SESSION 12: CONTINUOUS SPEECH RECOGNITION AND EVALUATION II*

Clifford J. Weinstein, Chair

Lincoln Laboratory, M.I.T.
Lexington, MA 02173-9108

This session featured a summary of the first dry run benchmark tests on the new Wall Street Journal (WSJ) continuous speech recognition (CSR) pilot corpus, and a description of the techniques used and lessons learned by the four sites who conducted the large vocabulary CSR tests.

For the first presentation, Dave Pallett distributed a handout with system descriptions and results. He credited the people involved, and indicated the tight schedule which was met. The tests included three training paradigms: speaker-dependent (SD); longitudinal speaker-dependent (LSD), with much more training speech; and speaker-independent (SI). Tests included 5K and 20K vocabularies, bigram and trigram language models, and recognition on speech collected with verbalized punctuation (VP) and with no verbalized punctuation (NVP). Data was shown indicating special difficulties with a few of the speakers. Results were presented on signal-to-noise ratios for both the primary and secondary microphones. The data are summarized in Pallett's Proceedings paper. Some comments on the results are given below.

The next four papers, on recognition of the WSJ data at Dragon, CMU, Lincoln Laboratory, and SRI, included the common theme that extending a CSR system to a much larger vocabulary and more general task domain required more than a new dictionary and language model. In particular, major increases in search time, computation for matching, and memory utilization required each site to make compromises or revise strategies in acoustic modelling, search, and matching strategies. Despite the preliminary nature of the work on this corpus, encouraging results were obtained and important issues were raised.

The Dragon paper was presented by Francesco Scattone, and described two recognition approaches that were developed and tested. The first method utilized unimodal phonetic elements (PELs), and the second a variation of tied mixtures very recently implemented at Dragon, which was used in Dragon's dry run evaluation test on the 5,000 word SD portion of the corpus. The tied-mixture models proved very effective in modelling the multi-modality of parameter distributions, and generally yielded better recognition results. Scattone indicated that future work will focus on further development of the tied-mixture techniques, including efforts to develop high-performance speaker-independent recognition techniques.

Next, Fil Alleva discussed the application of CMU's SPHINX-II system to the WSJ CSR task. An important change to SPHINX-II which was made to reduce running time was to use only left-context-dependent cross word models; in addition, a number of changes were made to the Viterbi search to reduce running time. Tests were run on a variety of conditions, including the spontaneous speech, and results are summarized in the paper.

The next paper, by Doug Paul, described substantial changes made to the Lincoln Tied-Mixture HMM CSR, to achieve effective operation for the large-vocabulary CSR task. The recognizer, which had previously used a time-synchronous beam-pruned search, was converted to a stack-decoder-based search strategy with an acoustic fast match. Cross-word models had not yet been included. The stack decoder strategy was shown to perform effectively for the larger vocabularies, and a variety of development test and evaluation test results were presented. In addition, a rapid speaker enrollment procedure was described, and positive (but preliminary) results on rapid adaptation (using the standard WSJ 40 adaptation sentences) were presented. A discussion followed, focusing on the language modelling, and on perplexity for closed and open vocabularies.

Hy Murveit described the application of SRI's DECIPHER system to the WSJ CSR task. He focused primarily on performance, since the CSR system used was essentially identical to the system used in ATIS. He acknowledged help from Dragon (Lexicon) and Lincoln (Language Models) in porting to the WSJ task. He described how DECIPHER was stripped down to reduce computation for the task. Tests on the secondary microphone were described, with about 40% increase in error rate. An experiment was described to investigate the effects of additional

*This work was sponsored by the Defense Advanced Research Projects Agency. The views expressed are those of the author and do not reflect the official policy or position of the U.S. Government.

SI training data. The experiment indicated that substantial increases in SI training data could produce significant reductions in error rate relative to those reported in the dry run evaluation tests.

