SPEECH RECOGNITION

Research undertaken by Ability Technology for the Australian Consumers' Association March 2001

Controlling your computer by voice – sometimes it seems like a train we see in the distance that never seems to get any closer. There have been improvements in recent years, but do the latest speech recognition packages bring the dream closer to reality? How do the latest speech recognition systems compare? We put the three systems currently available through some rigorous testing. The results were interesting…

What is speech recognition?

You speak into a microphone and the computer translates or interprets your words into text in your word processor or other application. That is the central promise of speech recognition. Early speech recognition programs made you speak in *staccato* fashion, insisting that you leave a gap between each word. You also had to correct any errors virtually as soon as they happened, which meant you had to concentrate so hard on the software that you often forgot what you were trying to say!

The new systems are certainly much easier to use. You can speak at a normal pace, without leaving distinct pauses between words. However you can't really use "natural speech", as claimed by the manufacturers. That would be too sloppy for a computer to interpret. You must speak clearly, as you do when you speak to a dictaphone or leave someone a telephone message. Remember – the computer is relying solely on your spoken words. It can't interpret your tone or inflection, and it can't interpret your gestures and facial expressions, which are part of everyday human communication.

The systems look at whole phrases, not just the individual words you speak. They try to get information from the context of your speech, to help work out the correct interpretation. This is how they can (sometimes) work out what you mean when there are several words that have the same sound (such as "to", "too" and "two").

Some of the new systems also allow corrections to be done later. They allow commands (such as opening programs) to be done by voice and also formatting to be done through voice commands (such as *Make the previous paragraph bold*). You can capitalise words, insert numbers and move to different parts of your documents. New speech recognition systems offer better features than their predecessors and they are far easier to use.

What speech recognition is not

Contrary to some popular beliefs, speech recognition is not really a shortcut to using the computer. While it can make computer activity, especially the entry of text, more efficient, it also presumes you know your way around your system. If you are entering text into a word processing program, you will still need to know how that program works in order to produce effective written material. Otherwise it would be like learning how to control a car without having ever been out on the road.

Speech recognition (SR) is best seen as a specialised application, not a shortcut. If you are new to computers, you would probable be better off becoming familiar with your operating system and word processor first.

You should also consider if you really want to dictate text into a computer. Some people find it difficult or unnatural to create written material this way, in the same way that many people don't like to record notes on a dictaphone. If you are half-hearted about speech recognition then it is likely you will not put in sufficient effort to produce a good result. Successful speech recognition requires some adjustments to the way we work, and not everyone is prepared to do that.

Getting started – what computer do I need?

If you believed what was written on the boxes, then you'd come to the view that the computer requirements of these programs are quite modest.

Dragon Naturally Speaking Preferred asks for at least a 266 MHz processor, 64 Mb RAM, Windows 98, 2000, Me or NT, and 150 Mb of free hard disk space. You will also need a CD-ROM drive and a 16 bit sound card.

IBM Via Voice Millenium Pro requires a 233 MHz processor with MMX, 48 Mb RAM, Windows 95, 98 or NT, and 340 Mb hard drive space. Plus a CD-ROM drive and a 16 bit sound card.

Philips FreeSpeech 2000 requires a mere Pentium 166 MHz processor with MMX, 48 Mb RAM (64 Mb when using *Microsoft Word*), Windows 95, 98 or NT, 100 Mb hard drive space, a CD-ROM drive and a Sound Blaster or compatible sound card.

Our testing showed that these specifications are underdone, particularly with regard to RAM. We discuss this later in our section on **Memory Tests.**

Two of the systems come with a headset microphone, whereas *FreeSpeech* has a handheld unit that incorporates a microphone and small trackball. While this was easy to hold, our testers became tired while holding this unit for lengthy periods. We tested the performance of these and several other microphones. See the results in our section on **Microphones** a little later.

Getting Started – Training

Unfortunately you can't just start speaking into the microphone. There are some preliminaries. Every person's voice is different, so the systems need some more information about the way you speak before they turn you loose on the task of dictation. Each of the systems requires you to undertake a process called 'enrolment'. You are asked to read some sentences and letters (in *FreeSpeech*) or some more interesting extracts from books (in the other two). This process takes as little as 5 mins (*Dragon*) or 10 minutes (*ViaVoice*), up to 45 minutes (*FreeSpeech*).

