SESSION 9: SPEECH III

Francis Kubala

BBN Systems and Technologies
10 Moulton Street
Cambridge MA 02138

This session consisted of five papers whose contents spanned a broad range of topics in speech recognition. They dealt with problems in the basic areas of acoustic modeling, statistical language modeling, and recognition search techniques, as well as adaptation of both the acoustic and language models to new data. All papers included experimental test results on well-known data sets and conditions where possible.

The first paper, presented by Jean-Luc Gauvain, formulated the training of mixture multivariate Gaussian HMM densities as a Bayesian learning problem. This formalism provides a unified framework for several basic problems in speech recognition—initial training of the HMM parameters, incremental retraining (adaptation), and parameter smoothing. Experimentally, this approach has reduced the SI recognition word error rate by about 10%, compared to AT&T's usual segmental K-means training algorithm, on a large test set of 34 speakers. Since these both were essentially Viterbi training procedures (estimated from only the single best state sequence), it would be interesting to compare the Bayesian formulation to the commonly used Baum-Welch ML training algorithm. In a speaker adaptation experiment, using 2 minutes of supervised adaptation data, a 32% reduction in error rate was reported on four test speakers. It should be noted, however, that nearly all of that gain was achieved by the two female speakers. It is not clear that this improvement would remain if (two) gender-dependent SI models were used as the baseline.

In the second paper, from CMU, Xuedong Huang presented three diverse techniques for supervised speaker adaptation—codebook adaptation, model interpolation and speaker normalization. The codebook adaptation procedure, which exploited the semi-continuous (tied-mixture) structure of the HMM observation densities in the CMU system, lead to a 15% error reduction. The second technique interpolated the baseline SI model with a speaker-specific one. To make the procedure more robust to sparse training, the HMM densities were clustered to a total of 500. Together, these procedures reduced the error by about 25% using 40 adaptation utterances from four test speakers. Interestingly, performance continued to improve as more adaptation data was used, and with 300 utterances it exceeded speaker-dependent performance with 600 utterances. In the normalization experiment, a multi-layer perceptron (MLP) was proposed to estimate a spectral mapping between two speakers. The procedure was evaluated by comparing cross-speaker recognition (train on one speaker, test on another) to cross-speaker with normalization. It appears that gender difference was the dominant effect in the control experiment, however, affecting two of the three test speakers.

The third paper was presented by Doug Paul from MIT/Lincoln. He reported on his experiences with backoff N-gram language models and a stack decoder. Backoff N-gram models have been used as a standard 'control' grammar in the recent ATIS evaluation, largely due to Paul's effort. In a summary study of bigram grammars at several sites, he found that, for the same test set perplexity, class-based N-gram models outperformed word-based ones. During the discussion, Fred Jelinek announced that the interpolated N-gram is now favored at IBM over the backoff model when the training is sparse. At the last DARPA workshop, Paul proposed an implementation of a stack decoder as a standard interface between speech and natural language. At that time, the decoder had only been tested under synthetic conditions. In this paper, he reports that the algorithm often fails when stochastic language models and real speech data are used.

Michael Riley, from AT&T, presented the next paper on the problem of finding the optimal word sequence, given a sequence (or more generally, a lattice) of phoneme labels and durations. Decision trees were used to estimate the label and duration likelihoods directly from automatically labeled training data. On a standard DARPA test set, with the word-pair grammar, this approach yielded 17% word error, even though the phonetic recognition rate was near 80%. Moreover, there was no gain for the duration modeling. It should be noted also, that the bottom-up lexical access problem, as posed here, is usually avoided by most systems employing HMMs, by constraining the acoustic search from the outset to the phoneme sequences found in a pre-defined lexicon.

The last paper was given by Salim Roukos from IBM on a dynamic (adaptive) language model. Here, the static parameters of a trigram language model were updated from a cache of N-grams computed from a fixed number of the most recently observed words. The IBM TANGORA isolated-word recognizer, with a 20K word office correspondence vocabulary, was used as a testbed. Five test speakers dictated 5000 words from 14 documents. The recognition word error was reduced by about 10% averaged over the test documents which varied from 100 to 800 words in length. It was observed that there was a very small improvement for using a trigram cache over a unigram cache even though perplexity predicted a larger difference. The interested reader should review a similar cache-based approach to adapting the language model, by De Mori and Kuhn, that was presented in session 7.