Speech Recognition – an Overview

The technology first started appearing in the 1950s, and the first active word recognition systems were deployed in the 70s. Why has speech recognition only recently emerged as a viable commercial proposition? Stephen Coates reports.

"Three", was what Walter Cronkite presenter of the American science television show *The 21st Century*, spoke into a microphone connected to a computer to demonstrate then current research into speech recognition. The printer attached to the computer, however, typed "four". Dismayed, Walter uttered "idiot", to which the computer responded "not in vocabulary".

Fast forward a few centuries and in some distant galaxy, while a Federation Starship is engaging a Klingon space ship, Captain Kirk wants a status report. Instead of asking a crew member, he simply states "computer", and a very human-sounding voice responds. He doesn't have to press '1' for warp engines or '2' for photon torpedoes, he simply starts his command with "computer" and, no matter how much background noise there is, his subsequent questions are answered.

Although it has yet to reach the state depicted in Star Trek, speech recognition has come a long way since the late 1960s when *The 21st Century* was aired and although the number of major commercial implementations is still very small, speech recognition is increasingly being deployed in a diverse range of commercial applications.

As with any technology that can be considered to be, at least from a commercial perspective, emerging, the uptake can be expected to grow modestly as market understanding begins to develop. But it is difficult to forecast growth to be anything but modest when one considers that speech recognition is a technology that has been in development for more than half a century.

1920s

Although patents for phoneme-based (more on that later) word recognition systems date back to the 1920s, the earliest speech recognition system was probably one developed in 1952 by Bell Laboratories (now part of Lucent Technologies). The earliest attempts to develop functioning commercial speech recognition systems appeared in the 1960s. According to Peter Theis, president of USA-based ConServIT, a subsidiary of Conservational Voice Technologies Corporation, IBM announced in the 1960s that it had fully developed a word recognition system for dictation, so that you could speak the text and it would be converted to digital format, but that system did not appear to ever reach the market.

Theis added that in the 1970s, Eastern Airlines in the US installed a speaker dependent word recognition system for baggage handling, allowing the agent to speak the destination of the bag which would be routed properly. However, due to the number of errors, the system never progressed beyond trials. Perhaps one of Walter Cronkite's bags was lost.

Some commercial applications were deployed during the 1970s, though. Conservational Voice Technologies Corporation's first in line (no branching) natural speech system in Australia was installed in about 1975 at Taxis Combined Services in Sydney.

Used for cab dispatching, that system was featured on an ABC program at the time.

According to Dr. Koval, scientific director of the Speech Technology Center in St. Petersburg, Russia, the first commercial deployment of voice command recognition in a stand-alone device was by Threshold Technology in the USA in 1978. The system, a VIP-100 was used for post parcels sorting and quality control at the assembly line.

From its inception until only a few years ago, speech recognition was largely the province of the research lab, the reasons being that technology was less developed and that this research cost far more than the returns from the applications that were developed.

Even in the 1980s, there were very few commercial applications. According to Theis, American long-distance carrier AT&T announced in about 1985 that its call prompter would be changed from touch tone to word recognition in three months time. It never happened.

There have, of course, been commercial applications and at-tempts such as these for some time, but historically, speech recognition systems were either able to accurately identify a large vocabulary of words spoken by persons for whom the system had been 'trained', or a small list of words spoken by a much larger population.

Dictation Systems

The above-cited applications largely used limited vocabularies. The primary application of the large vocabulary for trained speakers has been dictation systems, offered by companies like Dragon (Naturally Speaking), which was recently acquired by Lernout & Hauspie of Belgium, and IBM (ViaVoice).

Nevertheless, the predominant applications for speech recognition are those that are speaker independent. A survey by Datamonitor found that 87% of commercial applications of speech recognition in the USA were for speaker-independent applications, with those in call centres the most popular (see Figure 1 on page 32). Given that directory assistance is provided by staff in call centres, such applications account for over two thirds of all commercial applications of speech recognition.

FROM ITS INCEPTION UNTIL ONLY A FEW YEARS AGO, SPEECH RECOGNITION WAS LARGELY THE PROVINCE OF THE RESEARCH LAB...

