A HYBRID RBF-HMM SYSTEM FOR CONTINUOUS SPEECH RECOGNITION

W. Reichl and G. Ruske

Institute for Human-Machine-Communication,
Munich University of Technology,
Arcisstr. 21, D-80290 München, Germany

ABSTRACT

A hybrid system for continuous speech recognition, consisting of a neural network with Radial Basis Functions and Hidden Markov Models is described in this paper together with discriminant training techniques. Initially the neural net is trained to approximate a-posteriori probabilities of single HMM states. These probabilities are used by the Viterbi algorithm to calculate the total scores for the individual hybrid phoneme models. The final training of the hybrid system is based on the 'Minimum Classification Error' objective function, which approximates the misclassification rate of the hybrid classifier, and the 'Generalized Probabilistic Descent' algorithm. The hybrid system was used in continuous speech recognition experiments with phoneme units and shows about 63.8% phoneme recognition rate in a speaker-independent task.

1. INTRODUCTION

A hybrid system for continuous speech recognition, based on a neural network (NN) with Radial Basis Functions (RBFs) and Hidden Markov Models (HMMs), is presented in this paper. Neural networks show superior pattern classification performance in static classification tasks due to their discriminant learning algorithms, while the HMM structure is able to cope with the temporal distortions in speech, using the excellent temporal alignment properties of the Viterbi algorithm. Therefore a hybrid NN-HMM system is proposed to benefit from both advantages [1,2,3,5,8,11].

This hybrid NN-HMM system utilizes a neural net with Radial Basis Functions to approximate a-posteriori probabilities of HMM states. RBFs are chosen since initialization of mean and range of the basis functions is possible by cluster techniques, such as the LBG algorithm, and the fast learning capabilities of RBF networks [6,7,9,10,11].

A two-phase discriminative training technique is used to optimize the NN parameters. Initially the hybrid system is trained by the backpropagation algorithm, optimizing a mean squared error (MSE) function. This results in an approximation of a-posteriori probabilities of HMM states [1,10]. The total scores for the individual hybrid RBF-HMM models are calculated by the Viterbi algorithm, estimating Bayes probabilities of the models.

In a second training phase the misclassification rate of the hybrid classifier is approximated by the 'Minimum Classification Error' (MCE) objective function, which is finally optimized by the 'Generalized Probabilistic Descent' (GPD) algorithm and results in a classifier with minimum error probability [4,8]. Discriminant training techniques provide better performance as compared to Maximum Likelihood Estimation (MLE), if not enough training data is available or the modeling assumptions do no fit the data [2,11].

The performance of the hybrid RBF-HMM system is reported for speaker independent speech recognition experiments, where about 63.8% phoneme recognition rate was achieved.

2. THE HYBRID RBF-HMM SYSTEM

The hybrid RBF-HMM system consists of a neural network with Radial Basis Functions, estimating a-posteriori probabilities $p(q_{nm}|X)$ for HMM states q_{nm}, conditioned by the acoustic input X [1,10]. To improve the approximation capabilities of the NN a context window in the input layer of the NN can be incorporated [1,2,7,8]. These a-posteriori probabilities are used as discriminant local probabilities in discriminant HMMs [2].

One phoneme model W_n is made up of 4 to 6 states q_{nm}, which are assigned to the corresponding NN output nodes O_{nm}. The total score $g_n(X)$ of a first order Markov model W_n for a feature vector sequence $X = \{x_1, \dots, x_T\}$ is calculated by the Viterbi algorithm, using the optimal sequence of states $Q_n = \{Q_{n1}, \dots, Q_{nT} | Q_{nt} \in \{q_{nm}\}\}$:

$$g_n(X) = \prod_{t=1}^{T} O_{Q_{nt}}(t) = P(Q_{n1}, \dots, Q_{nT}|X). \qquad (1)$$