Each of the programs allows you to have some of your existing documents "analysed" to identify unknown words and peculiar expressions you use. This can speed up the process of allowing the system to get to know your speech characteristics.

Overall our testers felt *Dragon* was the easiest and quickest to set up. *ViaVoice* wasn't far behind, and probably had better introductory resources, such as a video on CD. But the training process was at times frustrating, as the marking of

errors seemed to lag considerably behind the user's speech. This seemed to cause the program to identify multiple errors and slowed the process down. *FreeSpeech* was not difficult to set up but the interface was not intuitive at first. Its manual was slim (63p as against 140p for *Via Voice* and 214p for Dragon). Its training passages were also boring. But its toolbar is customisable – quite a handy feature.

All of the programs have introductory training resources and on-line help.

Correcting Mistakes

The accuracy of these systems improves quite quickly *if you take the time to correct mistakes.* This is important. The voice model created after your 'enrolment' is constantly changed and updated as you correct misinterpretations made by the systems. This is how the programs "learn". If you do this properly then the accuracy you obtain will improve. If you don't, the accuracy will deteriorate.

There are three types of corrections you will need to make when you are editing text. The first is when you cough or get tongue-tied, and the word comes out nothing like what you intended. The SR systems will make an honest (if not sometimes humorous) attempt to translate your jumble. The best suggestion here is to select the word and then say the word again properly in place of the mistake. Or just delete the word and start over.

The second circumstance is when you simply change your mind. You said "this" but you now want to say "therefore". You can make these changes any time, because the speech recognition system has not made a mistake. You simply change the word (by typing or by voice).

In both of the first two cases the SR software has not made a mistake. In the third type of correction the software gets it wrong. If you said "this" and the system interpreted the word as "dish", then you need to go through a correction procedure to enable the system to learn from its mistake. You can't just backspace and try again, tempting though that might be.

Why is this? The reason is that modern speech recognition is not based on individual words, but on small sounds within the words. Information gained about the way you say "th" will be generalised to other words with a "th". Thus if the system misinterprets the word "this", it does not mean the error is restricted to that one word. If you don't correct the error, other "th" words might be affected. If you do correct the error, then the new information is also spread to other words, improving accuracy. Your voice model will always be getting better or worse. *Th*is is some*th*ing to *th*ink *th*rough *th*oroughly!

In each of the programs, once you enter the correction process, you are offered a list of alternatives in a correction window. If the word you said is in the list, you simply select the number next to it (by mouse or voice). If it is not there, you spell the word by keyboard or voice. New words are added to the active vocabulary automatically. You are automatically asked to train the new word in *ViaVoice*; it is an option in *Dragon* and *FreeSpeech*. If you are dictating a word that is unlikely to be in the program's vocabulary, you can enter a spell mode and spell the word out; this feature is available in all of the programs.

We found *ViaVoice* had the best correction system. If you are doing a series of corrections, *ViaVoice* allows you to make the correction window stay open. With the other two programs the correction window is opened on a word-by-word basis. In

DragonPad a *Quick Correct* window opens up automatically when you select a word (this feature can be turned off if you find it distracting).

ViaVoice and *FreeSpeech* allow you to save speech session data so that you (or another person) can make corrections later (this feature is only available in the more expensive *Professional Edition* of *Dragon)*.

In *Dragon* the default correction method is by voice. You select the word or phrase to be corrected, then say "Correct That". However some of our testers were then taken to the word "that" for correction. There is an option in *Dragon* to create a hot key (such as F12) to be the correction key, or to make the correction window pop up by double-clicking the word. This made correction a lot easier. In *FreeSpeech* you have to select the word and then click the Correction button on the toolbar (or choose it by voice) on each occasion.

