NLP and TTS to understand a user's intent. Google Voice is already showing results for complex questions that require the speech recognition engine to make related decisions. Critically, these advances are primarily made by the Cloud services – enabling continuous improvements to be delivered immediately to all devices connected to the internet. Security is also advancing for Cloud based services, and the legal issues that surround personal data are being openly discussed (and challenged) worldwide. This is to be expected of new technologies that significantly change behaviours – but ultimately, the best user experience will win the day.

Voice interfaces enable products that don't have screens, buttons, dials - it signals the end of those dreaded hierarchical menus, it liberates significant real-estate on product enclosures. As well as the user experience, NLI will change the way products look. Does the technology for effective NLI exist?

For consumers to accept and adopt products with natural language interfaces, the communication must be intelligent and reliable – at least 95% accurate from the start. A core part of that performance is dependent on the ability of the product to capture the user's speech and respond quickly. In normal conversation, we generally don't expect to have to wait for several seconds for a reply, and so it will be with voice interfaces; when we talk to a product we expect it to understand our questions and commands and respond immediately – even if that response is to clarify our intent.

3.1 Four dimensions of voice capture

There are four key considerations for a high integrity voice capture for a great NLI experience:

First the ability of the product to capture a very high Signal-to-Noise Ratio by reducing background sounds and eliminating echo. In an environment with numerous hard surfaces like a house or office (engineers refer this type of environment as acoustically 'wet'), multiple reflected versions of the voice signal will be presented to the microphone that captures the signal. This echo and reverb causes interference in the signal, which needs to be removed before it's passed to a speech recognition engine.

Second, the voice interface needs to identify one voice among many other sounds. In a kitchen we might have a radio playing, washing machine running and other people talking. In an office or conference room there are likely to be many people talking as well as air conditioning and other ambient background noise.

Third, voice interfaces must work effectively in the 'far-field' (typically regarded as at least five metres), enabling users to interact with their digital assistant across a room.

Finally the product needs to capture our voice commands as we move around a room, getting closer to the capture device and further away. It's no use if we have to stand in the same place all the time.

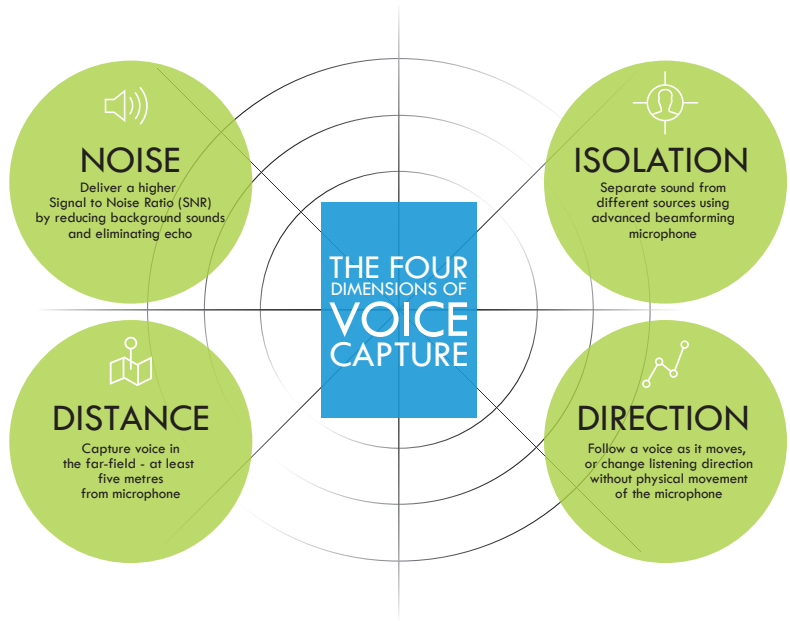

Figure 2: four dimensions of voice capture

The good news for voice developers, is that much of this technology already exists. Furthermore, whilst in the past the required compute performance was expensive, modern voice-user-interface enabled microcontrollers and Cloud services offer the potential for NLI to be affordably integrated into all smart home technology.

4 Conclusion

For all of us, language is the natural way to communicate with the many and varied smart devices connected to the Internet-of-Things. For early adopters of new technologies and younger generations, Natural Language Interfaces are already a reality, providing freedom to consumers who can be anywhere in range of a smart device. With NLI and voice capture technologies now economical for integration into consumer products, proliferation is inevitable. Existing electronic products will be updated with voice capture interfaces, while different types of products will emerge, many of which will have no visible interface at all. Just as language has enabled humans to communicate within a rich and diverse society, so natural language interfaces (NLI) where interactions are based on everyday words and phrases, will enable us to communicate with this infrastructure of IoT enabled products in an efficient and intuitive way. “How do I work this?” – the answer is simple:

Just talk!

XMOS®