



Is Embedded Speech Recognition Disruptive Technology?

Author:

Jordan Cohen,
CTO, VoiceSignal Technologies

Synopsis:

*This article deals with the question:
"Is phonetic embedded speech
recognition a disruptive technology?"*

Remember the unused "speech recognition" feature on your cell phone? For almost as many years as cell phones have existed, manufacturers have tortured their customers with stone-age "voice tag" speech recognition systems allowing you to call 10 or 20 people by name by voice, after a training session. These systems would work if you mimicked the way you said each name during training, but the systems tended to fail in noisy environments or anywhere unlike where you trained the phone. The acoustic matching technology used (dynamic programming) was an algorithm well suited to the primitive acoustic models available in the early days of automatic speech recognition, but it was neither effective nor efficient in doing the job at hand – dialing the phone by name.

Speech recognition technology has advanced substantially since those early systems were designed. You can buy "dictation" software for your PC from multiple vendors that will recognize thousands of words, and which will capture, edit, and print a document by voice. You can call a voice activated assistant on the phone for banking, travel, ordering, and many other activities. The speech recognition algorithms used in these applications are substantial beasts, consuming thousands of MIPs and hundreds of Mbytes of memory, but they perform their tasks much more competently than the "train before talking" cell phone applications.

A new technology is phonetic embedded speech recognition. You can find it in speech activated dialing in some cell phones, PDAs, and other handheld devices. New telephone services are emerging which will make small devices into powerful communication assistants, and a competent speech interface will

make these services easy to use and possible to remember.

It is interesting to consider whether this new technology is disruptive. Clayton Christensen, in "The Innovator's Dilemma" says, "Generally, disruptive technologies underperform established products in mainstream markets. But they have other features that a few fringe (and generally new) customers value. Products based on disruptive technologies are cheaper, simpler, smaller, and frequently easier to use." Disruptive technologies eventually dominate their markets, as well as the markets of their better-performing alternatives. Is embedded speech recognition a disruptive technology candidate?

During the last year, algorithms based on modern embedded speech recognition have become available on many cell phones. They allow phone dialing by name or by number using your voice, and often allow voice activation of other cell phone functions. In these new applications, it is no longer necessary to "train" the system to your voice. The application understands how names, numbers, and other words sound in a particular language, and can match your utterance to a name, number or command in the phone. Users have found this new functionality straightforward to use and easy to remember.

There are more than 10 million phones that include these modern embedded speech applications. They work very well at calling the numbers listed by name from your phone book, at allowing you to dial your phone by saying the phone number, and at letting you look up a contact entry, launch a browser, start a game, and more. The applications are easy to use. Moreover, these modern instantia-



tions are similar to other disruptive technological advances in the use of techniques and computation, which, while very competent, are not at the bleeding edge of their respective developments. It is interesting to look at some of the history of this new technology and the associated emerging market forces, and then to speculate about its future.

Modern embedded speech recognition depends on advances in high capacity low power computing - of which ARM microprocessors are the dominant cell phone entrants. It also depends on advances in speech recognition, largely sponsored by the Government in the US and elsewhere from 1970 to the present (with a time-out during the 1980s), and an associated improvement in memory technology. In addition, the creation of mobile telephone networks and their associated small screen, button-poor, compute-rich hand-held devices has created a market which cries for voice-assisted multimodal interfaces. (Christensen identifies cell phones as a disruptive technology themselves.) It is possible that phones of the future will not need buttons, can have a screen or simply a broadband connection that can be associated with a screen, and can be worn, pocketed, in a handbag, or be otherwise nearby.

They will have new multimodal communications capabilities, which arise because of their broad connectivity, but they will be useable because of their multifunctional user interfaces.

A Few Bits of History

So how did these new speech capabilities develop? Here are a few of the advances that have led to this performance.

Low Power Computing

Over the past two decades, Intel, the primary manufacturer of microprocessors, has been boosting the memory size, speed, and power dissipation of its processors. Moore's law suggests that memory size (and compute performance) increases by a factor of two each 18 months, and Intel's chips have, until recently, been following that projection to a very high degree of accuracy. This may slow down because, as can be seen in the Intel projections of Figure 1, the operating temperature of the chips, which are smaller and faster, appears to be headed towards a very unpleasant regime. Solving this problem is beginning to dominate the designs of future chips.

