Session 7: Speech Recognition I

Mitch Weintraub, Chair
SRI International
333 Ravenswood Ave
EK 180
Menlo Park, CA 94025

This session presented a number of interesting papers on a wide range of topics concerning speech recognition: two papers on noise robust signal processing algorithms, one paper on approaches to large vocabulary continuous speech recognition, two papers on algorithms for reducing computation time, one paper on adding new words to the vocabulary, and one paper on theoretical issues concerning vocabulary independent recognition.

The first paper, presented by Harvey Silverman, presented an algorithm for talker location using a linear microphone array. He advocated an algorithm called Stochastic Region Contraction for nonlinear optimization problems, and claimed it was very successful in microphone array work. Results were presented for determining the location of a single talker using a cross-correlation hyperbolic-fit algorithm. In the discussion following the paper, the issue of how to evaluate microphone arrays was raised, and it was suggested that improvement to speech recognition accuracy be used as the e valuation metric.

In the second paper, presented by Alehandro Acero, a new algorithm for joint noise suppression and spectral-tilt compensation was presented. The algorithm uses the EM algorithm for finding the ML solution using incomplete data (channel and noise characteristics). Recognition rates were presented for a number of simultaneous recordings using different microphone pairs. During the discussion, he explained that although some microphone characteristics are sensitive to the relative location of the talker, this is dealt with by reestimating the transfer function independently for each utterance.

In the third paper, presented by Laurence Gillick of Dragon, described their Phoneme-in-Context (PIC) modeling approach for large-vocabulary continuous speech recognition. He described how each PIC is made up from 1 to 6 "phonemic segments," with a maximum of 2000 phonemic segments used to construct all 30,000 PIC's. He also said that stress and duration were an important part of each PIC. The discussion focused on how many PIC's were necessary for modeling general English, and how the PIC differs from the commonly used triphone: the main difference being that Dragon uses Bayesian smoothing.

The fourth paper was also from Dragon and focused on their rapid match algorithm. Laurence Gillick explained that the signal processing consists of 3 smooth frames spaced over a 240 millisecond window, and was used to find collections of words whose beginnings are acoustically similar. Each of the features are assumed independent Laplacian distributions, and a special clustering algorithm is used for clustering the probability distributions to reduce computation time. There was some discussion of how this algorithm compared to IBM's fast match, but not enough details were known about either system to make a detailed comparison possible.

The fifth paper, presented by Vasilios Digalakis, focused on fast search algorithms for use in the stochastic segment model (SSM). He presented a new algorithm for joint segmentation and recognition usin g a variant of the split and merge algorithm. They reported phone recognition results on the TIMIT corpus using context independent SSM techniques that are similar to other reported context-dependent results. They were able to significantly reduce the computation of the SSM; however the resulting split and merge algorithm required over 100 iterations in processing each sentence. The discussion focused on their plans to extend this work to word and sentence recognition tasks, and whether there was an agre ed upon convention for training/testing sets and scoring algorithms using the TIMIT corpus (apparently not). It was suggested that NIST define a standard for training and testing using this corpus.

The sixth paper, presented by Michael Picheny, described IBM's approach to adding new words to the dictionary. The problem was formulated in a probabilistic framework, where the goal is to find the baseform string to maximize the probability given the language model and acoustic observations of that word. He summarized their approach to using decision trees for generating spelling-to-sound rules, and the difficulty of generating consistent baseforms. The discussion focused on the need for this approach (versus a large dictionary), and how users of the Tangora system pronounce words strangely, especially proper names. He also stressed the need for baseforms so that these words could be incorporated into the fast match algorithm.

The seventh and last paper presented in this session was presented by Doug Paul. He described how the training language models affects the acoustic models used by speech recognizers, and how this bias is illustrated by the use of corrective training with one language model and

testing with a different language model. He pointed out that current resource management (RM) corpus is a very poor source of acoustic information for porting to new vocabularies, as illustrated by the CMU experiments on vocabulary independence. He then presented a proposal for a new corpus of read speech based on the ACL/DCI corpus. Part of the discussion focused on the ability to develop language models from small task-related data sets, rather than from large non-task related texts. The other part of the discussion focused on understanding what the perplexity of RM task is.