The types of mistakes made by speech recognition programs are quite interesting. In our tests the SR programs nearly always got words like "erroneously" or "spokesperson" correct. These larger words are distinctive and therefore easier for the systems to recognise. But they all had problems distinguishing "or" and "all". And they nearly always interpreted "or ordered" as just "ordered", perhaps assuming a stutter at the start of the word. We were disappointed that in very few tests was the phrase "two invoices" interpreted correctly, in spite of the plural noun. We almost always got "to invoices".

Some humour

No study of speech recognition would be complete without a sample of the sometimes humorous output created by the systems. Part of our test was the first stanza of "Advance Australia Fair". The multicultural influence was noticeable, with "Thailand abounds in nature's gifts" and "And France Australia fear!" The business influence was also evident, with lines such as "of BT rich and rare". Then we got "our home is Goethe by sea", "endure a false drains", and the frivolous "with cold and soil and wealth for twirl" and "four-wheel a young and free".

Environmental concerns made an appearance, too, with "A land of dams in nature's gifts, of beauty which and rear, in his trees page, dead every stage". For *Advance Australia Fair* we got "an fence or straighter fear", "advance a stand of their", "events of stellar effect it", "advance bus journey affair" and "and ends Australia there".

What if I have impaired speech?

Modern speech recognition systems tolerate wide variations in speech types and accents. However people with speech impairments need to be careful when purchasing SR software. One of the staff at the Ability Research Centre, where the testing was conducted, has impaired speech as the result of cerebral palsy. He can usually be understood on the telephone, with some concentration by the other person, but he was unable to get to first base on any of the three systems. If you have a speech impairment, the lesson is – try the system before you purchase it.

How do I add my own words?

The SR programs come with pretty hefty active vocabularies that are loaded into RAM and used instantly available for dictation. They also have large backup dictionaries that come into play when you do corrections. *ViaVoice* has 100,000 words in the active vocabulary (expandable to 164,000), and a 260,000 word dictionary. *FreeSpeech* has 27,000 words in active vocabulary, expandable to 64,000. It also has a 290,000 backup dictionary.

But they all allow you to add your own words to the active vocabulary. In *ViaVoice* you can add up to 64,000 words to the active vocabulary, while in *FreeSpeech* you can add up to 37,000 new words. If you say a new word, then it will be added to your active vocabulary once you correct it. Dragon Naturally Speaking Preferred has 160,000 word active vocabulary, with a backup dictionary of 250,000 words.

Allowing the programs to analyse your documents to find new words is another way of adding your own words to the vocabulary. This is available in all three programs.

FreeSpeech and *ViaVoice* allow you to have additional custom vocabularies. This feature is only available in the more expensive *Professional* version of *Dragon Naturally Speaking.*

What are 'voice macros'?

As well as dictating text, commands and formats, these SR programs allow you to create and apply voice macros. These are where a single voice command inserts a block of text into your document or implements a sequence of keyboard or mouse actions.

All of the programs allow you to create text macros. For example, you may say "My-address", and out will pop the address you have previously stored for this purpose. You must be careful to say the macro name as one word: if you leave a pause between the words, you will get the words *my address* instead of your macro. Macros can be very helpful when you use standard text frequently. *Dragon* limits the size of text macros to 1000 words. No limit is mentioned by the other programs.

In *FreeSpeech* and *ViaVoice* you have the capability of creating your own voice commands involving keystrokes and mouse movements. In *Dragon* this feature is again reserved for the *Professional* edition.

How do the programs distinguish between dictation and commands?

It is important for the software to be able to work out whether what you are saying is text or a command. The three programs tested deal with this issue differently. In *FreeSpeech* the microphone unit has a button that activates Command mode. You must hold this button down while speaking a command.

Both *Dragon* and *ViaVoice* try to distinguish commands automatically. You leave a short pause before and after saying the command. They are surprisingly good at this, although they can be caught out at times. We had one experience where *ViaVoice* got locked in Command mode and wouldn't interpret any dictation at all. In acknowledgement of the possible problems, both allow you to enforce the command mode more definitely. In *Dragon* you hold down the CTRL key to force acceptance

of a command. In *ViaVoice* you can activate an option to preface all commands with the word "Computer", to help prevent ambiguities. You can also make some selected commands inactive.

