Graphics system aids development of synthesized speech

To teach a child to spell with a microcomputer, you can't have "Spell C-A-T" appear on the video display screen; it gives the game away. Instead, you give the computer a voice that issues the instruction. It's a logical solution but difficult to implement if you want a voice that sounds more human than C3PO's.

With "A Bee Cs," the alphabet game Commodore Business Machines will have in full production this fall for the Commodore 64 home computer, a child uses a joystick to fly a "bee" from one letter of the alphabet to another, taking his cues from the voice emanating from the system's speech synthesizer.

At Commodore's Speech Technology Division in Dallas, Texas, speech scientist David Covington and his colleagues are using the Model One/40 graphics system from Raster Technologies as a development tool to give the Commodore VIC-20 its voice. The Raster Technologies device produces spectrograms of individual words and sounds that facilitate the development of Commodore's educational games.

Covington explains: "With a spectrogram of the word 'yellow' on the screen, you have the word trapped in three dimensions. The verticle axis represents the frequency of the sounds contained in the word, measured in hertz. The horizontal axis represents time. Color, the third dimension, represents energy. Not only can you see exactly what is wrong with a word, but you can improve it with software manipulations.

"Of course, we could synthesize speech without a graphics system by just automating the whole process and hoping for the best, but we couldn't get the quality we're after."

Computerized speech synthesis can be accomplished in several ways. Two possible techniques for computerized voice output are waveform digitization and linear predictive coding, or LPC, synthesis.

Waveform digitization starts with a human speaking into a microphone for recording on a magnetic disk. The voice waveform is then sampled periodically and each sample is encoded into discrete levels. This point-by-point process results in faithful reproduction but demands enormous computer memory.

LPC is a means of reducing the amount of data required to store speech in a computer. Using specialized hardware for the LPC synthesis of the speech waveform reduces the memory data requirements by a factor that may range from 50 to 100.

The speech waveform is divided into many short time intervals, or frames, and LPC parameters are computed for each frame. The result is a series of sets of numbers that describe the frequency content of the speech over each frame. The quality of the waveform used to excite the synthesizer in each frame is also encoded (pitch, energy, etc.). Once the LPC parameters are computed, they can be interactively changed to improve the quality of the synthesized speech.

The quality of synthetic speech depends largely on the bit rate of the stored data. Waveform coding results in bit rates of 64,000 bps and up. The LPC technique results in about 1000 bits for each second of speech, and Commodore is experimenting to find even lower rates. As the encoding rate is lowered, less computer memory is required, but which portions are encoded and which are left behind becomes more and more critical.

Commodore's speech synthesis method requires six tools: a data collector (generally the microphone), a quiet recording studio, analog-to-digital and digital-to-analog converters, a host computer (in this case a DEC VAX-11/750), highly specialized applications software, and the Raster Technologies graphics system.

Words are recorded individually rather than by phrase or sentence because software combines words and changes inflection to make a declarative statement a question or an exclamation. Letters of the alphabet are also recorded individually. To record numbers, the speaker records all the numerals from zero to nine, all the teens, and all the multiples of ten; as a result, any number likely to be used in a game can be assembled.

Once the game vocabulary is recorded onto a disk, the bit patterns are compressed for storage in the host. From the host, the compressed data are fed to the synthesizer. If the synthesizer produces a word that sounds wrong, Covington or another member of the team produces a spectrogram to locate the problem.

The Raster Model One graphics system used to display spectrograms is connected to the VAX via a DMA interface. An array processor generates the picture display data. Through pseudo-color mapping, Covington can display a spectrogram on the Raster system's video monitor and assign different colors to the various energy levels to be pictured.

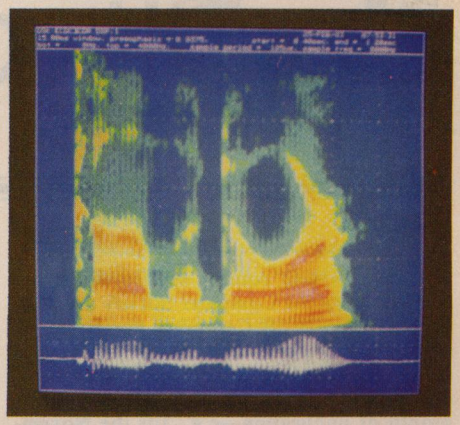
Using Raster Technologies' controller to provide visual representations of speech signals, Covington can change the sounds of a word, trim off unwanted silences, add or remove inflections, and even—if the software algorithms are complex enough—make a Southerner sound like an Englishman.

The Raster Technologies Model One/40 supports image memory configurations from 1024 × 1024 × one bit to 1024 × 1024 × six bits. Because it supports up to six image memory planes, it allows the simultaneous display of 64 colors from a palette of more than 16 million colors.

The system's resolution is so high that Covington can zoom in to examine all the information in one quadrant of the screen without the picture breaking up into obvious pixels. It can also support complex applications programs written for the lower resolution Model One/20 systems.

The Model One/40 has a 16-bit microprocessor configured with 64K bytes of PROM and 128K bytes of user-programmable RAM. In addition, it has an independent vector processor to offload vector computations from the CPU. The vector processor can generate vectors at a rate of more than 700,000 pixels a second.

Covington says, "The technology has a way to go, but we're making progress. After all, man's most natural interface with the computer is not fingertips on a keyboard; it's his own voice."



This spectrogram was produced on a Raster Technologies Model One graphics system. Brighter colors represent greater energy and greater amplitude of the waveform of specific frequencies. An experienced speech scientist can recognize a word from its spectrogram; this is a spectrogram of the word "Commodore."

Use order form on p. 105.