DISCRIMINATIVE TRAINING OF SELF-STRUCTURING HIDDEN CONTROL NEURAL MODELS

Helge B.D. Sorensen ¤, Uwe Hartmann #, and Preben Hunnerup #

¤ Department of Applied Electronics
Technical University of Denmark, DK-2800 Lyngby, DENMARK

# Institute of Electronic Systems
Aalborg University, DK-9220 Aalborg, DENMARK

ABSTRACT

This paper presents a new training algorithm for Self-structuring Hidden Control Neural (SHC) models, which we presented at ICASSP[1]. The SHC models were trained non-discriminatively for speech recognition applications [2]. Better recognition performance can generally be achieved, if discriminative training is applied in stead. Thus we developed a discriminative training algorithm for SHC models, where each SHC model for a specific speech pattern is trained with utterances of the pattern to be recognized and with other utterances. The discriminative training of SHC neural models has been tested on the TIDIGITS database [3].

1. INTRODUCTION

The SHC models are hybrid pattern recognition models combining advantages of Neural models and Hidden Markov models, e.g. the universal approximation capabilities and the temporal modeling capabilities, respectively. In our application we apply one SHC model for each pattern in the vocabulary. An SHC model automatically develops the model architecture during training avoiding the manual selection of the architecture before training. Each SHC pattern model realizes a finite state (left-to-right) model and each specific state is defined using a related binary vector code at the input terminals of the model. N states thus necessitate N different binary vector codes. In other words we have N possible mappings in the SHC neural model. For each state the SHC model is used as a non-linear predictor and the aim is to predict a spectral vector (consisting of cepstrum coefficients) at time t from one or more of the preceeding spectral vectors. Each SHC speech pattern model is trained in the proposed discriminative training framework described in the following.

2. DISCRIMINATIVE TRAINING OF SHC NEURAL MODELS

In the paper by Y. Liu et al. [4] a discriminative training algorithm is proposed for predictive neural network (PNN) models. Each PNN model models has been used to model a speech pattern. In a PNN model a finite (left-to-right) model is realized using a neural network for each state to establish a non-linear predictor in each state. We use only one neural network in the SHC model to realize the same number of states thus the PNN training algorithm can not be applied for our models. This fact inspired us to develop a new discriminative training algorithm for the training of each SHC model. The following two equations describe an SHC model [1]. For simplicity we have only written the equations for a model with one output. An extension to a model with more

0-7803-2431-5/95 $4.00 © 1995 IEEE

than one output is straight foreward. The two equations for an SHC model can be written as the following:

$$\tilde{y}_j = \Psi\left[\sum_{k=1}^{r} a_{jk}\, x_k - \theta_j + \sum_{k=1}^{j-1} a_{j,r+k}\, \tilde{y}_k\right] \tag{1}$$

$$f(x_p) = \sum_{j=1}^{q} a_{q+1,j+r}\, \tilde{y}_j + \sum_{j=1}^{r} a_{q+1,j}\, x_j - \theta_0 \tag{2}$$

where r is the number of inputs, q is the number of hidden neurons, x_k is an element in a spectral vector x_p from the utterance x (consisting of a sequence of vectors) or an element in the binary code (control vector) describing the states, $f(x_p)$ is the output and a- and θ-variables are the weights in the neural model. The output is normally a vector i.e. a prediction of a spectral vector.

The following figure describes a two state SHC model:

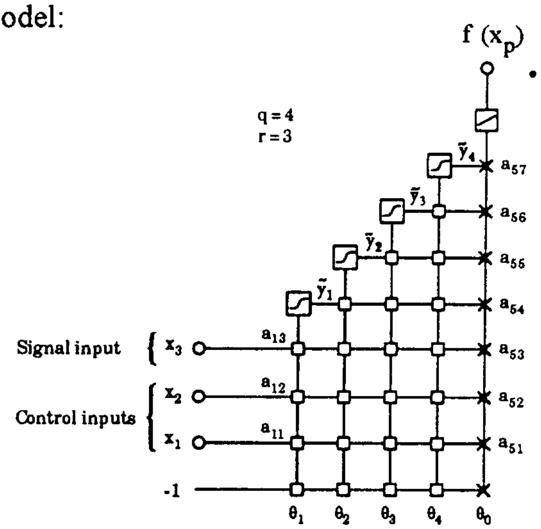

Figure 1 A single-input/single output SHC model with q hidden neurons. r + 1 is the number of input terminals. The hidden neurons have non-linear input-output relations and the output unit has a linear input-output relation. Some of the weights are frozen during training and this is indicated by the small boxes. The dots indicate the self-structuring capability (addition of neurons during training).