This total score is an approximation of the a-posteriori probability $P(W_n|X)$ of the model W_n for the acoustic input X, if only the optimal sequence of states is considered [8,2]:

$$\begin{aligned} P(W_n|X) &\approx P(Q_n|X)P(W_n|Q_n) & (2) \\ &\approx g_n(X) \end{aligned}$$

The second factor in (2) is independent of the acoustic input and reflects the structure of the models. It is assumed $P(W_n|Q_{n1}, \dots, Q_{nT}) := 1$ for all valid state sequences in phoneme model W_n. Additional transition probabilities are used in the Viterbi decoding, which are counted during the segmentation of the training data in the individual states. The scores $g_n(X)$ for the models estimate the a-posteriori

This work was funded by the German Federal Ministry for Research and Technology (BMFT) in the framework of the Verbmobil Project under Grant 01 IV 102 C6. The responsibility for the contents of this study lies with the author

Presented at the *Proc. ICASSP-95*, Detroit, USA, May 1995, pp. 3335-3338.

probabilities of the models, which are the optimal discriminant measures according to Bayes theory. The 1-stage DP algorithm can be utilized to achieve an optimal segmentation and classification in a continuous speech recognition task.

The utilized neural network is a Radial Basis Function network, consisting of one layer of Radial Basis Functions and additional layers of neurons with sigmoidal summing neurons. The purpose of the basis function layer is a statistically based transformation of the features into a high dimensional space [9], whereby the sigmoidal nodes are used in the final transformation to calculate the required a-posteriori probabilities. The activations of the sigmoidal neurons are computed according to the usual summing rule: $O_{nm} = f\left(\sum_i W_{nmi} o_i\right)$, utilizing the sigmoid transfer function $f(a) = (1+e^{-a})^{-1}$. The nonlinear sigmoidal transformation in the output layer is needed for a better approximation of the a-posteriori probabilities and the minimization of the classification error rate [7,2].

Each basis function i computes the Mahalanobis distance between the input vector $\vec{x}$ and its mean $\vec{m}_i$, using a diagonal covariance matrix C_i. The activation of the RBF neuron i is calculated by the exponentially weighted Mahalanobis distance:

$$o_i = \exp\left(-\frac{1}{2}(\vec{x}-\vec{m}_i)^T C_i^{-1}(\vec{x}-\vec{m}_i)\right). \qquad (3)$$

This definition is similar to Gaussian functions, applied in statistical methods such as HMMs. There mixtures of normal density functions are used to approximate multimodal probability densities. In (3) a different normalization is used, which limits the range of activation to $0 \leq o_i \leq 1$ and makes the maximum value of o_i independent of the covariance matrix C_i. Constant activations of RBF neurons are located on ellipses in the feature space.

The basis function neurons in the hybrid RBF-HMM system are organized in subnets for the different features, which are derived every 10ms from the acoustic preprocessing. The subnetwork for the 20-dimensional Bark-scaled loudness spectrum consists of 256 RBF nodes, the subnet for the delta-loudness spectrum (20-dim.) of 128 RBF nodes and the subnetwork for the total loudness of 16 RBF nodes. Altogether this results in 400 Radial Basis Functions for the hybrid system. Mean vectors and variances of the basis functions are initialized by the LBG clustering algorithm [7].

To calculate the a-posteriori probabilities of the states $p(q_{nm}|X)$ a normalization of the basis function outputs is needed. If Bayes' rule is used and multimodal distributions of the features are estimated with Gaussian mixtures, the Bayes probability for state q_{nm} can be expressed as a mixture of a-posteriori probabilities $p(i|X)$ of the basis functions:

$$\begin{aligned} p(q_{nm}|X) &= \frac{p(q_{nm})\sum_i p(X|i)p(i|q_{nm})}{p(X)} \\ &= \sum_i p(q_{nm}|i)p(i|X). \qquad (4) \end{aligned}$$

If the a priori probabilities of the basis functions $p(i)$ and the determinants of the covariance matrices $|C_i|$ of the basis functions are identical, the Bayes probability is approximated by the output nodes:

$$p(q_{nm}|X) \approx O_{nm} = \frac{\sum_i W_{nmi} o_i}{\sum_i o_i}. \qquad (5)$$

No sigmoidal transformation in the output nodes is needed for the estimation of Bayes probabilities by mixtures of a-posteriori probabilities of the basis functions.