What programs can I use with speech recognition?

All three programs work with Microsoft Word. *Dragon* and *ViaVoice* also have their own simple word processor (Speech Pad or DragonPad) as an option.

You can use these programs in virtually all Windows programs. They work reasonably well in spreadsheets like Excel, Internet Explorer and Microsoft Outlook. But the full features of the software are not always available in these other programs. Speech recognition is at its best in word processing.

ViaVoice has a handy feature whereby you can say "*What can I say?*" in order to be shown a list of available commands in your current program. In *FreeSpeech* you can go to the complete list of commands in the *Command Explorer*. You can even do a search for a command in this window, but its options are not contextual. Similarly in *Dragon* the command "View Command List" takes you to the full list of commands.

Speech recognition and RSI

Speech recognition is often touted as an option for people suffering from overuse injury, commonly referred to as RSI. It can be helpful in such situations, as it reduces the number of keystrokes required. However ambient noise can be an issue, as can privacy and possible disturbance to other workers in modern compact offices. The use of speech recognition in an office network can be problematic. Overuse injury can result just as easily from mouse activity as from keyboarding. And other factors need to be addressed, such as posture and ergonomic layout. Professional advice should be sought if you suffer from RSI. Speech recognition, on its own, is unlikely to be a solution.

Can I control my mouse by voice?

Both *Dragon* and *ViaVoice* allow control of mouse movements and clicks by voice. *FreeSpeech* only supports mouse clicks. *ViaVoice's Voice Mouse* is probably the most wholehearted effort at offering mouse control by voice. It is virtually a separate mode, activated with the command, "Begin Voice Mouse". Mouse movement and clicking (including click and drag) are supported. The speed of cursor movement can be varied. Accurate control takes time and practice but is possible.

Dragon uses grids to control the mouse. The screen is divided into 9 squares. As a square is chosen, it divides into 9 smaller squares. The process continues until the cursor is in the position you desire. It can be a slow and frustrating process. But again, with practice, it can be made to work.

Can I hear a replay of my dictation?

Each of the programs offers a replay facility. In *Dragon* this facility is only available while the document is open. In *FreeSpeech* the text is highlighted as it is read back - a very helpful feature. The playback feature in *ViaVoice* is limited to

1000 words. Only *ViaVoice* and *FreeSpeech* allow speech data to be saved with the document. This allows the dictation to be replayed later, by you or another person. But keep in mind that documents saved with speech data are very large files – possibly several megabytes!

Can they read my documents aloud?

Each of the programs offers the helpful facility of reading your documents aloud with synthesised speech. *FreeSpeech* has a basic facility with no option for changing the speed or pitch. *Dragon* has similar quality speech output, but with the option of varying the pitch, speed and volume.

But the best quality speech output comes through *ViaVoice.* Although the pitch and speed cannot be adjusted, the quality is excellent. Up to 1000 words can be read back at a time. It has a slight English accent.

Speech Recognition and People with Disabilities

Speech recognition has brought significant benefits for people with disabilities. People with quadriplegia especially have used speech recognition to regain productive involvements in their lives. Usually a more direct method of controlling mouse functions is preferred, however. These can include trackballs, joysticks, touch pads and switches. Alternative microphones may also need to be considered. An assessment to determine the best combination of access technology for a person with a disability can be arranged through an organization such as the Ability Research Centre (02-9907-9770).

OUR TESTS

Choice arranged for the three main speech recognition programs to be compared. Four adults were involved in the tests (two males and two females). Three computers were used for the tests:

- IBM Aptiva (AMD 533, 128 Mb RAM)
- Compaq Presario 1900 Notebook (PIII 500, 192 Mb RAM)
- No brand (Soundblaster Live! PIII 733, 128 Mb RAM)

The testers varied the sequence in which they did the tests. The microphones that were supplied with the packages were used. The minimum training as specified by each program was undertaken. A test passage (quite a difficult one, with a total of 316 words and commands) was then used and scored for errors. We then corrected the errors from the first test, undertook an additional segment of training, and then did a second test.

