IEEE Conference on NIPS, 1988.11.28, Denver, USA.

Modularity in Neural Networks for Speech Recognition

A. Waibel
ATR Interpreting Telephony Research Laboratories

June 13, 1988

Summary

In a previous study (Waibel et al., ATR TR-I-0006, 1987) it was shown, that "Time Delay Neural Networks" (TDNNs) can achieve superior recognition performance (99%) for difficult phonetic discrimination tasks (e.g., "B", "D", "G"). TDNNs achieve such high performance partially by their ability to inhibiting all incorrect outputs in addition to activating the correct output unit. Although this property leads to good performance for such subphonetic tasks, it raises serious doubt on whether such networks scale, more specifically, whether they can be extended to networks that can handle the full set of phonemes: Since all false categories must be considered during training, learning additional categories incrementally may not be possible and complete relearning on a full set may be computationally too expensive. To alleviate this problem we have explored techniques for constructing larger phonetic neural networks from component subnets, by exploiting the hidden structure of previously trained component subnets. We present results from experiments on a stop consonant recognition task, i.e., "B","D","G","P","T" and "K". A neural network performing this six category classification task was obtained by 1.) completely retraining a larger 6 class network, 2.) training only the combination of 8 lower layer hidden units from previously trained BDG- and PTK-nets, 3.) training the combination as before with the addition of 4 free hidden units as "connectionist glue" and 4.) training the combination of lower layer hidden units from a BDG- and PTK-net and from a separately trained voiced/unvoiced net. Recognition experiments on test data show, that networks constructed by the latter two methods achieve best performance. While learning time was kept low, the resulting nets could classify the 6 stops with an accuracy of 98.4%, i.e., with an accuracy as high as the original component subnets (98.8% for BDG and 97.9% for PTK). The feasibility of this approach has since been verified successfully on a stop/nasal task and on all consonants. It's success suggests that modular incremental learning in neural networks may indeed be possible and could provide the basis for further extensions of neural network based speech systems.