OVERVIEW: CONTINUOUS SPEECH RECOGNITION I
chairperson - Janet M. Baker

The Continuous Speech Recognition I session consisted of 8 tightly woven presentations, rich in technical content and experimental results. BBN led off with three talks, followed by two each from CMU and AT&T, and a final presentation from LL.

The first paper, BBN1, presented by R.Schwartz, refreshingly started by recounting a series of experiments the first of which had not or only minimally, improved system performance. Algorithmic methods discussed included Linear Discriminant Analysis, Supervised Vector Quantization, Shared Mixture VQ, Deleted Estimation of Context Weights, MMI Estimation Using "N-Best" Alternatives, and Cross-Word Triphone Models. The last of these proved most effective in reducing word errors. Although not all of these methods have yet been combined into one system, the error rate on the May 1988 Resource Management test set (using word-pair grammar) has been halved.

In the BBN2 paper, F. Kubala presented a method for speaker adaptation using multiple reference speakers. A more traditional approach typically pools multiple speakers into a single set of broad patterns. This approach differs in that it first normalizes speech from multiple speakers, separately performing spectral transformations to a common reference space, and then pooling these, as if they were from a single speaker. Preliminary results pooling normalized speech from 12 speakers appears quite promising in contrast to single reference normalization results. Additional control experiments and the use of many more reference speakers are anticipated.

R. Schwartz in BBN3, recounted initial experiments aimed at detecting that a speaker has used a word not in the known vocabulary. By comparing each of the words spoken with a general acoustics model for all words, in addition to "in-vocabulary" word models, one can apply thresholding on word match scores to help discriminate new words from in-vocabulary lexical items. Depending on the level of thresholding applied, the proportion of new words detected relative to a false alarm rate may be altered. Encouraging test results were obtained using Resource Management speech data where "new" words were created simply by removing a subset of in-vocabulary words from the standard system lexicon.

KF. Lee and F. Alleva jointly presented CMU1. Lee discussed CMU's present progress, including the use of semi-continuous hidden Markov models (SCHMMs) applied to the 1000-word speaker-independent Resource Management continuous speech recognition task. The SCHMM used here is derived from multiple VQ codebooks, whereby the probability density function for each codebook is determined by combining the corresponding discrete output probabilities of the HMM and the continuous Gaussian density functions for that codebook. Test results indicate superior performance with this SCHMM methodology in contrast to both a discrete HMM approach and the continuous mixture HMM.

Alleva's CMU1 presentation centered on automating new word acquisition by mapping acoustic observations in continuous speech, to appropriate standard English orthography. A 5-gram spelling/language model using 27 tokens (A through Z plus "blank"), was constructed from extensive (15,000 sentences) training data. Despite a low spelling perplexity in a test set, difficulties in accurately detecting word boundaries were believed a significant factor in observed high error rates. Future experiments will concentrate on using intermediate mappings from acoustics to phonetic units, and possibly syllables, prior to generating the corresponding orthography.

The CMU2 paper presented by HW. Hon, addressed research in constructing "vocabulary-independent" acoustic word models in an effort to avoid task-specific vocabulary training, thereby enabling the rapid configuration of new speaker-independent recognition tasks, incorporating new lexical items. This approach requires the extraction of flexible sub-word units from a large training database. The recognition results using generalized triphones are highly dependent on the size of the training set, from which they are derived. Errors decrease substantially (though showing no asymptote...), as the training set size increases from 5000 to 15,000 sentences.

Delivered by CH. Lee, the AT&T1 paper reviews acoustic modeling methodologies employed in conjunction with a large vocabulary speech recognition system being developed at AT&T Bell Laboratories. Based on the actual words in a given training set, acoustic descriptions are defined in terms of phone-like units, "PLUs". Many tests on the Resource Management task were performed with both context-independent (CI) PLUs (set size = 47) and context-dependent (CD) PLUs (set sizes range from 638 to 2340). The highest performance results were obtained using CD PLUs. Detailed error analyses were presented as well as recommendations for further work to include more detailed function word/phrase modeling, interword CD PLUs, corrective training, and multiple lexical entry acoustic descriptions where required.

In AT&T2, S. Levinson discussed a separate speech recognition system at AT&T Bell Labs. This approach is based on matching a phonetic transcription derived from continuous speech input, against the closest string of phonetic spellings for the constituent lexical items of grammatically allowable sequences. Although test results on the Resource Management task have been disappointing thus far, the author is encouraged by the quality of his speech synthesis of the phonetic transcriptions, effectively a 120 BPS coder. Audio tapes were played for the audience and are available from the author upon request.

The concluding paper of this session, LL1 by D. Paul, addresses the issue of employing "tied mixtures" for compactness in implementing a continuous observation HMM system running on very large vocabulary tasks. Resource Management test results indicated a modest improvement for speaker-dependent recognition without cross-word triphones models. Performance gains were not realized however for speaker-dependent recognition with cross-word triphones, or for the speaker-independent system. It was proposed that further work on smoothing of the weights for the tied mixtures may prove productive. Using these tied mixtures results in decreased CPU usage during recognition (1/2 x), but at a cost of increased training (2x). In commenting on the general issue of the Resource Management comparative evaluations, the author observed that inter-site test results show both high across-speaker standard deviations as well as poor correlation of best-speaker lists. Analysis indicates the need for much larger test speaker sets, as well as tests properly accounting for speaker variability.

The chair of this session strongly applauds the openness of the authors, and commends their candor in communicating their results, both negative and positive. Readers and audience, alike, are cautioned however to remember that negative results should not be construed as failures of the intended approach. Potentially positive results from constructive ideas and methodologies can easily be curtailed or negated due to limitations in the data provided (as realized with training and/or test set inadequacies), as well as the myriad of opportunities for Murphy's Law to intervene; e.g. program "bugs", etc.