Results - Accuracy

Dragon Naturally Speaking was the most accurate out-of-the-box, with an average accuracy of 89%. This compared with 87% for *ViaVoice* and 82% for

FreeSpeech. However *ViaVoice* was slightly more accurate after the second tests (following corrections and additional training) with 93% (compared with 92% for *Dragon* and 87% for *FreeSpeech*). We consider *Dragon Naturally Speaking* and *ViaVoice Millenium Pro* to be equivalent in terms of accuracy. Two of our testers did better with *ViaVoice* and two did better with *Dragon*.

The three computers varied slightly in the results obtained. Overall the *Compaq* notebook obtained a slightly better result (90% overall, compared with 89% for the unbranded computer and 88% for the *IBM Aptiva*). This was a surprising result, as conventional wisdom is that notebook computers aren't as good as desktops for speech recognition. Our result doesn't suggest that *all* notebook computers will be as good as the one in this test, but it does show that very good results are possible through the increasingly popular portable platform.

Another point worth noting is the rapid improvement in each of the programs after a very short period of additional training, and after having corrected the errors from the first test.

We obtained very similar results using the built-in DragonPad (*Dragon*) and SpeakPad (*ViaVoice*) compared with *Microsoft Word*.

One important point – our tests do not tell you how the accuracy of these programs will develop with further usage. Each of them improved with our corrections and short period of additional training; one would expect this process to continue.

Results – Ease of Use

Our ease of use scores were based on the following factors:

- Ease of installation and initial training (15%)
- Resources provided (manuals, computer-based instruction, tours, etc) (10%)
- Microphone comfort (15%)
- Built-in help facilities (10%)
- General operation of the program (50%)

ViaVoice was the overall winner in this area, with better help resources, microphone comfort and ease of general use.

Memory Tests

We conducted tests to determine the impact of memory (RAM) in the performance of the SR programs. We reduced the RAM in each computer to 64 Mb, then re-loaded each program in turn and observed changes.

The first thing to note is that some consumer level computers (like the Aptiva in these tests) have a nominal 64 Mb RAM, but part of this RAM is used to support video functions in the computer. This left about 56 Mb of true RAM available for the SR software.

Dragon seemed the most severely affected by the limited RAM. The main problem was that the program lagged increasingly behind the speaker. A short delay is normal in speech recognition, but when the delay extends to several minutes, then it becomes disconcerting and distracting. In DragonPad the delay ranged from 5 seconds (PIII 733) to 1m 25s on the Aptiva and 2m 5s on the Compaq. But in Microsoft Word the corresponding figures were 2m 50s, 6m 45s and 17m! These delays were extended further if another program was opened at the same time.

ViaVoice and *FreeSpeech* seemed less affected by the smaller available RAM. The Aptiva was affected more than the other computers.

But the conclusion is clear – unacceptable delays will result if you are limited to 64 Mb RAM. This is especially the case if you are dictating into Microsoft Word. We would recommend users increase their RAM to at least 128 Mb if they are planning to use speech recognition.

Microphones

We decided to test the microphones that came with the speech recognition programs, along with some other popular microphone options, to see whether there were significant variations in performance. We included a USB headset microphone, a USB desktop microphone, a telephony headset microphone and a cheap but well regarded headset microphone.

We used the *Dragon* audio setup program to evaluate the microphones. Dragon was the only program that gave an actual score for signal-to-noise ratio. All seven microphones were tested straight after each other, at varying sequences for each test. Two testers were involved.

The results were quite surprising. Best performer clearly was the cheapest microphone – the *Emkay VR3310* – closely followed by the *Dragon* microphone. The *ViaVoice* microphone was the third best in our tests, but a couple of points behind the other two. The *Philips* microphone was near last on our tests, vying for that position with the other non-headset microphone – the *Telex* USB Desktop microphone.

The USB microphones were disappointing. USB microphones are supposed to offer cleaner sound, by-passing the computer's sound card. Desktop or hand-held microphones will always involve a compromise, because the user's mouth is not maintained a consistent distance away from the microphone (as it is with headset microphones).

