Session 10: Large Vocabulary CSR

George R. Doddington, chairman

SRI International
Menlo Park, CA

OVERVIEW

This session comprised four papers on various topics in speech recognition, followed by a general discussion. The first two papers covered computational search techniques, while the last two papers addressed phonetic modeling issues.

The first paper, "Rapid Match Training for Large Vocabularies", was presented by Larry Gillick of Dragon Systems. This paper described an improved algorithm for building rapid match models for computational efficiency in continuous speech recognition. The technique, designed to accommodate variation in model parameters and phone duration, was demonstrated to provide significant improvement in the miss rate for the correct word. The miss rate remains relatively high however, about 5 percent for a list length of 250 words and a vocabulary size of 5000 words.

During the discussion on this paper, a question was raised regarding the use of a language model in the rapid match. The answer was that, yes, a unigram word probability was used.

The second paper, "An A* Algorithm for Very Large Vocabulary Continuous Speech Recognition", was presented by P. Kenny of INRS. This paper described a new A* stack search algorithm that is only about ten times more computationally expensive than isolated word recognition. Using a 60,000 word vocabulary, the CPU time required to run a perplexity 700 task was 120 times real time on an HP 720 workstation.

During the discussion on this paper, a question was raised regarding the manner in which the search path is extended. The answer explained that the phone endpoints were known and were independent of the search path.

The third paper, "Modeling Spontaneous Speech Effects in Large Vocabulary Speech Recognition Applications", was presented by John Butzberger of SRI. This paper described an analysis of speech recognition errors on spontaneous speech and concluded that the increased error rate on spontaneous speech is attributable to disfluencies and that fluent spontaneous speech exhibits the same recognition performance as read speech. It was also concluded that the use of spontaneous speech in training the recognition system is important for best performance.

During the discussion on this paper, a question was raised regarding how 70 percent of all errors could be labeled as disfluencies. The answer was that the notion of disfluency also comprehended natural phenomena such as vowel elongation and spontaneous speech grammatical constructs (low bigram probabilities).

The last paper, "Speaker-Independent Phone Recognition Using BREF", was presented by Jean-Luc Gauvain of LIMSI. This paper described a series of experiments on speaker-independent phone recognition using the BREF corpus of read speech as prompted using the French newspaper Le Monde. Phone-level performance of 31 percent error was achieved, which is comparable with results achieved on the English TIMIT corpus.

During the discussion on this paper, a question was raised regarding the use of a grammar on this task. The answer was that a grammar was tried but that the error rate was very high. (The perplexity of the grammar was about 500.)

DISCUSSION

The general discussion deviated from the topics covered by the papers and addressed instead pitfalls and issues related to the idiosyncrasies of speech corpora and their impact on speech recognition results and technology.

The SLS ATIS corpus was "attacked" by noting that results were a strong function of the identity of the site which supplied the data. The question was raised multiple times of what was the cause of these differences. Various sources were suggested, including consistent differences in speakers, differences in the acoustics and digitizing system, and differences in the task. Of these three, the last seems most likely, with even primitive measures (such as the number of words per sentence) showing striking differences from site to site.

A general complaint was lodged regarding the imbalance of training data between sites for the MADCOW corpus, with MIT supplying an inordinate fraction of such data. In a mitigating reply to this complaint, it was noted that the MIT sentences are significantly shorter than sentences from other sites, and therefore the imbalance (in terms of the amount of speech data) is not as great as is indicated by the sentence count.

One astute comment categorized speech signal variability as being of two distinct types -- systematic (modelable) and nonsystematic (random). It was further noted that the systematic variability is not handled properly by HMM technology and must be built into the system as a model of the speech process. Nonsystematic variability on the other hand can only be modeled as noise and the only way to handle such variation is through training with large amounts of data. A plea was made to isolate the systematic effects and model them explicitly, so that we don't continue to need more and more data.