RESEARCH IN CONTINUOUS SPEECH RECOGNITION

PIs: John Makhoul and Richard Schwartz

BBN STC, 10 Moulton St., Camridge, MA 02138
makhoul@bbn.com, schwartz@bbn.com

The primary goal of this basic research is to develop improved methods and models for acoustic recognition of continuous speech. The work has focussed on developing accurate and detailed mathematical models of phonemes and their coarticulation for the purpose of large-vocabulary continuous speech recognition. Important goals of this work are to achieve the highest possible word recognition accuracy in continuous speech and to develop methods for the rapid adaptation of phonetic models to the voice of a new speaker.

Major Accomplishments

- Developed context-dependent phonetic models based on the hidden Markov modeling (HMM) formalism to describe the acoustic variability of speech due to coarticulation with neighboring phonemes. The method resulted in a reduction of the word error rate by a factor of two over using context-independent models.
- Developed and demonstrated the effectiveness of the "time-synchronous" search strategy for finding the most likely sequence of words, given the input speech.
- Incorporated the various techniques in a complete continuous speech recognition system, called BYBLOS, and demonstrated it first in 1986. It was, and continues to be, the highest-performing continuous recognition system for large vocabularies. The basic methodology of BYBLOS has since been adopted by other DARPA sites.
- Developed a new formalism for phonetic modeling, called "stochastic segment modeling", which can model the correlation between different parts of a phoneme directly. Initial experiments with this model on context-independent phonetic units reduced the recognition error by a factor of two compared to the corresponding context-independent HMM models. However, the new method requires significantly more computation.
- Developed a novel "probabilistic spectral mapping" technique for rapid speaker adaptation whereby the phonetic models of a new speaker are estimated by performing a transformation on the phonetic models of a prototype speaker, using only a small amount of speech from the new speaker. Using this technique, the recognition accuracy with only 2 minutes of training from the new speaker is equal to that usually achieved with 20 minutes of speaker-dependent training or with speaker-independent training (which requires speech from over 100 speakers).