Although there are many reasons to implement speech recognition in a call centre, the primary one would appear to be to simplify the user interface with an application that, using DTMF (Dual Tone Multiple Frequency) commands alone, would be tedious and not customer friendly, yet retain the benefits

of having an IVR (Interactive Voice Recognition) system itself – primarily automating the self servicing of frequently requested transactions.

The most frequently-cited application is an airline call centre which may receive calls for bookings, departure and arrival times, lost luggage, special meals, frequent flyer points and much more. Using an IVR with only DTMF, a booking, for example, would take tens of steps.

Fortunately, most IVR applications are more straightforward and are able to be implemented using only DTMF commands, providing only numeric input is required. However, quite a number of corporations and government departments use customer numbers, account numbers and other key numbers that contain one or more letters as well as digits. The lack of a uniform set of letter assignments to numbers on Australian handsets and outright inconvenience are just two of the reasons that keystrokes are less than ideal as a means of entering such numbers.

Speech and DTMF

This brings us to an interesting point – speech recognition and DTMF input are not mutually exclusive, they are complementary. This is one reason that speech recognition is almost always implemented in an IVR platform. But there are others.

IVR systems have the audio circuits interfacing it to the PABX, allowing receipt of speech as well as DTMF commands. IVR systems have the host computer interfaces allowing information to be retrieved and updates to be performed. And IVR systems host applications that both allow callers to perform self-service and front-end call centre staff.

JUST ABOUT EVERY VENDOR OF IVRS WILL WAX LYRICAL ABOUT THEIR CAPABILITIES IN THIS AREA

So who offers speech recognition? Worldwide, there are between 20 and 25 developers of speech recognition software. According to Gartner, Nuance has the largest US-installed base of high-end applications and Philips has the largest number of European installations *(Speech Recognition for Contact Centers: Vendor Status*, 1999).

IBM, Philips and Speechworks are known to have a presence in Australia, as well as Syrinx which is based in North Sydney. Of these vendors, IBM and Syrinx include speech recognition within their own IVR/speech recognition systems whereas the other vendors license their speech recognition software to developers of IVR systems.

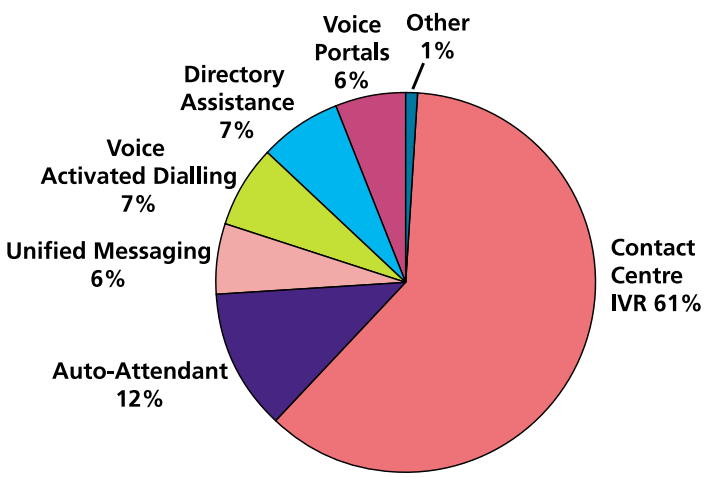
The next question is, who is actually using speech recognition? Although just about every vendor of IVRs will wax lyrical about their capabilities in this area, a survey of SR software vendors revealed very few signed contracts. The only such vendor who advised they had signed contracts was Speechworks, who had entered into contracts with Compaq Computer Australia, Information Technologies Australia and CTI Communications, each of whom distribute InterVoiceBrite, as well as VisibleVoice, Telemanagement and Pracom.

According to the report *CRM Strategy and Implementation in Australia and New Zealand* from Callcentres.net, about 4% of Australian call centres currently use speech recognition technology, about 28% are intending to adopt it in the next 24 months and 55% have not ruled speech recognition out. However, the SR vendors present in Australia could, amongst them, cite only ten functioning sites. Either there are other vendors here, or different means of counting have been used.

The other key question is, why, after 50 years of development, now? The answer is natural language processing.