In case of statistical independent features the Bayes probabilities of the states are the product of the independent calculated probabilities, conditioned by the individual features [10]. This multiplication can be approximated by the usage of the sigmoidal transfer function in the output nodes. The nonlinear mapping is able to create the required products of the individual Bayes probabilities, which can be seen by a Taylor expansion of the sigmoidal mapping [2]. The output nodes of the RBF net are then computed by the following equation:

$$p(q_{nm}|X) \approx O_{nm} = f\left(\sum_{CB} \frac{\sum_i W_{nmi} o_i}{\sum_i o_i}\right). \qquad (6)$$

The summation in (6) runs over all basis functions in all subnets or codebooks (CB), whereas an individual normalization of the basis function outputs to sum up to 1.0 for each subnet is used.

The structure of the hybrid RBF-HMM system is similar to tied-mixture or semi-continuous HMMs (SCHMMs) with separate codebooks for the individual features. The RBF layer is working like SCHMM codebooks, but the normalization for estimating a-posterioris of the RBFs is added. The actual basis functions for the sigmoidal units are the a-posterioris of the RBFs. These are used to form the approximation of the required mapping. The weights of the summing neurons resemble SCHMM mixture coefficients, but are not constrained to sum up to one. The sigmoidal transfer function supports the computation of a-posteriori probabilities [2] and is not common in HMMs. The sigmoidal units sum up all the a-posteriori estimates from the different codebooks (6), no explicit multiplication of probabilities takes place. Additional layers with sigmoidal units can be added to allow a more complex transformation by the neural network. By the application of a context window in the neural net the processing of acoustic information from more than one acoustic vector is possible, exploiting correlations in the input data [2].

In contrary to MLE optimization for HMMs, NN training procedures are inherently discriminative. Bayes probabilities of all states and models are calculated simultaneously by the RBF network, while HMMs estimate separate densities for the individual states or models. During training all parameters of the hybrid system are updated by the presentation of any utterance in the training data, independent of the class membership. Discriminant training techniques estimate class boundaries and not the parameters of assumed model distributions. Although both types of training are theoretically equivalent (if sufficient classifier parameters and enough training data exist, if Gaussian mixture assumptions are appropriate and if a priori probabilities are known) discriminant training techniques provide better performance if these requirements are not met [11].

3. TRAINING OF THE HYBRID RBF-HMM SYSTEM

The discriminative training of the hybrid system occurs in two phases. In the bootstrap phase the usual backpropagation algorithm is used to optimize a quadratic error function (MSE). The parameters of the NN are updated after the presentation of each pattern by a standard gradient descent rule and the optimization is stopped when the performance for an independent test set decreases (cross-validation) [2,7]. The target values for the output nodes of the NN on frame level were assigned by state labels, derived from the Viterbi decoding in SCHMMs [3,11,8]. The basis function parameters are initialized by the LBG cluster algorithm and kept fixed during training. A retraining of the RBF means is difficult, due to the normalization of the RBF layer. Every a-posteriori estimate of a RBF node is very dependent on the outputs values of all other nodes in his subnet and therefore the individual RBF nodes can't be treated separately. The a-posteriori estimates are also very sensitive to variations in the means and variances of the basis functions and hence retraining of these parameters is likely to be instable.

The optimization of the MSE leads to an approximation of a-posteriori probabilities of individual states, conditioned by the actual feature vector x_t: $O_{nm}(t) \approx P(q_{nm}|x_t)$ [1,10]. The total scores are estimates of the a-posteriori probabilities $P(W_n|X)$ for the models W_n (2). The neural net is trained to reproduce the given sequence of states, which is derived from a Viterbi decoding in SCHMMs. This sequence is not guaranteed optimal for the classification, and hence an embedded optimization for the hybrid models is utilized in the second learning phase to minimize the phoneme error rate [8]. The parameters from the MSE optimization are used for initialization. The 'Minimum Classification Error' (MCE) objective function is used to approximate the misclassification rate of the hybrid classifier. The optimization of this error function by the 'Generalized Probabilistic Descent' (GPD) algorithm results in a classifier with minimum error probability [4].