At the same time, at Acorn, a British company, in 1983, a project to design a RISC computer came to fruition. The ARM

family of microprocessors was born and went through several iterations over the next decade. The ARM processor has become the ubiquitous small microprocessor and, with the introduction of the ARM6 in the 1990s, an immensely successful line of low-power high-performance chips and chip designs became available. These designs are now used in VLSI worldwide, and the ARM7, ARM9, and ARM10 are in current production in a multitude of cell phones. The ARM7 is similar in power to the PC of the early 1990s, with clock speeds between 20 and 50 MHz, and associated memories of 8 to 64 Mbytes. The difference is that they are much smaller and use much less power, and the memories are slow and non-volatile (mostly e-prom). We find them in all sorts of consumer appliances, from thermostats to home entertainment systems, but most of all they are in a very large proportion of cell phones.

Speech Recognition

During the early 1970s, the Defense Advanced Research Projects Agency (DARPA, an agency of the Department of Defense) established the Speech Understanding Research project to develop a computer system that could understand continuous speech. The \$3M per year funded groups at CMU, SRI, MIT, SDC, and BBN, represented the largest speech recognition effort ever, and it moved speech recognition from an university and laboratory curiosity, to a serious engineering endeavor. The research community in 1970 created a very early speech recognizer which recognized some words for some people hundreds of times slower than real time, and, although falling short of the original goals, it demonstrated that computers could recognize speech. During the 1980s, speech recognition techniques advanced from dynamic-time-warped acoustic pattern matching (like those early cell phone voice tag systems) to sophisticated algorithms based on statistical hidden Markov models. These HMMs are the ancestors of today's Phonetic Recognizers, and these techniques form the heart of almost all large scale dictations systems. Support continued through the 1990s.

As the community became better at actually accomplishing speech recognition, the tasks that they tackled became hard-

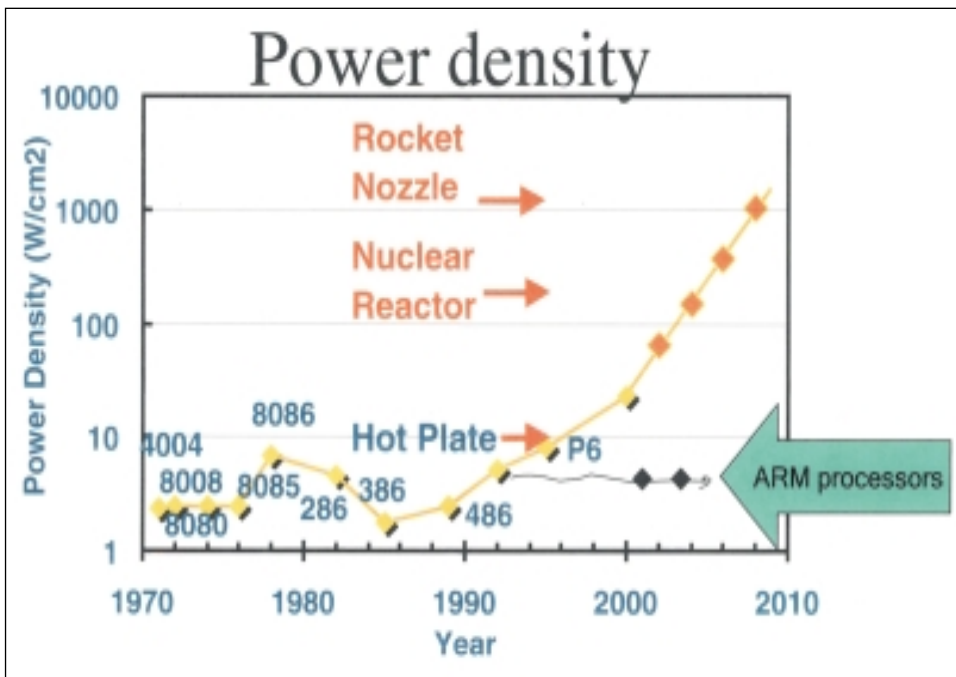


Figure 1: Power Density of Intel Microprocessor Chips with projections through 2010. Courtesy of Intel

