CSC 248/448 Assignment #3 Due: November 4th, 2003

This assignment involves building and testing a highly simplified speech recognition system. We assume that the speech has already been processed and the input is classified into a very small codebook containing the following symbols (each with an intuitive description of what type of sound it represents): S (silence), H (hissing sound), U (unvoiced consonant), V (vowel sound) and W (another vowel sound).

We will be doing word based recognition, and assume the following HMM template for each word.



a) We are given one instance of training data for the word "sad". Here we indicate where the actual state transitions from S1 to S2 to S3 with a "/".

HSWSCH/VWWHCHSCVHUUUV/CSV/

Using this fully tagged data, what transition and output probability distributions would you get for the HMM using the MLE technique? What is the entropy measure of the training corpus with this distribution?

- b) For the Baum-Walsch re-estimation technique, could you use the probability model developed in part (a) directly as the initial probability estimate? If not, why not, and develop some way to get around the problems you identify. Give your initial revised initial probability estimate using your revised technique. What is the entropy measure of the training corpus with this measure? Did it change from part (a) in a way you expected?
- c) Implement the forward and backward algorithms, and the Baum Walsch re-estimation formulas. You are going to compute first three iterations of the algorithm for two different starting distributions. The first is the one you just computed in part (b). For the second, assume a uniform distribution throughout. Give the probability distributions produce by each iteration (i.e., 6 in all) on the following three new training instances for the word "sad".

HHHHVVVSVVVWWWUWC HHSHHHCWVWVCVVWCWSUCCHV SWUHSVUSVCHWWUSCCS Explicitly describe your strategy for combining the three training instances in your estimation formula. Give is the entropy measure on the training corpus for all six probability models produced.

- d) Using your model in (c), develop a "state tagging" system that given a new instance of the word sad, will indicate the most likely path through the HMM that generated the sequence. Give your results on the first training instance in part (c). Reusing your code to compute the forward and backward probabilities, compute the most likely state that the HMM is in at each time point on the same training instance. Are there cases where the most likely state at time t is not the same as the state on the most likely path at time t? Finally, Submit your tagging program in the usual way for testing on new test data.
- e) (448 students, or extra credit for 248 students) (continuous word recognition) Extend the model developed in part (c) to now do continuous work recognition with a three word vocabulary. You may use the models you've already developed for "sad" in order to help bootstrap the continuous model. *Make sure you document all the design decisions you make, and the experiments you performed in order to optimize your solution.* The documentation and analysis will count for a major part of your grade. Here we are given the following corpora of short sentences (file is directly accessible from the web page):

ONE SAD BUS

UHCUHWWWSWVHCCWUSVHSWWWWHSSCCHUVUSVVVHSHUUC **BUS ONE** CHHCCWWVWUHUSHHCWHUHWWVUHVVVWCHV **ONE BUS** CHUUHVVVVUCWVUCCHWSWVWHHCHH ONE SAD SAD BUS CUVUVCUCWVVVSUVSVWVWVVHWCWCHSSHUHHHHVUCCWVWCC CWHHHCWVSVVCHUWWCSCCVCVVVUVVVHHCHU SAD ONE HHUSCHHUHWHSCSCSSUHCCCCUUCCHVUWUWWCUC BUS ONE BUS CHHUVSHHSCWWWVSHVUHHHVVSHUSSVSWCCVVWSCCWCCHVHC VWUWVHHHUHH **BUS ONE SAD BUS** VCWVWWSVWVHWHHUUHHHSSVSVUVUWVCHCCCCHHHWCVUUCU **CUUHCVUVCWWSVVHHHH**

Submit your program as usual for evaluation on new test data.