One further comment – we conducted some of the tests on the Compaq notebook with the power cable connected, and other tests with the computer running on battery power. We obtained better results from the notebook when the power cord was not connected, i.e., when it was running on battery power. This is something to keep in mind if you use a notebook computer for speech recognition.

Ambient noise

We conducted some tests to determine whether ambient noise would influence the performance of the systems. Scores on a test passage were recorded without background noise initially. Then three further readings were conducted, with three sets of noises introduced. The first was a loud conversation, about one metre from the computer. The second was a loud telephone ring, about one metre from the computer.

Finally we used a musical keyboard to play car screeching, sirens and horns tooting in a set sequence, about 1.5 metres from the computer. Each sound came from a different angle. The sounds were reproduced consistently for each test.

We were surprised at the resilience of the systems to extraneous noise. They didn't flinch at loud talking and phone ringing. Only *FreeSpeech* was affected, and then only by the rather extreme car noises.

Our results do not mean the SR systems are not vulnerable to any external noises. In our view you should still be careful if you plan to use a system in a noisy environment. Test it out if possible beforehand. But we were pleasantly surprised at the way the microphones generally seemed to ignore extraneous sounds.

Young person

We arranged for a young user (13 yo male, unbroken voice) to undertake the same tests as the adults. He was closely supervised in this task, to ensure consistency of effort.

We cannot generalise from this one tester's experience. But his results were lower than for the adults, with an overall average of 80% as against 87-92% for the adults. His best results were with *Dragon Naturally Speaking.* In one test he achieved a score of 95%, suggesting that good scores are achievable for young people in the right circumstances.

Mac options

There is only one speech recognition option available for the Mac – *ViaVoice.* There was talk of a version of *Dragon Naturally Speaking* for the Macintosh, but this hasn't materialised.

We put the Mac (iMac, 333 Mhz with 160 Mb RAM, System 8.6) with *ViaVoice Millenium Mac* through the same tests we did for the other SR programs. The tester was the one who obtained the best scores previously on *ViaVoice* for Windows. The Mac version achieved similar accuracy scores that this user obtained with the PC version of *ViaVoice* (93.5% v. 93%).

The Mac version of *ViaVoice* only allows you to dictate into the SpeakPad word processor. You cannot dictate directly into Word or another application. An *Enhanced Edition,* which supports direct dictation into Word and AppleWorks, has been released in the US, but only supports US English. A local version has not yet been released.

Conclusions

There is no doubt that speech recognition systems have made significant advances in recent years. They are capable of highly accurate results – as high as 99% in our tough tests – in the right conditions.

Dragon Naturally Speaking Preferred is a competent product, easy to set up and train, with high accuracy. However it lacks several features that are only

available in the much more expensive *Professional* edition (such as adding custom vocabularies, saving speech data with documents, and custom commands). These features are available in *ViaVoice* and *FreeSpeech.*

FreeSpeech did not perform well in our accuracy tests. A main factor in this poor performance was the handheld speech microphone, which obtained low scores in our microphone tests. It otherwise had some very good features.

Overall the best performer was *ViaVoice*. It had high accuracy and a full range of features. It was also easy to use.

Good solutions are now available for speech recognition. But the main variable now is – the user. If you are prepared to correct errors, so that your speech model will improve, then you will most likely benefit from increasing accuracy. If you desire (or need) to use speech recognition, then your initiative will carry you through the early stages of training and lower results, where the faint-hearted will be tempted to give up. If you are prepared to discipline your speech, to give the speech systems a fair chance, then your results are likely to be rewarding.

It all boils down to whether or not you really want to create text this way. If you do then, with some care in the selection of equipment, speech recognition is ready for you.

Postscript

Since undertaking our tests a new version of *ViaVoice* (Release 8) has become available. It has a larger active vocabulary (150,000 words, expandable to 214,000 words). It supports Windows 2000 and Windows Me. Minimum requirements are now Pentium II 300 (Pentium III 600 for Windows Me), and 510 Mb of available hard drive space.

These tests were carried out at the Ability Research Centre in Sydney. Ability is deeply involved in providing computer technology services for people with disabilities. They have been testing speech recognition programs for over a decade.