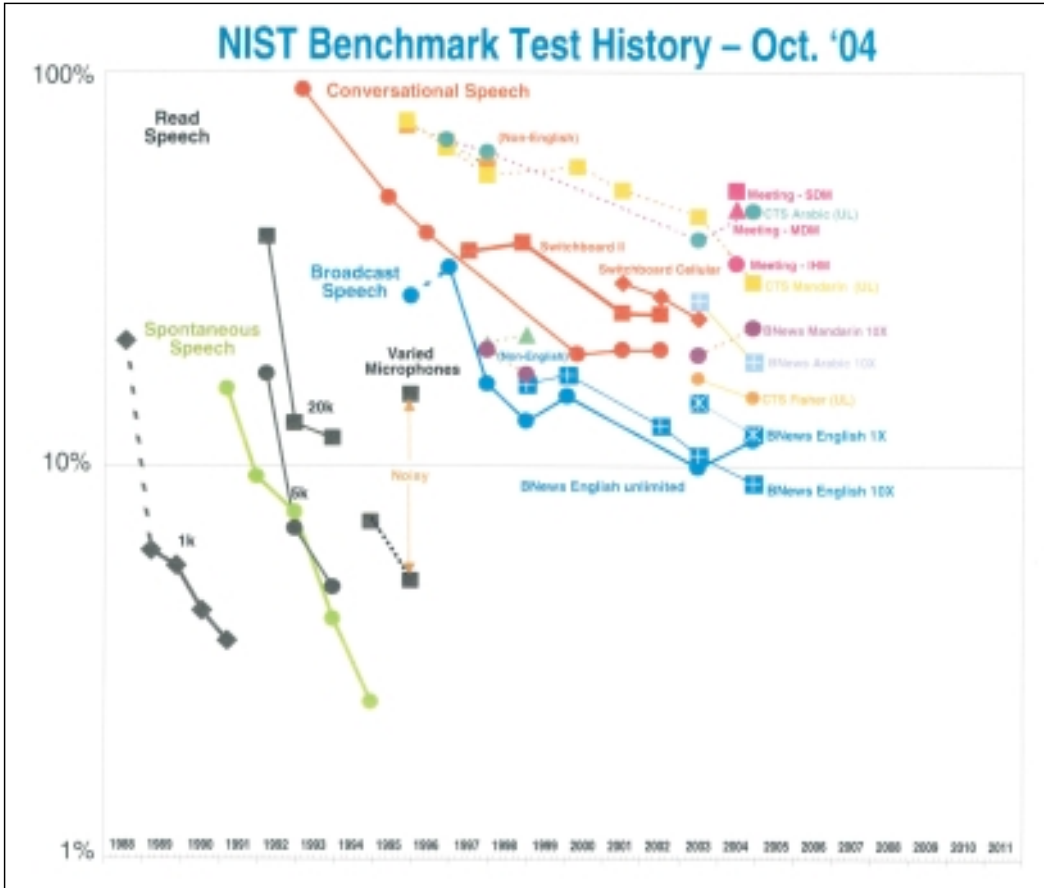
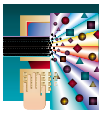


Figure 2: Speech recognition performance as a function of time for many government sponsored speech recognition projects

er. (It is difficult to improve a system that is 99.5% correct, and even harder to actually measure its performance.) As you can see from Figure 2, the community has continued its creative investigation of speech recognition algorithms by tackling increasingly harder tasks. On the other hand, in each increasingly difficult task, the researchers discovered statistical methods, ways to use data, and computational efficiencies that made previous tasks more tractable.

In our cell phones, we pose relatively simple questions to be answered by the speech recognizer. Although we did not know in the 1990s all of the tricks we know today, we can use 1990s-like computing resources (in other words, the ARM processor in your modern cell phone) to good advantage to compute a task which would have been difficult in 1990, but is simpler today because of our technical advancements.

Why hasn't everyone done it? Taking advantage of all of the knowledge learned to date, but using computation power equivalent to that from a decade ago, is not a simple task. It takes a dedicated research team and an equally dedicated engineering staff to make appropriate use of the modern algorithms. It requires the developers to internalize a nuanced feel for which advanced algorithms are possible to approximate, which are impossible, and to investigate the utility of each algorithm. Finally, you must develop a solution that is efficient and effective. In some sense it is an exercise in retrospective technology transfer, but it yields exciting returns. A few companies, such as VoiceSignal and ART (Advanced Recognition Technologies) have taken up this challenge. The speech recognition algorithms are running on hardware similar to that of a decade ago (except that is very power efficient and small), but the speech algorithms are almost state-of-the-art.

Once again, as in the case of the computation above, we have retreated from the cutting edge of speech technology, keeping in place those recent discoveries which can apply, and revisited a less demanding computing task with a less-than-cutting-edge (but still very capable) speech recognition. This shares the earmarks of a disruptive technology.

Is there a market?

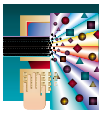
This technological achievement will be a success only if there is a market to receive the advances. The market is defined by a capability to deliver the technology, and a demand from the users of that technology which they are willing to pay to satisfy. Both of these situations exist.

I have argued above that we now have the technological capability to provide high quality speech recognition in your cell phone. Are there enough cell phone users to create an attractive marketing opportunity?