Human speech is comprised of phonemes which are the minimum

FIGURE 1: TELEPHONY-BASED SPEECH RECOGNITION REVENUES BY APPLICATION: 1999



Source: Datamonitor

sound units that comprise individual words, some of which are roughly equivalent to single letters. However, according to Dr. K. Prasad of ICS Software in Bangalore, India, the English language uses 44 phonemes, requiring about 350 pronunciation rules to convert English text to synthesised speech. Speech recognition is inherently far more complex. Dr. Prasad added that all Indian languages can be synthesised from a total of 62 phonemes and that most Indian languages are phonetic, with a one-to-one correspondence between written letters and phonemes, making them easier to synthesise from written text to speech. Of European languages, Finnish is notable for being very phonetic. Presumably, speech recognition for such languages is also easier.

Phonemes and Languages

Donna Stapleton, Syrinx's marketing and communications manager, noted that some languages, such as some Polynesian languages have as few as 15 phonemes, and have simpler phonetic structures that make establishing speech recognition solutions easier. Stephen Lewis, general manager of Philips Speech Processing noted that "in general, speech recognition should be easier for 'regular' languages with only few pronunciation variants, for example Italian and Spanish. However, in practice, evaluation results do not reflect this. The only area where a difference can be seen in the quality is in the algorithms used for producing an automatic phonetic transcription of a word. This is more difficult to do in English than other languages".

Nonetheless, English is far from being the most difficult language to recognise. Sarah Neale, the marketing executive with UK-based Vocalis, Peter Theis, and Donna Stapleton nominated Chinese languages and Arabic, the Asian, Slavic and Russian languages, and Slovakian, respectively, as being more difficult to recognise and interpret.

Dr. Koval singled out Russian as having articulation that is more sluggish so the speech sounds are more smoothed, slurred and more variable and unstable. Except that Russian is the language of the synthetic type, where all grammatical forms are given by means of numerous, typically unstressed and therefore hardly recognised endings. Some Russian sentences may be understood only after voice melody detection and its classification.

Traditional speech recognition attempted to identify phonemes and from them, determine what words were spoken. An improvement on this has been to match pairs of phonemes to those known to be commonly spoken within the language. "Based on each phoneme that is said, algorithms analyse that as well as the previous and following phonemes that were said," says John Meiling, Philips' vice president of global marketing. "It then assigns a probability of what was said and matches it to words inside the host recogniser."

> WHEN DONE, AND DONE PROPERLY, SPEECH RECOGNITION CAN DELIVER IMPRESSIVE RESULTS.

Although different persons will give slightly different definitions, consultant Sara Chesters, speaking at the Executive Call Centre Congress in October 1997, defined continuous speech processing as the ability of a speech recognition engine to decode speech spoken continuously, in real-time, and natural language speech processing (NLSP) as the ability of a speech recognition engine to decode continuous speech that included pronouns. Although many vendors claim to support NLSP, the day one of them replaces their own telephonists with a speech processing system that greets callers "welcome to XYZ corporation, how can I help you?", I'll believe them.

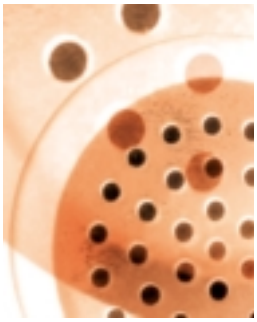
TABLE 1 – TELEPHONY-BASED SPEECH RECOGNITION REVENUES ($US M) BY REGION: 1999

	1999	2000	2001	2002	2003	2004	CAGR
North America	87	136	227	367	566	795	55.8%
Europe	36	53	87	150	239	352	57.4%
Rest of the World	10	14	22	39	80	156	74.5%
Total	**133**	**202**	**336**	**556**	**885**	**1,303**	**57.9%**

Source: Datamonitor

But they are getting close. Far more widely used is context-specific speech recognition. With this technology, the SR engine attempts to match the phonemes against a set of words and phrases that might be said in response to the prompt. This is a major step forward from the earlier directed applications which prompted a caller to "for sales, say 'sales', for service, say 'service'" and so on. And although context-specific speech recognition might not appear to be quite as user-friendly as true natural language SR, it has the advantage of, by using prompts, advising the caller of what the system can do. Prompting can also be phrased to discourage use of colloquialisms.

