Speech Recognition Moves from Software to Hardware

Linda Dailey Paulson

Speech recognition has long promised a natural way to improve user interaction with computers, cars, and other devices. During the past 30 years, researchers have gradually upgraded the technology to the point that it is used in a number of these settings.

However, because of limitations in processing power and other factors, the applications typically have been relatively simple, and speech recognition has not been widely used, despite the growing desire to implement it in PCs, cell phones, applications that automate home utilities and entertainment devices, and other systems.

This has occurred in part because the technology, which is computationally intensive, traditionally is implemented only in software executed by the host device's main processor, explained Todd Mozer, CEO of speech-chip pioneer Sensory Inc. This makes the process inefficient, slow, expensive, and power-hungry and thus unsuitable for many uses, especially in mobile and other small devices.

Researchers have been working on implementing speech recognition in dedicated processors for about 20 years, but the chips still have limited capabilities and work with only relatively small vocabularies. As such, few companies sell speech chips.

Now, though, scientists are interested in developing high-end speech chips that work with large vocabularies of words and that recognize continuous speech.

"Our central argument is that speech recognition needs to be liberated from a software-only form," said Carnegie Mellon University professor Rob Rutenbar, lead researcher for CMU's In Silico Vox speech-chip project.

Rutenbar predicted that speech recognition will follow in the footsteps of graphics technology and move to silicon. "Nobody does graphics in software anymore," he said. "All graphics are now done in hardware."

Despite its promise, speech-chip technology faces technical and marketplace challenges. For example, speech recognition in general is not yet a commonly accepted user application, noted Carl D. Howe, principal analyst at the Blackfriars Communications consultancy.

DRIVING SPEECH TO HARDWARE

A key challenge to developing effective speech-recognition chips has been the inability of past processors to handle such complex functions quickly, effectively, at a reasonable cost, and without excessive power consumption.

In addition, Rutenbar said, a key stumbling block to faster advances in silicon-based speech technology has been hesitancy by researchers who have worked exclusively with software to consider moving speech to hardware.

Now, though, several factors are driving speech recognition to silicon.

Potential uses

Vendors see a potential demand for using speech-recognition technology in tasks such as interacting with and issuing complex commands to PCs, cell phones, and PDAs; using automation controls for lighting, audiovisual gear, automobiles, and other devices; working with automated services like those in airline-reservation and customer-care systems; and conducting online activities such as searches of the Web and large video or audio files.

In addition, audio search and audio mining could use advanced speech recognition. For example, fast, accurate speech technology would let users search a video for a specific line of dialogue.

The US government has also acknowledged the importance of advanced speech recognition. About a year after the 11 September 2001 terrorist attacks in the US, Rutenbar said, investigators still hadn't been able to listen to 113,000 audio intercepts that could have provided useful intelligence, a problem that faster, more accurate speech technology could have helped solve.

Advances also could expand speech recognition's important role in adaptive technology for amputees, arthritis sufferers, and others unable to use their hands to input information into computers and devices.

Today's technical shortcomings

Today's primary approach, in which a host system's main processor runs speech-recognition software, is not efficient or accurate enough for reliable usage in a variety of systems. This is the case even in high-quality systems with trained users working with good microphones in quiet environments, according to Rutenbar.

Achieving even acceptable levels of performance using the software approach would require extremely fast, powerful, large, expensive, and energy-hungry processors, according to Rutenbar.

A dedicated chip would offer a faster, more efficient approach because it would perform only speech recognition—unlike a CPU, which must handle many tasks.

Moreover, Rutenbar said, an optimized hardware-based approach would lower energy consumption, which would be critical for implementing speech technology in battery-powered mobile devices.

Technical advances

Proponents say ongoing advances in speech recognition and processor functionality have made developing effective speech processors more feasible.

The initial speech processors were limited in speed, vocabulary, the ability to determine meaning based on context, and other ways, noted Sensory's Mozer.

However, he added, general increases in processing power and decreases in per-MIPS costs have begun to overcome these limitations. According to Mozer, cost will be a huge factor in speech recognition's success because consumers don't want to pay a lot more for such added functionality.

INSIDE THE TECHNOLOGY

Various academic researchers began working on the basic technology behind today's speech-recognition products in the early 1970s.

The first software products, released in the 1980s, included Dragon Systems' Dragon Dictate and IBM's ViaVoice. Sensory released the first commercially successful speech-recognition chip in 1995.

Speech recognition 101

Speech recognition works basically the same for all languages. However, vendors modify the software to account for differences such as the pitch inflection that is critical to recognizing speech in languages like Chinese and Thai.

> **Dedicated chips handle the parts of the speech-recognition process in parallel.**

In a typical system, a microphone captures speech as electrical signals, which an analog-to-digital converter then processes for further handling.

