Large Vocabulary Speech Recognition Prototype

Janet M. Baker, Ph.D.

Dragon Systems, Inc.
90 Bridge Street
Newton, MA 02158

Objective

The objective of this project is to develop a system for real-time recognition of natural language continuous speech. To satisfy the objective of "natural language", the system must recognize the sentences actually produced by users in a realistic application -- no artificial grammatical restriction will be allowed merely for the purpose of the speech recognition. On the other hand, the recognition system may be specialized just to recognize speech in a single well-defined application environment and the vocabulary may be restricted (the minimal objective in vocabulary size is 1,000 words). The performance objective is an operational one. The system must include an interactive user interface which includes error correction capability. The error rate must be sufficiently low and the remaining errors must be easy enough to correct that the users in a realistic application are more productive through the user of the speech recognition system.

Approach

The basic approach to speech recognition at Dragon Systems is to model speech as a sequence of random variables or stochastic process. More specifically the speech is modeled (with some variation in details) as a probabilistic function of a Markov process using techniques that are now common among high performance speech recognition systems, where this model is often called a Hidden Markov Model (or HMM). Rather than recide the features which are common to most HMM systems, this summary will list some of the features that are different from other HMM systems.

1. The system uses a "rapid match" algorithm which computes a preliminary score for each word in the vocabulary using models which are much simpler and faster than the HMM models.
2. The acoustic-phonetic system is aimed at modeling the full English vocabulary rather than the vocabulary of a single application task. The acoustic-phonetic training is semi-automatic using human supplied expert knowledge rather than fully automatic.
3. Phonetic duration is modeled in a computationally efficient manner that, for pragmatic reasons, violates the strict theoretical framework of the Markov process. Expected duration is derived partly by rule and partly by automatic training.
4. The system uses not only phoneme moedels, but also sub-phoneme acoustic segments as a unit whose statistics may be shared by different word models.

In general, however, the similarities between the Dragon continous speech system and other HMM based systems are greater than the differences.

Progress

The basic milestone of a real-time implementation of 1,000 word vocabulary continous speech recognition has been achieved. Furthermore, the goal of reducing the amount of computation to fit on a host plus 4 RISC processors has been substantially surpassed. The initial demonstration system runs real-time on an 80386 host plus a single RISC processor, or even on an 80486 host alone.

Recent Accomplishments

- Implementated basic speech pattern matching routines in a fashion designed for ease of portability across algorithms (these routines have been used in both the Stack Decoder and the Time Synchronous Decoder mentioned below), compilers, processors, and operating environments.
- Implemented a Time Synchronous Decoder. This is a system which computes the probability distribution of the hidden Markov process (given the accoustics) in steps strictly increasing in time. The probability of all active states is computed for each time frame before beginning the analysis of the next time frame. This is the most common method of computation for HMM recognizers.
- As an experimental alternative, implemented a Stack Decoder algorithm. The Stack Decoder can potentially be more efficient by analyzing the most promising paths throught the Markov state space first, rather than synchronizing at each time frame. Because of the greater complexity of the Stack Decocer, the current real-time implementation uses the Time Synchronous Decoder.
- Built a probabilistic language model for the dictation of radiology reports.
- Implemented a rapid match algorithm which reduces the amount of computation for a 1,000 word vocabulary by an order of magnitude.
- Ported both the rapid match and the full match algorithms to run on a high speed RISC proccessor, the AMD 29000.
- Developed a simple parallel processing implementation allowing the recognition computation to run in parallel on the 29000 and a host processor.

- Accelerated computation to achieve real-time recognition on either the combination of the 29000 with an 80386 host processor, or even on an 80486 processor by itself.

FY-91 Plans

- On-line adaptive training will be implemented.
- The signal processing will be improved by systematically evaluating the performance of a wide variety of signal processing algorithms.
- The robustness will be improved by modeling alternate word pronunciations.
- The performance on short function words will be improved by developing techniques specifically modeling the phenomena which occur for such words.
- The performance will be improved in general by detailed post-analysis of specific errors.
- The overall error rate will be reduced by a factor of two compared to the current system.