Your Voice Will Command The Cloud

Peter Marshall
CEO and co-founder, MeMeMe

Most people still think of speech recognition as just a geeky thing; something in the scientific realm, not really appealing or important to regular people. They may be right about today's consumer perception, but that's about to change dramatically. Speech recognition is nothing less than the next big thing in computing: about to become as prominent in our everyday lives as the Internet and mobile phones.

Surprisingly, the biggest blind spot regarding the status of speech recognition may be right within the expert community of speech recognition professionals: a cozy insider club which has been prophesying a boom in speech-enabled applications for so long that perhaps they've become immune to their own evangelism. Good speech recognition has always been just around the corner – a little like the Jetson's flying cars – but the science and technology have proven stubbornly difficult to make good enough for real people to use. As a result, the experts have become a little gun-shy, a little cynical, with a built-in mindset that it's still realistically a few years away, and built-in lowered expectations about the impact when it finally arrives. They have all the inside knowledge of the little successes and failures over the years that have made IVR ubiquitous but universally despised, and not very impactful except in lowering costs and making consumers angry. They know all the challenges that remain: all the real-world difficulties of accents, dialects, noise, homonyms, the rapid evolution of language, the problem of recognizing made-up names for products and titles, and much more – and that insider perspective blinds them to the approaching tipping point when we will all talk to our systems and forget how we could ever have done it any other way.

You see, fundamentally, speech recognition isn't just about replacing typing with talking. The experts have been trying to make that happen for so long that some may have forgotten what the real point was. Yes, it is true that talking is a more convenient mode of interaction than typing for many uses, especially when people are out and about and trying to do things where keyboards are difficult to use, such as on their mobile phones. And it's true that speech recognition has slowly evolved to the level of quality that real people would often be able to talk instead of type to send an email or search for a song or restaurant. It's even true that this added convenience will actually have a major impact on usage levels of many mobile phone apps, where typing is a disabler of usage, and voice interfaces are the needed enabler. This is, in fact, the tactical focus of my company – MeMeMe – and we, along with a few others like Vlingo, Nuance, Google and Microsoft are driving a real change in the expectations that mobile users have about how they use mobile apps.

All well and good, but in the end, it's still nothing more than an uptick in convenience if all that changes is that you sometimes type with your voice instead of the keyboard. That's not an insignificant gain, but it's missing the real opportunity. ***The real opportunity is how speech enables a revolution in the very nature of applications; a giant leap forward in how applications meet human needs.***

Everyone working in technology is aware that cloud architecture enables the intelligence of systems to be remote from the user, reversing the short-lived trend of the PC era that put system capability inside the user's personal computing device. The future is clear – almost all of the intelligence and capability of applications will reside in central servers, universally accessible over networks, and open to interconnection with each other in infinitely flexible combinations via APIs.

What is much less well understood, however, is how the cloud model is naturally tied to the evolution of smart mobile devices as the universal access point for applications, and the emergence of human language as the most powerful and flexible system to operate it all. When general spoken language becomes the operating system for the cloud, it enables the full potential of that cloud architecture. The utility of applications that are discrete, modular, and distributed is maximized when they are operated as a universe of completely interoperable Lego pieces, using an extremely flexible command language that can call for any action, respond to any prompt for commands and input, and describe any information to be passed from one module to another. The operating system has to be capable of dealing with an infinite and unforeseeable set of possible application commands and information requirements, because each modular unit can evolve very quickly, and independently of the operating environment, and respond to ever-changing opportunities for interoperation and integration. The only system capable of meeting this requirement for human beings is human language.

Many leaders in speech technology believe that achieving the potential of natural language interfaces requires major advances in natural language understanding. They foresee a point where speech will not only be recognized, but understood, which will guide the way systems respond to it. But this is a based on a basic misreading of the nature of applications, and how they use commands and input.

The fundamental advance of cloud architecture has been to decentralize applications: to enable them broadly to evolve in much the way living things do, in unforeseeable ways, and with an infinite diversity of implementations. And to be useful, applications must do something that requires special, proprietary information or processing; something which is not obvious, not freely available to everyone with no effort. They must contain and deliver some specialized information. This means that it will never be possible for a general-purpose model for understanding language to do a good job of handling the contextual usage of each application. Each useful application, almost by definition, will require its own proprietary model for understanding, and in fact the general model will break down at exactly those points where the application is of any use, i.e. where it encapsulates specialized information.

All applications are interactive – conversational if you will – because that is how people consume functionality. The user initiates an activity, the application responds, and based on context, prompts for further input, responds to it, moves to the next state, and continues the cycle until it is dismissed. It is not necessary, or even appropriate, to apply a general model for understanding the user's language as they interact with applications. All that is required to deliver the best possible speech interface experience is to apply the right contextual model for that specific application, and the right adaptation for the individual speaker, while maintaining the ability to call and exchange information with other applications. This is well within the grasp of current speech recognition technology; it requires only the right architecture. The speech interface must be available between any user and any app, i.e. right in the device that connects them – the new very personal computer: the mobile device. It must enable tight integration with the specific content of each application, but with a loose, open architecture that adapts easily to evolving applications and platforms. And it must be personalized to the speaker, to maximize recognition accuracy for the enormous range of voices, dialects, speech styles, and other individual factors.

At our founding, MeMeMe made two fundamental decisions, based on our understanding of where speech recognition fit in the overall architecture of applications, and what would be required for it to be effective.

First, we decided not to try to develop a Me-branded application to do things that people wanted to do with speech, like sending emails, doing searches, enabling a purchase, or transcribing notes. Because fundamentally, speech can be an interface to any application at all, and we could never be the best at everything. There will always be many better applications than we – with our primary focus on speech recognition – could deliver, and we don't want to compete with better apps. We want to give people the power to operate any app by talking.

Second, we decided that MeMeMe would be distinguished by adapting its recognition to each speaker and also to each application – the user and what they are using – because those are the two important pieces of information that help to understand what is being said. Each person has a distinctive and different voice and vocabulary. And each application has it's unique, specific commands, content, and language usage. Together, these MeMeMe adaptations, in a mobile speech interface with the right architecture for integrating to applications, can enable anyone to find and use any application anywhere just by talking. They voice-enable the world.

Speech recognition has the power to make the user into the operating system of the cloud – not because it replaces typing into apps with talking into the same apps, but because when speech recognition works well, it can deliver all the flexibility and power of language itself. Each application has just the capabilities that are coded into it, but speech enables an end user to have a dialogue with each application, and to command and combine them in any way that can be described. Speech is the most powerful and general tool that exists to define a useful interaction with the universe of applications, giving the user the power to create new behaviors on command. The entire model for computing changes when users gain all the power of language to operate their universe of applications.

About Peter Marshall:

Peter Marshall is the co-founder and CEO of MeMeMe (www.mememobile.com): a leading speech recognition platform, which provides highly accurate real-time voice operation of any application, adapting to each individual speaker and to any application they are using. Peter has a strong track record of successful leadership and innovation, and a foundation in advanced technology development and management. His previous leadership roles include: CTO of Peracon, the world's leading platform for commercial real estate sales; co-founder and CEO of The Identity Guardian, an identity theft protection service; and co-founder and CTO of Cipient Networks, inventing and building their breakthrough distributed event-detection and knowledge management platform. Peter also led and grew advanced technology consulting practices at KPMG Consulting and Siebel Systems, and managed IT projects and teams at The Walt Disney Co. and Cisco Systems. He studied Mathematics at UC Berkeley.