The system subsequently divides digitized words into segments, analyzes the component frequencies, and compares them to a database of digitized signals that represent various words and phonemes, which are a language's single distinctive sounds. The computer then determines the probability that captured words and phonemes correspond to valid entries in the database.

Hidden Markov modeling, a probabilistic statistical approach, is the backbone of basic speech-recognition technology. Other technologies such as neural networks—a predictive approach in which interconnecting processing elements in a network work together, often in parallel—provide acoustic, phonetic, and other types of pattern recognition.

There are various speech-recognition approaches, including isolated-word and continuous-speech systems, speaker-dependent systems designed to work with a single person, and speaker-independent systems that work with multiple users.

Moving to silicon

Dedicated chips implement speech recognition the same way as software. However, speech-chip makers are designing for speed in addition to the accuracy that is software developers' principal goal, noted Shlomo Peller, CEO and founder of speech-chip maker Rubidium.

Speech chips typically include a preamplifier to increase the strength of input signals from the microphone, analog-to-digital and digital-to-analog converters, direct-memory-access units for better data handling, and a vector accelerator and hardware multiplier for improved performance.

Hardware handles the various parts of the speech-recognition process in parallel, which speeds up the work.

Speech chips generally are relatively inexpensive—between $2 and $3 each—and slow, running at up to 50 MHz. Their vocabulary capabilities depend on their memory size.

Systems can work with speech chips if they have or add a battery, speaker, microphone, and a bit of additional circuitry.

IN SILICO VOX

CMU's In Silico Vox (voice in silicon) research project is looking at two approaches to someday building massively parallel, energy-efficient speech chips to attain high performance, Rutenbar explained.

One approach uses field-programmable gate arrays (FPGAs) and recognizes 1,000 words. It uses a Xilinx Virtex II-based chip with a clock rate of 50 MHz and has 1.12 Mbits of on-chip static RAM and access to 3 Mbytes of off-chip dynamic RAM, with 200 Mbytes per second of memory bandwidth.

The processor operates about 2.3 times slower than real time—meaning it takes 23 seconds to process 10 seconds of speech—because of memory-bandwidth limitations.

The CMU team is also working with application-specific integrated circuits. The ASIC approach recognizes about 5,000 words. The chip has a 100-MHz clock rate, 2 Mbits of on-chip SRAM, and access to 70 Mbytes of off-chip DRAM, as well

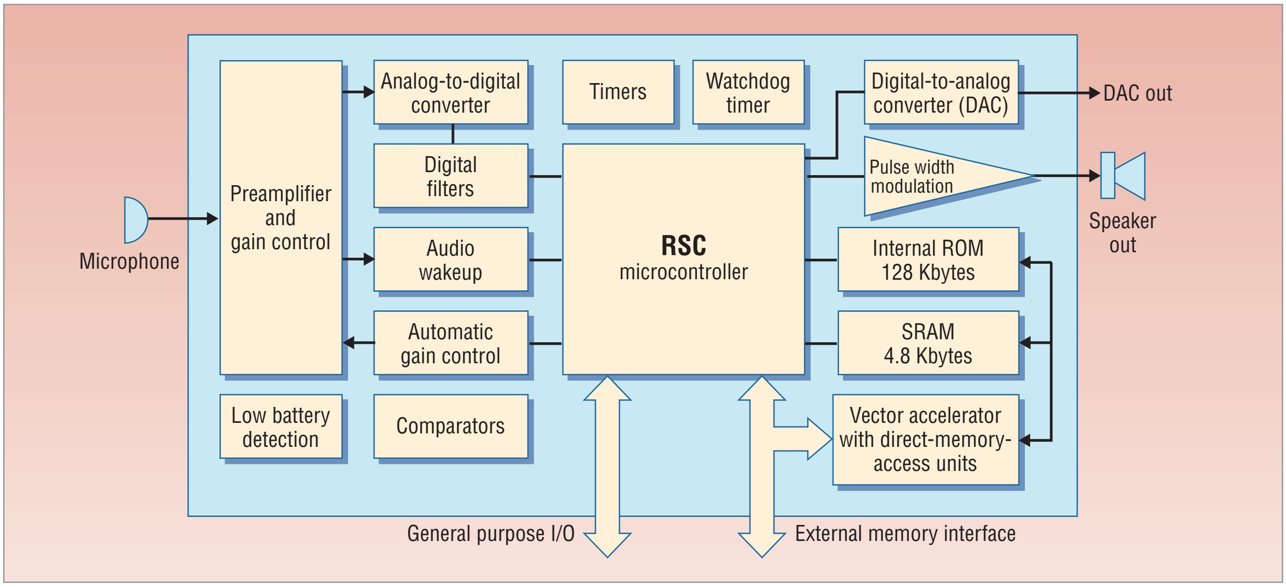

Figure 1. Sensory Inc.'s RSC-4128 speech chip has several key elements in addition to the main speech-processing unit. For example, the chip includes a preamplifier to increase the strength of input signals from the microphone, analog-to-digital and digital-to-analog converters, direct-memory-access units for better data handling, and a vector accelerator for improved performance.

as 900 MBps of memory bandwidth.