The weights in the SHC model are updated using the quick propagation algorithm [5] but the algorithm is not applied to perform the classical error minimization. In our discriminative training framework we changed the algorithm to maximize the global probability of the SHC pattern models making the right classification. The global probability of the classifier making the right classification is defined as:

$$L(\Lambda) = \sum_{j=1}^{J} \sum_{x \in s_j} g_j(x, \Lambda) \tag{3}$$

$\Lambda = \{\lambda_1, \lambda_2, \ldots\ldots, \lambda_J\}$. Λ describes the complete parameter set of the classifier and λ_j represents the parameter set of the j-th SHC model i.e. the model for the j-th speech pattern class. The function g_j (x, Λ) is the probability of the speech utterance x belonging to the j-th class. s_j is the subset of the training set belonging to the j-th class. The gradient to be used in our discriminative training algorithm is the gradient of the global probability function L(Λ). This gradient is defined in [4] for PNN models, but it can not be applied directly for our SHC models due to the fact that a PNN model contains a set of neural networks, one for each state, in contrast to our SHC model that has only one neural network. We have derived the following gradient to be used in the quick propagation algorithm. A component of the gradient is given as:

$$\begin{aligned}(\partial L(\Lambda)/\partial\lambda_j)_{m,i} = {} & -\alpha \sum_{x \in s_j} g_j(x;\lambda)\,[1-g_j(x;\lambda)] \\ & \times \sum_{p=1}^{T(x)} \left(f(x_p)_m^{j,c(\pi)} - d_{p,m}\right) x_{p,i} \\ & + \alpha \sum_{k \neq j} \sum_{x \in s_k} g_k(x;\lambda)\, g_j(x;\lambda) \\ & \times \sum_{p=1}^{T(x)} \left(f(x_p)_m^{k,c(\pi)} - d_{p,m}\right) x_{p,i}\end{aligned} \tag{4}$$

This is the partial derivative with respect to λ_j

(and index m and i) for discriminative training of a SHC model. i is the index of the i-th input to the model and m is the index of the m-th input. α is a positive constant. $d_{p,m}$ is the desired output at the m-th output of the model and the f-expression is the measured m-th output. $c(\tilde{\pi})$ is refering to the sequence of control vectors which results in minimum prediction error for the SHC model explained by the fact that this model is functioning as a one-step non-linear predictor. The sequence $c(\tilde{\pi})$ is found by using the Viterbi-algorithm. π is a path and $\tilde{\pi}$ is the optimal path found by the Viterbi-search algorithm.

The discriminative training of the SHC models can briefly be described as using the gradient in equation (4) in the the quick propagation algorithm to maximize the global probability probability (see equation (3)), of the classifier making the right classification.

The discriminative training algorithm for SHC models is now developed, and each SHC model for a specific speech pattern can now be trained with utterances of the pattern to be recognized and with other utterances, see equation (4). The discriminative training of SHC neural models has been tested on the TIDIGITS database as described in the next section.

3. EXPERIMENTS

Preliminary results are very promising e.g. the recognition rate based on female utterances (words) from the TI-DIGITS database was 99% (on training data) and this is better than our non-discriminative approach [2]. The SHC Models trained with the new discriminative training algorithm was evaluated using ESPRIT SAM evaluation methods [6].

The training set contains utterances (the digits 1 to 5) from 57 women and the test set contains utterances from 55 women different from the group of women in the training set. The sampling frequency is 8 kHz and based on an 8-th order LPC analysis 12 cepstrum and 12 Δ cepstrum coefficients were calculated every 10 msec based on a 20 msec window. The number of states in each SHC model is 8.

A few initial sets of experiments were performed using different stop criteria for the maximum number of neurons to be generated during training in the SHC models. The best result achieved on the training set was a recognition rate of 99% and the best results on the test set were rates 97%-98%. These preliminary experiments indicated that performance can be improved by fine tuning the settings of the stop criteria in the SHC models.

4. CONCLUSION

The discriminative training algorithm for SHC neural models has been derived and tested. The preliminary results are promising and indicate that the approach is a relevant alternative to non-discriminative training of the models. Further experiments including the application of different stop criteria in the SHC models are expected to improve the test results described in the previous section.

Some of the advantages of the SHC models compared to the PNN models [4] are: (1) Only one neural network is applied to realize a finite N state model in stead of N neural networks in each PNN model. (2) The model architecture is generated automatically during training which typically results in a more efficient model architecture. (3) The proposed discriminative training of an SHC model is more efficient due to the fact that this model only consists of one neural network.

REFERENCES

[1] Sorensen H.B.D., Hartmann U., "Pi-Sigma and Hidden Control based Self-structuring

Models for Text-independent Speaker Recognition" , in IEEE Procedings ICASSP93, Minneapolis, USA, 1993.

[2] Sorensen H.B.D., Hartmann U., "Self-structuring Hidden Control Neural Models for Speech Recognition", in Proceedings ICASSP92, San Francisco, USA, March 1992.

[3] Texas Instruments and National Institute of Standards and Technology, "Studio Quality Speaker-Independent Connected-Digit Corpus (TI-DIGITS)", NIST Speech Discs 4-1, 4-2 and 4-3, February 1991.

[4] Liu Y., Lee Y.C., Chen H.H., Sun G.Z., "Discriminative Training Algrithm for Predictive Neural Network Models", in IEEE/INNS Proceedings IJCNN92, Baltimore, USA, June 1992.

[5] Fahlman S.E., Lebiere C., "The Cascade-Correlation Learning Architecture", CMU-CS-90-100, Carnegie Mellon University, Pittsburgh, February 1990.

[6] A.J. Fourcin, G. Harland, W. Barry, W. Hazan, "Speech Input and Output Assessment, Multilingual Methods and Standards", Ellis Horwood Books in Information Technology, 1989.