NOW HEAR THIS...

Voice-controlled computers are no longer science fiction, thanks to the latest in speech-recognition technology. We report on this fascinating subject, review the latest software and look into the future

It's hard to believe but speech recognition and the idea of talking to computers to make them work has been around for at least 15 years. In much the same way that OCR (optical character recognition) was originally confined to large companies with very deep pockets, speech recognition was, because of the expense, confined for ages to highly specialised markets, such as the medical profession. We've all seen film of pathologists dictating their notes into a mike hanging over the morgue slab and, in fact, this was one of the original key markets for this leading-edge technology.

History

There were some early attempts to bring speech recognition to the masses but these were doomed to failure. Apricot produced a portable PC in 1985 with a hinged microphone by the side of the display, and users were able to perform simple dictated tasks in DOS. Well, at least that was the theory – most people could never get the thing to work and spent many a frustrating hour shouting 'DIR!' into its microphone.

The first truly usable PC-based speech-recognition programs came from speech pioneers Kurzweil (now owned by L&H) and IBM, with its VoiceType package. Both required special hardware in your PC and could only cope with 'discrete' dictation, that is, speaking with slight pauses between each word. While they worked well and delivered good recognition accuracy, having to talk like a Dalek to your PC for long periods of time wasn't a pleasant thing to do and so discrete was eventually superseded, in 1996, by continuous speech recognition.

Continuous recognition lets you dictate in a more natural, almost conversational style and speeds of up to 120 words per minute are achievable.

This way of speaking is much more difficult to 'understand' from the computer's point of view but the current speech products offer even better recognition rates, so speech technology has come on in leaps and bounds.

Most speech-recognition packages originally confined themselves to straight dictation. You'd dictate your words in a simple word processor, something resembling Windows 95's Notepad or Wordpad. Once completed, you could then transfer the text to a proper word processor for final editing and fancy formatting.

The latest speech-recognition packages go beyond this: most let you dictate straight into Microsoft Word 97 and many other Windows programs. Tight integration with a word processor also allows you to perform formatting commands by simply speaking them, eg, 'Select paragraph; make bold'.

The new breed of speech packages also offer 'command and control' of your PC. This lets you use Windows 9x and, say, Internet Explorer 4, simply by talking to your PC – you say 'Load Quake' and Quake loads.

How it works

The basic technique behind speech recognition is pattern recognition: you say something into a microphone; your speech is quickly converted into digital data and the computer compares the patterns it sees with those stored in memory.

The more memory it has the more words it can recognise quickly. The basic process hasn't changed in over a decade: your PC compares what you've just said to everything it knows in order to work out what you are on about.

All speech-recognition software uses a series of very complicated mathematical formulae or 'algorithms' to crunch your voice data. Some use a simple technique called 'template matching' but most rely on something a lot more sophisticated, with the fearful name of the Hidden Markov Model or HMM.

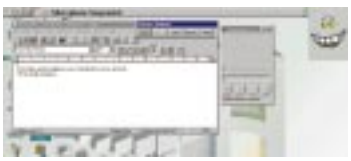
Mind-boggling

Developed in the late 1970s by the US Institute for Defense Analysis and IBM, HMM is what's known as a 'predictive model' – it uses mind-boggling statistical techniques to determine the probability of something happening.

Speech-recognition products

IBM ViaVoice 98 Home

ViaVoice 98 is available in two versions, the Home Edition and the £140 Executive Edition. Both have the same 32-bit recognition engine and support continuous speech recognition. They also offer direct dictation into Word 97 and the IBM WordPad lookalike, SpeakPad. ViaVoice also offers natural language commands for Word 97.



It is perhaps the most sophisticated speech package on the market, eg, it will correctly format 'twenty pounds' as £20. It's also 'modeless,' meaning you can switch from dictation to command and control simply by pausing. However, it is resource-hungry and enrolment can be a lengthy process. But you'll get high recognition rates with ViaVoice 98.

