Computerworld US 21.09.1992

Guten tag! Bonjour! Good day!

Carnegie Mellon to demo Janus speech recognition and translation system in January

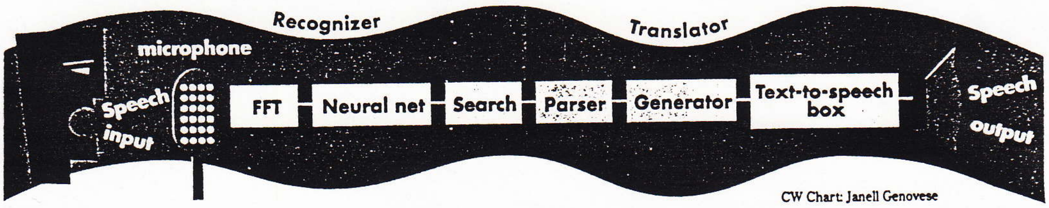
8846

BY MELINDA-CAROL BALLOU
CW STAFF

Imagine this: At your Boston office, you speak English into your workstation, and your Tokyo-based colleague hears the Japanese translation. She responds in Japanese, and you hear her answer in English.

As world economies become more integrated, the need for multilingual technologies is more acute than ever before.

Research into speech recognition and translation has been going on for more than 30 years. But the requirements for speech recognition have been so computationally intensive that even simple tasks, such as the recognition of single words, were laborious and time-consuming when computing power was less available.



CW Chart: Janell Genovese

The translation of text from one language to another has progressed more rapidly than speech recognition because the text translation is less consumptive of CPU power. The combination of both technologies is now beginning to yield results.

The birth of Janus

In January, members of Carnegie Mellon University's Center for Translation, Siemens Corp., the University of Karlsruhe in Karlsruhe, Germany, and Japan's Advanced Technology Research consortium are scheduled to conduct an intercontinental demonstration of Janus, a speech recognition and translation technology.

Janus will allow members of the group to speak with one another in English, Japanese and German but under more constrained conditions than the scenario described above.

Janus, which was designed as a conference registration system, has a vocabulary of only 400 words. The translations themselves will occur in almost real time, taking about one and a half times the actual speaking time in the best cases.

The current version of the system runs primarily on Hewlett-Packard Co.'s HP 9000 Series 720 Unix workstations, in conjunction with a massively parallel system from Maspar Computer Corp. in Sunnyvale, Calif., using Digital Equipment Corp.'s DECtalk as a speech output device.

Janus addresses several complex problems: receiving and translating speech into digitized text, translating that text into the speech of the target language and outputting it via a speech synthesizer.

"Translating broad-based speech into accurate, digitized text is a very hard problem, and translating text into a target language is also very difficult. Combining them compounds two very difficult problems," said Bill Meisel, president of TMA Associates, a market research firm specializing in speech technologies based in Encino, Calif.

One problem is that computerized recognition of spoken words can often introduce errors.

The largest market for speech recognition and speech synthesis systems will be for systems capable of translating spontaneous speech. But the complexity of addressing those translation issues is far beyond the capabilities of current technologies.

In order to make them nearly interactive, "almost all speech recognition and translation systems must limit themselves now to a specific set of words and tasks," McNair said.

"It is a hugely difficult technical problem to handle an unconstrained conversation like the one you and I are now having," Meisel agreed.

"There are challenges all along the way," said Arthur McNair, a research programmer at the Neural Network Speech group at Carnegie Mellon. "Someone might say 'Hello, how are you?' and our recognizer might come back with 'Hello, how is it?' "

To overcome ~~these~~ difficulties, Janus offers a language model built specifically for the task of conference registration. Hypotheses created by the recognizer are parsed and matched against that model. The text is then analyzed to find the most appropriate choice.

Neural exploration

Those involved with the Janus project have been exploring the use of neural networks to help discriminate among sound patterns.

"Our language model is written in such a way that it is expecting the sentences that we have and not very much more at the moment," McNair said.

"One of our current tasks is to try and extend our system to spontaneous speech. We want people to be able to express things differently, rather than being limited to the actual sentences that we have in the system. We're trying to write a more complicated grammar for the parser so that it will accept many more types of grammatical structures and sentences," he explained.

The researchers are also attempting to increase the speed of the recognition process and to support more people. They would also like to increase the vocabulary size beyond the 400 words and thus increase the range and scope of the content it addresses.

To allow for common speech, a system must have a vocabulary of at least 1,000 words; a more general discussion requires at least 10,000, according to McNair.

Janus mechanics

When a user speaks into Carnegie Mellon's Janus system, the workstation receives the input as wavelengths that are digitized by an analog-to-digital converter.

Then Janus uses Fast Fourier Transforms (FFT), an algorithm that was designed to take the waveform input to create a frame for each fragment of speech and convert this data into a spectagram. The FFTs also compact the data so that it is easier to feed through the system. These frames are then fed into Janus' speech recognizer. The recognizer arrives at a hypothesis of what was said, which is then sent to the translator. The translator parses the sentence to determine if it is a valid utterance according to the Janus language model.

The use of neural networks in this parsing process increases the accuracy of the Janus system as well as its efficiency because the system is able to learn parsing via examples, instead of requiring complicated grammatical structures for all possible linguistic alternatives.

If the sentence is valid, it is sent to the output language generator, which produces the translation, and the speech is output.