Certainly the carriers (Sprint, Verizon, and AT&T) thought so, because they provided voice activated voice-dialing services to their customers by subscription. Companies like Phonetic Systems provide these services to large organizations using large servers, and they have found a lucrative business providing directory services by voice. People find that voice access to dialing services makes sense.

One critical question must be "how many cell phones are there?" That is, what is the user population who might want this service in their phone?

There are many ways to count. Worldwide, the cell phone manufacturers have created about 600 Million new cell phones in 2004. The market analysts suggest that the number of new phones (possibly phones and PDAs, although it is not clear what the mix will be) will level off at about 900 Million per year, supporting about 2 billion cell phone users, in 2008. The majority of the phones will be replacements. These numbers could be



Modern Phones with Multimodal Speech Centric Interfaces



Figure 3: Some cell phones supporting Voice Dialing. The i700 is the second from the top.

conservative if the Chinese economy continues to boom.

The calculus from the technology perspective suggests that there are enough users. If the additional utility from speech is great enough to be worth something to each user, then something times 600 million can actually amount to real money. That makes a market.

Who needs Voice dialing?

The only remaining issue is whether or not people want voice dialing. For this discussion, we need only look at the local cell phone store to see what is happening to the technology of cell phones themselves.

Figure 3 shows several cell phone styles. Newer phones tend to be small. In fact, they are becoming so small that it is difficult to successfully dial a number using the keypad. As the world population ages, the ability to manipulate small devices decreases, making those keypads even more difficult to use.

For dome form factors, the keypads have all but disappeared. In the Samsung i700, a PDA/phone combination, the only numeric keypad is a touch screen that mimics a dialing touch pad. While it is possible to use and see that screen, one handed or no handed use is all but impossible. The voice interface is a substantial improvement. In the recent Xelebi line by Siemens, one of the phones (the Xelibre 3) has only one button! While it was possible to dial a number with that button, the

user interface can best be described as baroque. The speech recognition system became a requirement for any kind of usability.

Finally, in many states and several countries, it is illegal to hold your cell phone while driving. Since the majority of cell phone calls in the United States are made from an automobile, some form of hands free dialing will be essential. An embedded speech recognition voice dialing system would seem the right solution at the right time.

The convergence of low power computing, cell phone networks, smaller devices, and competent speech recognition thus offers a great opportunity. The sole remaining requirement for this technology is that it is worth something to the user – in other words, it actually must work! While I cannot speak for all consumers, it is interesting to hear what reviewers say about these services. I offer two reviews for your perusal.

Time Magazine, in its Gadgets column at Time.com, offered this view on February 4, 2004:

You know that voice-recognition thing on your cell phone that you don't use because it's too complicated and buggy? The one that asks you to record names, and then only auto-dials the names if you say them exactly how you recorded them? Now there's something infinitely better: it's called "speaker-independent" voice recognition, and the best one out by far is from Voice Signal Technologies... Each phone has additional

voice-recognition perks worth checking out, and the silly thing is, this million-dollar technology might already be in your phone. Have a look. If not, trade it in, because this stuff is awesome!

Further, in Engadget, Peter Rojas said on May 19, 2004

The speaker independent voice recognition MobileBurn describes sounds identical to the incredible software on my Samsung i700 Pocket PC Phone called VoiceSignal. ... VoiceSignal works so well its creepy. It has never needed a stitch of training and responds just as well to names like "Zolnowski" as it does "Smith." What I would like to know is this: how the heck are they able to achieve this on a handset while Via Voice and the others have been so bad for so many years — and on a desktop processors, no less?

Does that sound like a disruptive technology in the making?

It is clear that customers and reviewers are enthusiastic, and projections suggest that more than 20 million of you will be the owners of the new Voice Enabled phones before the end of 2004. If you want to try these new systems for yourself, go to your Sprint, Verizon, SK Telecom (Korea), or Bell Mobility (Canada) stores and find one of the new handsets. Information about the models may be found at www.voicesignal.com, or at www.artcomp.com, or at the web sites of the various carriers. Voice Signal's web site also contains a try-it-yourself demonstration.

And in the future...

As cell phones become the standard communications medium, their increasing capabilities will lead to even more interesting communications services. Voice services will include dialing, as well as application launching, web navigation, and even speech-to-text for messaging and email. The pressure for more capable user interfaces, coupled with the continued push towards smaller phones and lower power operation, suggest that the future is bright for speech recognition and its associated services. If these algorithms spread to other devices and to a wider industrial base, it could be disruptive. Will this be the technology of the future? Let us know what you think.

Jordan Cohen, jrc@voicesignal.com