- *£49 (inc VAT)*
- *IBM: 01705 492249*

www.software.ibm.com/is/voicetype

IBM ViaVoice 98 Home

Overall ★ ★ ★ ★ ★

Dragon Systems NaturallySpeaking

The latest release of the award-winning NaturallySpeaking is available in several flavours; Standard, Preferred and Professional. All three feature Dragon's new BestMatch technology, which offers improved recognition accuracy dependent on your processing power, plus support for natural language commands in Word 97 – this lets you phrase the same command in several ways and it will still be recognised. Unusually, WordPerfect 8.0 is also supported. It is not quite as clever as ViaVoice but just as accurate and somewhat easier to set up and enrol.

- *£92.82 inc VAT (Standard version)*
- *Endeavour Technologies: 01932 827324*

www.dragonys.com

Dragon Systems NaturallySpeaking

Overall ★ ★ ★ ★ ★

Dragon Systems Point and Speak

A cut-down version of NaturallySpeaking, Point and Speak offers continuous speech recognition only, no command and control.

- *£45.82 (inc VAT)*
- *Endeavour Technologies: 01932 827324*

www.dragonys.com

Dragon Systems Point and Speak

Overall ★ ★ ★ ★ ★

Philips FreeSpeech 98

FreeSpeech98 is a well-specified debut for Philips. It's a general-purpose continuous speech package that lets you dictate text straight into most applications, not just Word 97 or a WordPad lookalike. It can also a function as a command and control package, too, letting you navigate Windows 98 without touching the keyboard.

Unusually, FreeSpeech 98 is available as a 30Mb download (no, really) from the Phillips Web site. This is a strictly time-limited version, restricted to just seven days of life. Another novelty is that FreeSpeech 98 is being sold exclusively over the Internet and won't be available as a retail product per se. This means, of course, no bundled microphone headset and no printed documentation, bar a 45-page manual in Word format you can print out.

- *$39 (only available via the Internet)*

www.freespeech98.com

Philips FreeSpeech 98

Essentially, HMM is used to make educated guesses about the true shape of an audio sound wave when it doesn't have all the sound data. For example, we can all make out the words when someone speaks to us in a noisy pub but a computer can't discriminate in the same way a human can. HMM is a way of subtracting the noise from recorded sound to leave just the speech.

But merely identifying word sounds isn't enough – there are loads of words that sound the same but are spelt differently, for example 'rowed' and 'road'. Therefore, speech-recognition programs also have to rely on the context of the word in a sentence.

So in this example it would know from its grammatical rules and language constructions that 'I rowed my bicycle' would be wrong. You can see this context analysis happening all the time when you use a modern continuous speech-recognition program – the program never 'commits' to your last spoken word and will typically change it, often several times, depending on what you say next, as the meaning becomes clearer.

For this reason it is perhaps best not to look at the screen while dictating, as you can be several sentences down the road before your PC has finally made its mind up about what you have actually said.

Training

All speech-recognition programs rely on you training them to recognise your voice, a process known as enrolment. Only when the PC is fully acquainted with your pronunciation and the way you speak will it be able to successfully recognise your words. In fact, for most programs, enrolment is a never-ending process because to maintain good recognition accuracy, it is essential for you to correct every mis-recognised word. So every time you use a new word, this has to be added to the vocabulary. Most training takes the form of you having to read out a series of sentences, which can take between 30 to 60 minutes. This is a tedious process – not only is it boring but you feel very self-conscious talking to your PC! Once this is completed your voice data has to be analysed and crunched – this can take anywhere from 30 minutes to several hours.

Recognition performance and hardware

As well as having clever recognition software, there are other things you need in order to get high levels of speech-recognition accuracy. Once upon a time, speech recognition was heavily dependent on having the right sort of hardware in your PC. Not only did you need a fast processor to perform the awesome number crunching required, but you often needed a special sound card, too, typically one fitted with a Digital Signal Processor (DSP).

Times have changed and every PC that ships today has more than enough processing oomph to cope with speech recognition. (One point here: although the minimum hardware specifications quoted for the software might be a Pentium 133MHz, we would recommend nothing less than a Pentium 200MHz MMX – MMX support being particularly important). Another thing that's gone by the board is the need for special sound hardware – today just about any decent sound card will do, though make sure you turn off any special effects, eg, 3D surround-sound.

