SESSION 8: SPEECH II

Kai-Fu Lee

Apple Computer Inc.
20525 Mariani Avenue, MS 71AB
Cupertino, CA 95014

This session contains four papers that describe new techniques and recent advances in acoustic modeling. This is an extremely important area of research. Throughout the past twenty years, as computers became more powerful and speech data more abundant, new directions in acoustic modeling further advanced the state of the art. The first two papers describe novel techniques that may lead to paradigm shifts. The second two papers propose extensions to current techniques, in order to deal with the variability found in very large vocabularies.

The first paper, presented by Steve Austin of BBN, described a method for integrating neural networks (NNs) and hidden Markov models (HMMs). The motivation of this work was to overcome two problems of HMMs: their conditional independence assumptions and the difficulty in integrating segmental features. These problems are more easily addressed using neural networks that examine one segment rather than one frame at a time. However, a full search strategy using segmental models is currently prohibitive, as discovered by BBN and BU from their work on stochastic segment models. Austin proposed to combine HMMs and NNs by using HMMs to propose the N-best sentence hypotheses. These hypotheses were rescored using both HMMs and NNs. Finally, the HMM and NN scores were linearly combined to determine the final top choice. The HMMs were the standard BBN context-dependent models. A single NN was constructed to discriminate all context-independent phones, using a fixed-length segment resampled from the actual segment proposed by the HMM N-best algorithm. The linear combination weights were trained from a set of tuning sentences not used in the training. Although the resulting NNs performed substantially worse than the HMMs, the combined result was slightly better than the HMMs. One of the major contributions of this paper is a new paradigm in integrating heterogeneous knowledge sources (the same strategy was used by Ostendorf, et al. in a paper in Session 2).

Mari Ostendorf from BU presented the second paper, "A Dynamical System Approach to Continuous Speech Recognition." The motivation of this work is very similar to the first paper—improved time correlation modeling. The proposed approach makes the assumptions that speech is modeled as a Gaussian process at the frame-rate level, and that the underlying trajectory in phase space is not invariant under time-warping transformations. The speech model used is then based on a stochastic linear system model which incorporates a modeling/observation noise term. This system was evaluated on the TIMIT database, and slight improvements over previous techniques were reported. Because the number of system parameters was constrained by the correlation invariance assumption, it appeared that this approach has greater potential with increased speech coefficients. In response to a question, Ostendorf pointed out that the training and test sets included different speakers and sentences.

The third paper, presented by Hsiao-Wuen Hon of CMU, described recent improvements in the vocabulary-independent work at CMU. The goal of this work is to develop acoustic models that work well on any task, without task-specific training. This requires rich acoustic models that generalize to new words. Previously, Hon has reported about a 30% increase in errors for vocabulary-independent recognition. In this work, he incorporated second order differences, additional training, inter-word triphones, and decision-tree clustering, and obtained a 13% reduction of errors from a vocabulary-dependent system. One of the major findings was that inter-word triphones are effective even when training and testing tasks are disjoint. This disproved the suspicion that inter-word triphones are effective because they capture grammatical constraints. A second finding was that decision tree based allophones (similar to the final paper) reduced errors substantially, while an earlier study from CMU found little benefit. The main difference is that this study started with many more detailed but poorly trained models, which benefited from the generalization capabilities of the decision tree. The final result of a lower vocabulary-independent result was surprisingly good, but it remains to be verified on the same speakers. Also, it remains to be shown that the latest vocabulary-dependent techniques (semi-continuous models, sex-dependent models) are effective under vocabulary-dependent conditions.

The final paper, presented by P.S. Gopalakrishnan of IBM, gives a detailed treatment of decision tree clustering of allophones. Their approach involves first collecting a large corpus of speech, and then automatically segmenting into phone labels, and storing the five left and five right phonetic neighbors. This ensemble of very detailed phonetic segments were then clustered using a decision tree that asked questions about the classes of phonetic neighbors. The leaf nodes of this tree were used as the final allophonic units in a 5000-word continuous speech recognition system. During recognition, simpler models were used until the next word is hypothesized. At that point, the current word is rescored with the appropriate models. The allophonic models yielded a substantially better result than phones and an IBM implementation of within-word triphones. It was also shown that extending contexts to five left and right neighbors gave some improvement, and that the current training dataset could support about 45 allophonic models per phone. The algorithm in this paper differed from the previous paper in several minor ways. Both showed that as the vocabulary increased, decision tree based algorithms that utilize a priori knowledge about context can improve generalization. (An MIT paper in session 2 also used decision tree clustering.) Some questions were raised about the differences among the standard triphones, IBM triphones, and IBM allophones that ask only about one left and right neighbors. Compared to the standard (inter-word) triphones, IBM triphones are within-word only and do not utilize a complicated smoothing algorithm, while the IBM allophones that ask only about one neighbor are clustered models (similar

to CMU's generalized triphones). So the latter might be a better baseline compared to the current DARPA systems, which makes the contribution of the 5-neighbor allophone system smaller, but still appreciable.

With only a few minutes left, the discussion centered around the issue: is context-dependent modeling better than complex context-independent models? The consensus was that while context-independent models are more easily trained in more detail (more mixture densities, states, etc.), they lack the constraints of context-dependent models. The powerful contextual constraints make context-dependent models sharper and more accurate. The last two papers and other earlier work clearly illustrated this point. However, the first two papers used only context-independent models and still achieved good results. This suggests that the new approaches in the first two papers are very promising. On the other hand, their applicability relies upon their extensibility to context-dependent modeling.