Although the potential to support multiple languages using speech recognition has its attractions, after one overcomes the small issue of how to simultaneously prompt the user to speak in the language of their choice, most implementations in Australia will be in English only.



When done, and done properly, speech recognition can deliver impressive results. At the above-cited congress, Chesters cited Charles Schwab, the largest direct stock broking firm in the United States, which had installed speech recognition in their call centres about two years prior to that year. At the time of her talk, Charles Schwab's speech recognition was processing 30,000 calls a day, with callers being able to choose any of 430,000 menu selections. The system was 97% accurate, correctly recognising speech on the first attempt. With only 10% of callers asking to speak to an agent, often because the system could not understand their accents, the average hold time had reduced from 60 seconds to 0 and the average call duration from 150 seconds to 142.

And accents may present a challenge for some time. Note that while Captain Kirk regularly spoke to the Enterprise's computer, Scotty did not.

There have also been a number of implementations in Australia, including:

- ComSec, a subsidiary of the Commonwealth Bank (Syrinx);
- Credit Union Australia (Speechworks);
- Regent Taxis (VeCommerce);
- TAB Queensland (VeCommerce);
- TD Waterhouse (InterVoiceBrite from CTI Communications, using Speechworks);
- WATAB (Syrinx).

Put the Menu in the RFT

What is common to all of these applications is that they automate a defined set of functions for a large call volume. Speech recognition is not for all call centres. It is expensive, with some vendors estimating a cost in the order of double the cost of a comparable DTMF-only IVR application, and that is on top of purchasing the hardware and software. The result is that speech recognition will be most applicable to larger call centres with high volumes of transactions that have the potential to be automated.

The ideal means of implementing a speech recognition application into an IVR system is not that different from that of implementing a DTMF-only IVR application. This commences with a com-

TABLE 2 – VENDORS OF SPEECH RECOGNITION SOFTWARE

Advanced Recognition Technologies (Chatsworth, USA, www.artcomp.com)
BeLingua Systems GmbH (Halle Germany, www.belingua.com)
Conversational Voice Technologies (www.cvtc.com, www.conservit.com)
Digitronic Computersysteme GmbH, Holm Germany (www.digitronic.de)
Entropic (Cambridge, UK, www.entropic.co.uk)
ETeX Sprachsynthese AG, Frankfurt, Germany (www.etex.de)
Fonix (Salt Lake City, USA, www.fonix.com)
IBM (Armonk, New York, www.ibm.com)
Lernout and Hauspie (Ieper, Belgium, www.lhs.com)
Locus Dialogue (Montreal, Canada, www.locus.ca)
Lyrix (Tewksbury, USA, www.lyrix.com)
Nuance Communications (Menlo Park, USA, www.nuance.com)
Philips (Netherlands, www.philips.com)
Speech Technology Center (St. Petersburg, Russia http://stc.rus.net)
SpeechWorks, previously Applied Language Technologies, Inc. (Boston, USA, www.speechworks.com)
Syrinx (North Sydney, Australia, www.syrinx.com.au)
Vocalis (Cambridge, UK, www.vocalis.com)

plete and comprehensive specification of the entire application, which is largely the menu, along with the processing of caller responses. Once this has been tested by role playing with agents, it can, preferably as part of a formal specification, be issued to a number of potential vendors, either informally or via a more formal tender process.

The advantages of this approach are:

- The buying organisation will have to determine and specify its requirements prior to making a purchase;
- The vendor will provide a price to implement a total, turn-key solution;
- The buying organisation will not have to concern itself by comparing various SR vendors of material on accuracy, success stories overseas or technical fact sheets – only the ability of the IVR vendor to deliver a working system need be assessed.

As straightforward as this might appear, it should not be assumed to be easy. Yet it is do-able. The author has prepared just such a specification for a client, but when vendors submit their proposals, they will be given the opportunity to recommend changes to the specified application.

Probably the most significant barrier to a greater uptake of speech recognition is a dearth of market expertise. It's new to vendors, it's new to consultants and it's certainly new to users. Nonetheless, as is detailed in Table 1 on page 33, Datamonitor forecasts significant growth in the adoption of speech recognition, with the 'rest of the world' grouping enjoying the highest compound annual growth rate.

Stephen Coates is an independent telecommunications consultant. He can be reached at swcoates@dot.net.au