RESEARCH IN CONTINUOUS SPEECH RECOGNITION AT DRAGON SYSTEMS UNDER THE DARPA SLS PROGRAM

Janet Baker, Larry Gillick, and Robert Roth

Dragon Systems, Inc.
320 Nevada St.
Newton, MA 02160

PROJECT GOALS

The primary long term goal of the speech research at Dragon Systems is to develop algorithms that allow us to achieve high performance large vocabulary continuous speech recognition. At the same time, we are concerned to keep the computational demands of our algorithms as modest as possible, so that the results of our research can be incorporated into products that will run on modestly priced personal computers.

RECENT RESULTS

In the past year (1991) Dragon has greatly modified its signal processing and its modeling, with a focus on achieving more accurate speech recognition performance.

At the beginning of the year, Dragon was using 8 spectral parameter signal processing and was training up Resource Management speakers via adaptation of a reference speaker's models. Each PIC (phoneme-in-context) was modeled as a linear sequence of unimodal output distributions (PELs), with the sequence of PELs being chosen once and for all, based on one reference speaker.

The next step in our research was to allow the sequence of PELs to be determined in a speaker dependent way: we called this process "respelling the PICs". Dragon then went on to experiment with several new signal processing representations involving the use of 32 parameters. To make use of the new signal processing parameters (which were a superset of our original 8), it was necessary to have an automatic way of generating a new set of possible output distributions (PELs). Previously PELs had been generated through a sort of human-aided clustering algorithm – a spectrogram labeler initialized a new PEL when he observed an acoustic event that he deemed to be new. An automatic clustering algorithm was developed as an alternative to this labor intensive task. This new process was called "rePELing." Based on these changes and numerous more minor ones, we were able to lower our word error rate (without rapid match) on the RM1 development test data from 5.1% last February to 2.3% by August (using respelling, rePELing, and inverse Fourier transform based cepstral and difference cepstral parameters).

We decided at this point to address the limitations inherent in using unimodal representations for our output distributions, and we embarked on a project to develop a new set of training programs that would be far more flexible and would allow us to more accurately model the true variability of speech. Thus Dragon spent the last third of 1991 implementing a Baum-Welch training algorithm that estimates tied mixture output distributions for each state of each PIC (for an arbitrary choice of independent streams of parameters). The code supports Bayesian smoothing of the mixture weights, which is a key element of the algorithm since there is rarely enough data for the MLE alone to be an adequate estimator. We have debugged our system by focusing on the special case of 32 independent streams, with equally spaced univariate basis distributions for each parameter. We have also simultaneously developed a new training strategy for our rapid match models, based on the idea of manufacturing the rapid match model for a word from the Hidden Markov Model for the word.

PLANS FOR THE COMING YEAR

We plan to continue working with our new training programs for HMMs and Rapid Match, with a strong focus on building speaker independent models. We will be exploring a variety of strategies for choosing basis distributions and for choosing streams. So far we have not tapped the ability of tied mixtures to model the statistical dependence among the parameters. We also plan to work on reducing the memory requirements for our tied mixture models by clustering the PICs and the PELs, and to generally improve the computational efficiency of our implementations. Another major project for the coming year will be the development of a user interface (with error correction facilities) and the addition of the capability for adding words to the vocabulary on the fly.

*This work was sponsored by the Defense Advanced Research Projects Agency and was monitored by the Space and Naval Warfare Systems Command under contract N00039-86-C-0307.