Summary of Session on Hardware for Spoken Language Demonstrations

Richard Schwartz

SUMMARY

Three talks were given during the session, followed by some general discussion of the needs of different research groups for demonstration hardware. The talks were given by:

Hy Murveit, SRI International, discussing a high-level design of a speech recognition system using special purpose accelerators;

Roberto Bisiani, CMU, describing a CPU board that supports simple parallel implementations;

Elizabeth Mailhot, AT&T Bell Labs, describing the DSP-3 parallel processor speech workstation.

Hy Murveit described a system in which the different operations necessary for spoken language understanding were performed by different hardware. The different hardware would initially be on different boards, connected together in a VME card cage. In particular, the signal acquisition, signal processing, and vector quantization are performed on a board with two TMS320C25 chips and sufficient static RAM to support them. The HMM processing is performed on a special purpose VLSI board that is capable of updating one HMM state in every cycle. The initial design runs at 5 MHz and can therefore process up to 3,000 words in real time with no pruning. The next version will run at 10 MHz, and allow pruning, thus greatly increasing its capacity. The design is general enough to allow for most of the features of all the discrete HMM systems. It will not allow for HMMs that use continuous densities or shared mixtures. It also may not allow for phonetic models whose context extends beyond the word. The next project will be aimed at replacing the statistical grammar processor with a natural language filter driven in a time-synchronous manner.

Roberto Bisiani presented a processor board that supports parallel processing with other similar boards. The initial board goes on a MacIntosh, and there may be a version that goes on a VME bus soon. The CPU chip being used currently is the Motorola 88000 which is about twice as fast as a SUN4. The board also contains 8 MB of DRAM for current memory chips, which will increase to 32 MB for the next generation of 4 Mb chips. Bisiani pointed out that processors are already too fast for their memories, so faster processors won't help. Instead, he recommends parallel processing. Therefore, this board has special purpose hardware for allowing several boards to access shared memory that might exist on a different board. Bisiani has an optimized version of the HMM decoder that can perform the search almost in real time on one board for the basic algorithm.

Elizabeth Mailhot from AT&T Bell Labs presented the latest version of the ASPEN multiprocessor, which is called DSP3. The system is based on the DSP32C processor. Each board will contain 16 DSP32C nodes and each processor can have 64 KB of static RAM for current chips and 256 KB for next generation memory. The DSP32C has a theoretical peak of 25 MFlops. Two significant changes that have happened since the previous version (BT100) is that the different processors can be (software) configured in many different arrangements, and is not limited to a binary tree. In addition, the communication between the processors is at a much higher rate of 40 MB/s instead of 1 MB/s. A video tape demo of real-time recognition of the FCCBMP domain on the BT100 was shown. Finally, a speech workstation is being designed around the DSP3. It contains a SUN, A/D-D/A hardware, a WORM drive, and the DSP3.

After the presentations, there was some discussion of what hardware they would want to have for

their demonstrations. A few sites (BBN, CMU, and SRI) said they would want a copy of the VLSI discrete HMM board, even though it was not completely general. Several more sites said they would like to look at the 88000 board, since it was more general and might eventually support research as well as demonstrations.