Research in Continuous Speech Recognition

PIs: John Makhoul, Richard Schwartz

BBN STC, 10 Moulton St., Cambridge, MA, 02238
Makhoul@@bbn.com, Schwartz@@bbn.com
(617) 873-3332, -3360

The primary goal of this work is to develop improved methods and models for acoustic recognition of continuous speech. Most of the work has focused on deriving statistical models for speech recognition that can capture the acoustic-phonetic phenomena that occur in speech with the constraint that the models can be adequately estimated from a reasonable amount of training speech.

Most of our work in phonetic recognition and word recognition over the past six years has involved hidden Markov models (HMMs). One significant contribution in this area has been a technique we developed for modeling the effects of phonetic coarticulation in a robust way. The technique is based on estimating context-dependent models of each of the phonemes. We have shown that this basic model (and its extensions) has resulted in a significant improvement in word and sentence recognition accuracy.

We have also developed the "stochastic segment model", which can model the correlation between different parts of the phoneme directly. Initial experiments with this model on context-independent phonetic units reduced the recognition error by a factor of two lower than for the corresponding context-independent HMM models. However, the new method requires significantly more computation.

In general, collecting a large amount of speech (about 30 minutes) from the particular speaker who will use the system will result in the highest recognition accuracy. However, one may need to minimize the amount of training speech from a new speaker without losing performance. We have developed a "probabilistic spectral mapping" technique for adapting a model from one speaker to a new speaker based on a small amount of speech. Using this technique, the recognition accuracy with only 2 minutes of training from the new speaker is equal to that usually achieved using 20 minutes of speaker-dependent training.

In this project we have combined various knowledge sources together to produce the BYBLOS speech recognition system. The system was tested on the DARPA Resource Management Database under several grammar conditions and resulted in higher recognition accuracies than had previously been reported for tasks of this complexity.

In the area of real-time speech recognition we have pursued two activities: the implementation of our speech recognition algorithms on a general-purpose parallel processor and a joint effort with UC berkeley and SRI to design and build a special-purpose board-set capable of real-time continuous speech recognition with a large vocabulary (3000 words) and a statistical language model. In this latter activity, we provided the BYBLOS recognition code, and consulted on the changes that would be appropriate for a special purpose VLSI implementation. The first prototype of this board set is expected to be completed by mid 1989.

The parallel implementation research used the BBN $Butterfly^{TM}$ parallel processor. To achieve near linear parallel efficiency on large configurations (97 processors) required some tuning of the communication and synchronization procedures. The final result was a factor of 79 increase in speed on the 97-processor machine over the speed on a single processor. The near-real-time BYBLOS speech recognition system on a 32-processor $Butterfly^{TM}$ was demonstrated and used in several "live" tests.