The chairman initiated the discussion period which followed this final presentation by presenting a plot of error rate vs perplexity for the WSJ dry run tests, the previous best resource management (RM) results, and CSR dictation results which had been presented by IBM at ICASSP-89 (Bahl, et.al., Large Vocabulary Natural Language Continuous Speech Recognition). For perplexity-80, the WSJ error rates ranged from 9.0% (LSD) to 12.9% (best SD) to 16.6% (best SI). These error rates were considerably higher than the most recent perplexity 60 RM results (1.8% for SD) and (3.6% for SI), but not as much higher than the perplexity-90 SD IBM results (an 11% error rate was reported in the ICASSP-89 Proceedings paper, and an improved error rate of about 5% was presented at the ICASSP-89 talk). With the understanding that results obtained in these different tests are not directly comparable, still a fair conclusion which could be drawn is that the WSJ corpus is a sufficiently-challenging one (especially when 20K vocabularies, spontaneous speech, and secondary microphones are considered), and that the results of the first dry run test were quite encouraging.

Most of the ensuing discussion focused on the WSJ corpus and evaluation issues which George Doddington had listed in his earlier CCCC talk. These are summarized below by topic.

In summarizing the discussion, an attempt is made to sample the range of comments and issues raised.

MULTIPLE EVALUATION CONDITIONS

The issue was raised of whether the large number of evaluation conditions was a good idea.

Doug Paul: The multiple conditions added richness and were appropriate for a pilot and for covering the varied goals of different sites.

Francis Kubala: The sampling of conditions had worked out well; the number of conditions should probably be narrowed a bit over the next few months.

Victor Zue: There is a concern over too much splitting of the corpus into multiple parts to support the different tests.

Patti Price: Expressed concern about too many baselines.

Rich Stern: Suggested settling on a few common conditions.

VERBALIZED PUNCTUATIONS VS NON-VERBALIZED PUNCTUATION

First, a sampling of the comments in favor of continuing to collect data with a split between VP and NVP.

Janet Baker: People using real dictation systems use VP, so any recognition system for dictation must handle VP.

Doug Paul: Both NVP and VP are needed to support both general recognition and dictation; reading with VP may be awkward at first, but not hard to get used to.

Michael Picheny: Might as well use VP since it is easier for the recognizer, and people who dictate do not seem to mind. Emphasized his strong support for VP.

Second, a sampling of comments generally against a lot more collection of VP data.

Rich Schwartz: Given recording problems with VP, would be happier with NVP for general recognition.

Dave Pallett: Doesn't like the split; would like to reduce handling costs.

Rich Stern: Doesn't see value in perpetuating VP.

Victor Zue: The VP speech, based on his listening experience, is highly distorted; also people hate to read it.

General follow-up comments.

George Doddington: SRI (Jaret Bernstein) is going to ask people to dictate naturally; let's see what they do.

Patti Price: All test data should be spontaneous speech.

John Makhoul: Would the recognition techniques we develop depend on whether we collect VP or NVP speech? If not, who cares.

PROMPTING TEXTS: PREPROCESSED VS NATURAL

Francis Kubala: Questioned the idea that the acoustic training data must match the language model.

Doug Paul: Acoustic modelling is a priority of this CSR effort, so it's very important scientifically to have a correct language model, as shown in paper by Paul, Baker, and Baker at the 1990 Speech and Natural Language Workshop. Prompting texts are a pragmatic way to do this.

Jordan Cohen: Prompting should be an empirical issue — do real dictation experiment and see what people do.

Bob Moore: Preprocessing is a small effect in the 20K language model, so it should be possible to generate language models from text without constraining the prompts.

Victor Zue: Cited the MIT study which showed the variability of responses from unpreprocessed prompts; also raised the issue (not discussed further) of selection of limited vocabulary.

SPONTANEOUS VS READ SPEECH

Many agree that real interactive data and testing is needed, but expensive. Many also agree that it is important to continue collecting read speech.

Roger Moore: Relates study showing that 2% error drives dictation users to isolated words.

Mike Picheny: Concurs — accuracy is more important for CSR.

George Doddington: Suggests simulating error-free dictation system.

John Makhoul: Let's concentrate on read speech now, while keeping alive the effort on exploring paradigms for spontaneous speech collection.

Janet Baker: Emphasizes that interactive simulations are expensive, and that collection of read speech is valuable and cost-effective.

Victor Zue: Agrees with John Makhoul.

FINAL REMARKS

At this point, Charles Wayne noted that two great landmarks had been achieved: the collection of the pilot corpus and the dry run evaluation, and that both were major accomplishments for the Spoken Language Program.