Perhaps the biggest hardware requirement these days is memory. In order to speed up recognition, these programs store their vocabularies in memory rather than on disk, simply because it's faster to get at them. Most current speech-recognition programs have 50,000- or 60,000-word active vocabularies held in memory and will run happily in 32Mb of memory; 48Mb if you want to run Word or other applications.

However, in order to increase recognition accuracy, vocabularies will have to be expanded and with it, PC memory. So, if you want to run Word 97 plus your specialist medical or legal vocabulary, you're going to have to start slamming in the RAM, perhaps to as much as 64Mb, maybe more. Some experts have predicted that to handle a truly comprehensive vocabulary of 200,000 to 500,000 words a PC needs no less than 100 to 500Mb of memory!

Microphones are also very important and to ensure good results, most speech-recognition software comes complete with a microphone headset. The microphones must have high sensitivity and fidelity but unfortunately these virtues also make them very susceptible to the vice of noise. For the best results, especially in noisy environments, you need to use highly directional noise-cancelling microphones. This need will become even more pressing when speech recognition starts turning up as a standard feature in cars.

Overall accuracy depends on many factors but it's possible to hit 95% accuracy with a little perseverance – sadly, 100% isn't achievable at present. Initial accuracy will be lower than this but because speech-recognition packages continually learn, the accuracy scores will improve the more you use a program, provided you always correct mis-recognised words.

The shape of things to come

We're now so used to seeing people talk to computers in the movies that we take it for granted that it is one of the more believable science fiction predictions – after all, 2001 isn't that far away. And to be honest, today's speech-recognition technology is almost 'there', almost good enough for everyday use. And it isn't just some 'blue sky' gee-whizz technology either, it is a genuine productivity tool – if you're a two-finger typist, it'll boost your typing speed no end. And if you suffer from RSI or are disabled in some way, then speech recognition will allow you to fully use a PC as though you were bashing away at a keyboard.

Today, speech recognition is getting closer and closer to being an everyday thing. You can now buy a Philips Genie cellular phone that will dial 'Mick's' number when you say 'Call Mick'. And it won't be long before domestic appliances such as VCRs, hi-fis and microwaves will also respond to spoken commands. Perhaps the greatest speech developments are taking place in cars – by 2000 many cars will feature voice-control options – you'll be able to perform simple commands by simply saying them, eg, 'open sunroof' or 'turn heating on'. Another area of our lives where speech recognition will make a big impact will be automated telephone services. Pretty soon, when you call your bank, airline reservations centre or even BT, it will be hard to tell if there's a real human at the other end of the line or some very clever interactive voice response system.

Have a nice day

The US department store, Sears, Roebuck & Co is rolling out speech-recognition automated switchboard systems in 700 stores across America. When you call Sears, your call is answered on the first ring. You're then asked to 'speak the name of a department' you want to be put through to. If your response isn't recognised it'll ask you 'did you say books?' If a customer says 'no' or the system doesn't recognise a command, the customer will be immediately routed to a live operator.

According to Sears, the system will work properly about 60% to 70% of the time, rising to about 90% when customers say words the system is programmed to recognise.

This side of the Atlantic, BT is currently experimenting with an automated directory enquiry service, called Brimstone.

The cost savings in this area alone can be huge – in the US, AT&T's partially automated directory service has an automated voice operator that asks callers 'What city? What listing?' before connecting them to a human operator. This saves AT&T some $200 million to $300 million a year in labour costs.

These speech-recognition systems will work slightly differently to PC-based systems in that they don't attempt to recognise every single word. Instead they try and recognise as many words as they can but then base their response on certain key words they pick up in a sentence. This means a caller can express the same instruction in several very different ways and still be understood, eg, 'I want to fly to Rotterdam' and 'I want to book a return to Rotterdam'. This in turn can prompt further questions by the system to make sure it's understood the caller perfectly.

As well as recognising instructions and requests, such systems will also eventually be able to positively identify you from your voice, a feat that has major implications for things like online banking. You'll be able to ring up your bank and rather than have to remember the third letter of your password and your mother's maiden name, you can just speak your instructions and the system will simultaneously verify your identity and carry out your commands.

Roger Gann