theguardian

Voice recognition: has it come of age?

Apple's voice-controlled 'personal assistant' Siri has got everyone talking (in more ways than one). Is this just the start of a computer revolution?

Charles Arthur
guardian.co.uk, Sunday 20 November 2011 20.00 GMT

Scotty gets to work in Star Trek IV. Photograph: Allstar/PARAMOUNT/Sportsphoto

The man sits down in front of the computer and says, affably: "Computer!"

Nothing happens. In a now-hear-this tone, the man repeats: "Computer?" Still nothing happens. Puzzled, he picks up the mouse and speaks into it: "Hello, computer?"

Beside him, the impatient owner says: "Just use the keyboard." The first man replies, "A keyboard?" Then, slightly annoyed: "How quaint."

The scene comes from the 1986 film Star Trek IV, where Scotty, the engineer, and the rest of the crew have flown back in time from the 23rd century; Scotty needs to get some work done on the computer, and, of course, in the 23rd century they all work by voice command, unlike those 1980s throwbacks. Ha, ha.

Yet if the crew were to land 25 years later, in the present day, Scotty would still be just as puzzled at the computer's lack of responsiveness – unless, that is, he picked up one of the latest breed of smartphones, where being able to respond to the human voice has become the new frontier in interaction.

Since October, people have been buying and using Apple's new iPhone 4S, which comes with a function called Siri – a "voice-driven assistant" which can take dictation, fix or cancel appointments, send emails, start phone calls, search the web and generally do all those things for which you might once have employed a secretary.

Siri isn't just a "voice recognition" tool, though it can do that (so you speak some words and it turns them into text, and sends them as an email or text message). You can also ask it things such as: "How's the weather looking tomorrow in London?" and it will come back with the forecast for London ("England"). It'll do currency conversions or give stock

prices. Or try asking it: "Why is the sky blue?" and, after a little thinking, the screen will show an explanation: "The sky's blue colour is a result of the effect of Rayleigh scattering." (There is more, but we all know about Lord Rayleigh's work on molecular refraction in the troposphere, don't we?)

Lots of people think Siri isn't anything new. We've had voice-dialling on our phones for ages (where you say a name to the phone, link that to a phone number, and when you say it again it dials it). Google already offers a voice-driven web search app, which it backed with a big poster campaign around London earlier this year, with phonetic spellings of searches you might want to do - "pih-ka-di-lee sur-khus" or "tak-see num-buhz".

But experts say that Siri – and what it represents – might be as subtly revolutionary as the iPhone's multi-touch screen was when unveiled in January 2007. That's because Siri isn't just "voice dialling" or "voice recognition" (which tries to turn speech into its text equivalent); it's "natural language understanding" – NLU, in the lingo.

I've been testing an iPhone 4S, and found lots of uses for Siri: for doing currency and other conversions (sq ft in sq m, anyone?) or finding companies' stock values and market valuations (useful in stories) when a web search would distract from writing, and – in dodgier parts of town – to play songs or playlists (instructed via the headset microphone) without taking the phone out of my pocket. Better to be thought a bit strange than have the phone nicked.

Siri grew out of a huge project inside the Pentagon's Defense Advanced Research Projects Agency (Darpa), those people who previously gave you the internet and, more recently, a scheme to encourage people to develop driverless cars. Siri's parent project, called Calo (Cognitive Assistant that Learns and Organizes) had $200m of funding and was the US's largest-ever artificial intelligence project. In 2007 it was spun out into a separate business; Apple quietly acquired it in 2010, and incorporated it into its new phone.

When you ask or instruct Siri to do something, it first sends a little audio file of what you said over the air to some Apple servers, which use a voice recognition system from a company called Nuance to turn the speech – in a number of languages and dialects – into text. A huge set of Siri servers then processes that to try to work out what your words actually mean. That's the crucial NLU part, which nobody else yet does on a phone.

Then an instruction goes back to the phone, telling it to play a song, or do a search (using the data search engine Wolfram Alpha, rather than Google), or compose an email, or a text, or set a reminder (possibly linked to geography – the instruction, "Remind me to call mum when I get home," will work), or – *boring*! – call a number.

NLU has been one of the big unsolved computing problems (along with image recognition and "intelligent" machines) for years now, but we're finally reaching a point where machines are powerful enough to understand what we're telling them. The challenge about NLU is that, first, speech-to-text transcription can be tricky (did he just say, "This computer can wreck a nice beach," or "This computer can recognise speech"?); and second, acting on what has been said demands understanding both of the context and the wider meaning.

The demonstration that computers have cracked this – just as their relentless improvement cracked draughts and then chess – came in February this year, when IBM's Watson system competed in the American game show Jeopardy!, where the quizmaster provides a sort-of answer, and you have to come up with the question. (For

example: "Originally called What's the Question, it got its name when a producer at testing said it needed more jeopardies.")

Except Watson didn't compete against just anyone. It was ranged against two humans who between them had scores of wins, and won millions of dollars in prize money from the game. They battled it out over "answers" such as "William Wilkinson's An Account of the Principalities of Wallachia and Moldavia inspired this author's most famous novel." Yes, of course – Bram Stoker. Watson, and the humans, answered correctly, but Watson had more points, and won. Watson wasn't competing on absolutely level terms; it was fed the questions as text at the same time as they were read to the other two players. Still, its victory led to plenty of tweeting from people welcoming our new robotic overlords. And it wasn't even connected to the internet; it just relied on a huge database of information stored on its system.