A generalized distance function is used as discriminance measure between the log score $r_c(X) = \log(g_c(X))$ of the correct model c and the log scores of the incorrect models $n = 1, \ldots, N; n \neq c$ [4]:

$$d_c(X) = -r_c(X) + \log\left\{\frac{1}{N-1}\sum_{n;n\neq c} e^{r_n(X)\eta}\right\}^{\frac{1}{\eta}} \text{ with } \eta > 0. \tag{7}$$

A false decision results in $d_c(X) > 0$, while $d_c(X) < 0$ indicates a correct classification: $g_c(X) > g_n(X); \forall n \neq c$. The discriminance measure is continuous with respect to the classifier's parameters and therefore suitable for a gradient-descent optimization. The following cost function is approximating the error rate of the classifier:

$$l_c\left(d_c(X)\right) = \frac{1}{1 + e^{-\gamma d_c(X)}} \text{ with } \gamma > 0. \tag{8}$$

This is a smoothed 'zero-one' cost function, counting the classification errors. The optimization of this objective function with respect to the parameters results in a minimum error classifier [4].

The parameters of the hybrid RBF-HMM system are adjusted proportionally to the negative gradient of the objective function (8). The training is concentrated on likely confuseable phonemes, because the derivatives of cost and discriminance measure disappear for secure classifications. The error is back-propagated along the Viterbi alignment into the NN by the successive use of the chain rule [8]. The training algorithm can be used for the adaptation of weights in the sigmoidal nodes and in principle for the update of the means and variances of the basis functions. The convergence of the GPD algorithm to the optimal Bayes classifier is ensured with proper initialization and appropriate selection of the learning step size [4]. After the MCE-GPD training the NN outputs can no longer be interpreted as probabilities. No segmentation of the training data in states is required within the phoneme models. The algorithm can be extended to the embedded training of whole sentences for known sequences of phonemes.

The basic idea of the MCE-GPD training algorithm, which is normally used for HMM training [4], is similar to the embedded time alignment in MS-TDNN [3] or the 'Figure Of Merit' training for the hybrid RBF-HMM wordspotter in [5], but no fixed phoneme or word level targets are used. In contrast to HMMs in the RBF-HMM a-posteriori probability estimates are processed for the computation of the model scores. Therefore the additional normalization and the sigmoidal transfer functions are included in the RBF net structure.

4. EXPERIMENTS

A database of 100 German speakers (Phondat "Diphon"-database) was used for the speaker independent phoneme recognition experiments. We applied about 7700 sentences from 67 speakers for training and about 3300 sentences from 33 other speakers for the test of the hybrid RBF-HMM. The speech data was sampled at 16kHz, and a 256-point FFT with Hamming window was calculated every 10ms. The power spectrum was combined in 20 critical bands. This Bark-scaled loudness spectrum was normalized to sum up to one. The total loudness and the delta-loudness spectrum were added as separate features. Every feature is processed by a separate subnet in the RBF layer. 41 phoneme models (included silence) with 3 to 6 states were utilized and resulted in a neural net with 169 output nodes (total no. of HMM states). The targets for the initial MSE training were provided by SCHMMs, delivering the required state labels within fixed segment boundaries. In the first phase the RBFs were trained to reproduce the sequence of states. The MCE-GPD algorithm was used for the final optimization of the phoneme error rate.

