RESEARCH IN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION*

Janet Baker, Larry Gillick, and Robert Roth

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

PROJECT GOALS

The primary long term goal of speech research at Dragon Systems is to develop algorithms that are capable of achieving very high performance large vocabulary continuous speech recognition. At the same time, we are also concerned to keep the demands of those algorithms for computational power and memory as modest as possible, so that the results of our research can be incorporated into products that will run on moderately priced personal computers.

RECENT RESULTS

This past year's effort has been devoted to further work on speaker independent training and recognition, as well as to the problem of adapting to new speakers and new microphones. Important projects in the past year have included the incorporation of a trigram language model into the forward pass, the use of phonetically tied mixture models, studies in adaptation, and experiments involving time resolution issues including the effects of the frame rate, the number of nodes per triphone model, and the skipping of nodes.

We implemented our trigram language model in the forward pass of our time synchronous decoder. This was done in an attempt to avoid the search errors being reported at the last HLT meeting when using them only in the final pass of a multipass algorithm. This architecture, however, required much more memory than a multipass implementation. As a result, we hit against the memory limitations of our machines while we still had a significant percentage of search errors. In spite of this, we obtained a substantial reduction in the error rate with trigrams. We also developed our own code for constructing trigram language models from a body of training text.

Our experiments with phonetically tied mixture models showed that these models appear to have an advantage over ordinary tied mixture models in terms of their speaker independent recognition performance as well as in the area of speaker and microphone adaptation. By adapting the phonetically specific basis distributions, one can, in essence, obtain phonetically specific nonparametric transformations of acoustic space so as to quickly capture important speaker or microphone characteristics and correct for them. In addition, phonetically tied mixture models generally require less memory since they tend to need fewer components per mixture.

Unlike most other sites, Dragon continues to favor the use of 20 ms frames over that of 10 ms frames. Our recent experiments indicate that as long as we retain the ability to skip nodes in our triphone models, there is little advantage to the faster frame rate. We also have generally found that there is an advantage to allowing the number of nodes to be variable across triphones. The use of as many as 5 nodes per triphone has been found to be useful for some contexts of some phonemes.

A number of attempts to improve the signal processing of our system led to negative results. We have not yet obtained any benefit from the use of second difference parameters, nor did we obtain any advantage from the application of linear discriminant analysis to a series of concatenated frames. On the other hand, we did find that it was important for us to use gender dependent linear discriminant based transformations.

PLANS FOR THE COMING YEAR

We intend to explore the use of decision trees and general mixture models. In addition we hope to compare the value of full Baum-Welch training to that of the method we currently use, which relies on explicit phoneme alignments. Multipass algorithms will also be implemented, as a vehicle for a more efficient implementation of trigrams and other longer context language models, as well as for speeding up the recognizer. Since the phoneme alignments off of which we build our models do not correspond as closely to human judgement as we might have expected, we plan to explore the question of how difference parameters may influence alignments, especially when mediated through linear discriminant analysis based transformations.

*This work was sponsored by the Defense Advanced Research Projects Agency under contract number J-FBI-92-060.