"While competing at Jeopardy! is not the end-goal," its IBM engineers noted drily, "it is a milestone in demonstrating a capability that no computer today exhibits – the ability to interact with humans in human terms over broad domains of knowledge." And, of course, in understanding what people mean when they say something.

Having gained that glory, Watson has been shunted off to work on healthcare problems – with the addition of Nuance's speech-to-text technology. You can imagine it popping up on some future episode of House, solving some gnarly medical conundrum (and perhaps banjaxing the series' reason for existing).

But much more interaction like Watson's is likely because of the growing availability of "cloud computing", where huge amounts of processing power is available ad hoc over the internet. Amazon, for instance, now adds almost as much computing power per day to its cloud computing systems as it used to need for all its site in 2000. Other companies are doing the same.

That means NLU is starting to seep into our daily lives: we've grown used to computers getting better at understanding us in limited contexts – where you phone an automated system and can pay bills just by reading out your payment card number, or use booking systems that don't have humans. Now, it's spreading wider, being used for "semantic analysis" of Facebook and Twitter postings by companies eager to figure out whether people are saying positive or negative things about them online. Feed in the text, and the computer figures out whether people are happy or annoyed with you. It could even work on your car: no more fiddling with the dashboard.

"The key thing is NLU – understanding what you mean and what you want," says Neil Grant of Nuance. "Historically, cars have voice elements but historically you had to learn a huge long list of commands. As NLU progresses, you can say what you want in a way that's natural to you."

Norman Winarsky, who coordinated the funding for the Siri company, says it is "real AI [artificial intelligence] with real market use", adding that "Apple will enable millions of people to interact with machines with natural language. [Siri] will get things done and this is only the tip of the iceberg. We're talking another technology revolution. A new computing paradigm shift."

Francisco Jeronimo, smartphones analyst for the research company IDC, is less sure: "We've had voice recognition on phones for a few years, but it never really took off. There are two reasons for that: there was no brand or company that really pushed it the way that Apple is doing. And there's no other company with a brand like Apple. It was only when Apple launched the new iPhone with Siri that I tried voice search on my

Android phone – I'd never tried it before." Sounding a little surprised, he adds: "It worked really well."

In fact, what's happening now with NLU – which is spreading far beyond just the iPhone – could just be the beginning of a revolution in how we use computers (particularly smartphones) as big as that which came with the iPhone in 2007, when suddenly multi-touch screens (where you can do more than just prod the screen) were de rigueur. It's taken almost five years, but now multi-touch screens are everywhere: hundreds of millions of touch screen phones and tablets have been sold, and even the next version of Microsoft's Windows – due in about 12 months' time – will offer swooshy touch-screen operation.

Horace Dediu, who runs the consultancy Asymco, having formerly worked for the mobile phone maker Nokia, thinks the time is ripe for a new way of interacting with our computers. He points to how Apple drove changes in interaction before: the mouse and windows in 1985, the iPod's click wheel in 2001, multi-touch in 2007. And he also points out that Siri has some interesting similarities with what people thought about the original iPhone touch screen: "It's not good enough; there are many smart people who are disappointed by it; competitors are dismissive; it does not need a traditional, expensive smartphone to run, but it uses a combination of local and cloud computing to solve the user's problem."

Certainly there's been no shortage of people who say that Siri (which Apple meticulously points out is a "beta", or unfinished product) "isn't good enough" because it can't yet deal with every accent, or every possible query. They also say that the whole idea of asking a disembodied computer questions is ridiculous: "You won't catch me saying things into a phone" is typical of the sort of reaction. (An online poll at the website Ars Technica found 36% saying "you won't catch me talking at a machine in public".) Which is odd when you think about it, since phones are designed for talking into, and nobody seems to get embarrassed about announcing that they're on the train to a carriageful of people, or reading their credit card details out loudly and slowly for all to note. Perhaps there's an element of self-consciousness: we're more concerned with our own feelings about asking questions of a machine than anyone else's about hearing them.

The thing about NLU is people have been expecting it to happen for ages. In 1996 I watched Bill Gates announce that by 2011 we would have computers that could recognise their owners' faces and voices. That's this year! And he's right – if you count smartphones as computers which they are, really – about as powerful as the laptop computers of 2001. The newest Android phones can be unlocked by showing them your face. And the voice thing – well, we're working on it.

Not that things are perfect now either: Siri's servers have already had a number of outages. At Nuance, Grant thinks that will be solved: "Time will sort out the connection problems," he says.

Even so, it's not clear whether in our offices we'll talk to our computers, Scotty-style. Apart from anything, the noise would be maddening. Wouldn't it? Then again, that's probably what people thought in the 19th century about introducing telephones to offices. And we already have offices where people spend the whole time on the phone. We know them as call centres. Just hold on for a few years, Scotty. The computer will hear you soon.

• This article was amended on Monday 21 November 2011. The original said that 1986 was 35 years ago, rather than 25.