The processor operates about six times faster than real time, meaning it takes 10 seconds to handle 60 seconds of speech.

It is also very small—about 10 sq. mm. excluding the external DRAM. By comparison, Intel's Core 2 Duo chip is about 150 sq. mm.

The issue between the two CMU designs is that the FPGAs are slower, larger, and more power-hungry than the ASICs but also are programmable and thus more flexible and useful for prototyping, Rutenbar explained.

He said his team is experimenting with various approaches to reach its ultimate goal of developing a speech chip that operates 1,000 times faster than real time. He said the technology might be ready for commercial use in about five years.

COMMERCIAL PRODUCTS

Numerous large companies—such as AT&T and Oki Semiconductor—worked on commercial speech chips over the years but abandoned the effort upon discovering there was not a big enough market to generate adequate revenue, said Sensory's Mozer.

More companies aren't developing speech chips because they aren't familiar with the technology, according to Rubidium's Peller. They don't want to spend the time and money needed to gain the expertise and develop the products when they fear there is not yet enough demand to generate a reasonable return on investment, he said.

Sensory Inc.

Sensory makes the RSC-4x family of chips, which are mixed-signal processors that provide speech recognition, synthesis, and system control. They can also perform music synthesis, speaker verification for security purposes, and other tasks. Vendors use the chips in products such as toys, light switches, and even controls for a massage chair.

Sensory's most recent speech chip, the RSC-4128, which Figure 1 shows, performs at up to 8 MIPS and contains 4.8 Kbytes of SRAM and 128 Kbytes of ROM.

Sensory can program the chips' vocabulary for use in specialized applications.

Vendors can build a complete system with the chip by adding a battery, speaker, microphone, and a few resistors and capacitors.

Users can also buy Sensory's VR Stamp speech module, which already contains a complete system's major elements, said Mozer.

Rubidium Signal Technologies

Rubidium makes the RDE50 and RDE100 dialogue-engine processors. These devices—used for voice-control in settings such as household appliances, cars, and telephone systems—perform a wide array of related tasks, including compression of speech data for easier network transmission and noise suppression for clearer input.

"The special thing is that they include all the components of a speech system in a single, very low-cost chip," said Peller. Also, he added, the devices recognize words and expressions, rather than phonemes, are speaker independent, and can work with large quantities of data.

The chips typically have 4 Kbytes of RAM and 64 Kbytes of ROM and perform at up to 16 MIPS, according to Peller.

Rubidium can program the processors' vocabulary for use in specialized applications.

SPEAKING OF HURDLES

Building speech processors for small, price-sensitive mobile devices such as cell phones is a challenge because greatly reducing the chips' size, power consumption, and cost can also reduce performance and accuracy, Rutenbar said.

Speech-recognition chips have very narrow markets now, such as automotive voice control, toys, and other novelty items, said Bill Meisel, president of TMA Associates, a speech-technology market analysis firm. This limits the immediate opportunity to make money from the technology.

Dedicated processors wouldn't be as flexible as the software approach, Meisel added, because it's easier to change algorithms and then implement them in a general-purpose chip than to make changes to a hard-wired speech chip.

The fundamental problem facing speech chips is the lack of large-scale demand for speech technology in general, stated Howe. Consumers haven't expressed much interest in using the technology, even where it's already available, such as in cars. There's just no killer app for the technology, he said.

Rubidium's Peller predicted that speech chips will run faster and cost less within five years. Future chips will thereby support more sophisticated and natural man-machine dialogue at an affordable cost and thus will appear in more consumer appliances and applications, he explained.

Consumer expectations will be one key to the technology's future. "People want *Star Trek*," explained Mozer. "To become ubiquitous, [speech technology] will have to work like in *Star Trek*."

The technology will indeed reach that point, according to CMU's Rutenbar, but it will take time and research. After all, he noted, it took 20 years for speech recognition to achieve 90 percent accuracy.

These advances might require faster speech chips, said Howe. In the process, the chips' cost and energy consumption must not increase.

"Maybe it's going to take a huge investment in computational power to be where it needs to be," Howe concluded, "but the advances are not going to happen in software." ■

__Linda Dailey Paulson__ is a technology writer based in Ventura, California. Contact her at ldpaulson@yahoo.com.

Editor: Lee Garber, *Computer*, l.garber@computer.org