Abstract

This thesis examines how artificial neural networks can benefit a large vocabulary, speaker independent, continuous speech recognition system. Currently, most speech recognition systems are based on hidden Markov models (HMMs), a statistical framework that supports both acoustic and temporal modeling. Despite their state-of-the-art performance, HMMs make a number of suboptimal modeling assumptions that limit their potential effectiveness. Neural networks avoid many of these assumptions, while they can also learn complex functions, generalize effectively, tolerate noise, and support parallelism. While neural networks can readily be applied to acoustic modeling, it is not yet clear how they can be used for temporal modeling. Therefore, we explore a class of systems called *NN-HMM hybrids*, in which neural networks perform acoustic modeling, and HMMs perform temporal modeling. We argue that a NN-HMM hybrid has several theoretical advantages over a pure HMM system, including better acoustic modeling accuracy, better context sensitivity, more natural discrimination, and a more economical use of parameters. These advantages are confirmed experimentally by a NN-HMM hybrid that we developed, based on context-independent phoneme models, that achieved 90.5% word accuracy on the Resource Management database, in contrast to only 86.0% accuracy achieved by a pure HMM under similar conditions.

In the course of developing this system, we explored two different ways to use neural networks for acoustic modeling: prediction and classification. We found that predictive networks yield poor results because of a lack of discrimination, but classification networks gave excellent results. We verified that, in accordance with theory, the output activations of a classification network form highly accurate estimates of the posterior probabilities $P(class|input)$, and we showed how these can easily be converted to likelihoods $P(input|class)$ for standard HMM recognition algorithms. Finally, this thesis reports how we optimized the accuracy of our system with many natural techniques, such as expanding the input window size, normalizing the inputs, increasing the number of hidden units, converting the network's output activations to log likelihoods, optimizing the learning rate schedule by automatic search, backpropagating error from word level outputs, and using gender dependent networks.