Phoneme recognition rates for training and test data of different hybrid RBF systems are reported in table 1. In the first line the results for a NN, consisting of one normalized RBF-layer with 400 basis functions in three different subnets and one output layer with 169 sigmoidal neurons, are shown. No hidden layer and no context window was used for the estimation of a-posteriori probabilities. 59.2% of the phonemes in the independent test data are correctly classified after the MSE training, while the phoneme recognition rate for the MCE-GPD trained system is 60.8%.

RBF		MSE		MCE-GPD	
Context	Hid. Units	Train.	Test	Train.	Test
1	-	60.8	59.2	62.3	60.8
1	100	57.3	55.8	60.4	59.3
3	-	64.7	62.5	66.0	63.8

Table 1: Phoneme recognition rates for different hybrid RBF-HMM systems; all values in %.

In line two the results for a NN with one additional hidden layer of 100 sigmoidal units are depicted. The recognition rates for both training procedures are inferior to the net without hidden layer. This is mainly contributed to the relatively small hidden layer with 100 nodes, which acts like a bottleneck in computing the state probabilities. The information from 400 RBF a-posteriori probabilities is compressed in 100 hidden node scores and expanded to calculate the 169 state probabilities. The number of trainable parameters (weights) for this system is about 57,000 and for the baseline system without hidden layer about 68,000. The enhanced transformation capability from the hidden layer was of no additional use for this task.

In a third experiment contextual information from the adjacent feature vectors is used for the calculation of state probabilities in the net. The output nodes refer to the a-posteriori probabilities of the RBF nodes from 3 frames. This expands the hidden layer to 1200 normalized scores and results in about 203,000 trainable parameters. Since the computing of the delta-loudness incorporates the processing of five frames for every feature vector, the usage of the 3 frame contextual window results in acoustic information from 70ms in total. The phoneme recognition results for this RBF net are printed in line three of table 1. 62.5% of the phonemes in the test data are correctly classified after the MSE optimization and 63.8% after the MCE-GPD training. The incorporation of contextual information in the estimation of a-posterioris leads to some improvements in performance.

The minimum error (MCE) training in the second optimization phase is started after no more increase in recognition performance on the test data for the MSE training occurred. The optimization of the MCE objective function, which is more related to the classifier error rate than the MSE, leads to some additional improvements in performance. The best result (63.8%) for the RBF-HMM is achieved, exploiting contextual information. SCHMMs with the same number of prototypes and model structure were trained with MLE for comparison. The SCHMM phoneme recognition rate for the training data is 58.5% and 57.9% for the test data. The improvement of about 6% for the RBF-HMM is attributed to the discriminative structure based on Bayes probabilities, the discriminative learning techniques and the incorporation of contextual input.

5. CONCLUSIONS

In this paper a hybrid NN-HMM system was presented. Radial Basis Functions were used in the neural network, because of the possibility for their good initialization and their similarities to HMMs. Two objective functions were used in the optimization of the NN parameters. The MSE training needs information about the state distribution within the models, while MCE-GPD minimization is based on model error rate approximation and requires model level supervision. The proposed RBF-HMM is similar to discriminatively trained HMMs, but is based on a-posteriori probability estimations. The hybrid system was used in speaker independent phoneme recognition experiments and shows good performance with 63.8% phoneme recognition rate for test data from unknown speakers. Furthermore our intention is to integrate this RBF-NN in a continuous speech recognition system with a lexicon and a beam search to obtain word and sentence recognition rates.

6. REFERENCES

[1] H.Bourlard, C.J.Wellekens, *Links Between Markov Models and Multilayer Perceptrons*, IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 12, no 12, pp. 1167-1178, December 1990.

[2] H.Bourlard, N.Morgan, *Connectionist Speech Recognition: A Hybrid Approach*, Kluwer Academic Publishers, Boston, 1994.

[3] P.Haffner, M.Franzini, A.Waibel, *Integrating Time Alignment and Neural Networks for High Performance Continuous Speech Recognition*, IEEE Proc. 1991 Int. Conf. Acoust. Speech Signal Process., Toronto, pp. 105-108, May 1991.

[4] B.H.Juang, S.Katagiri, *Discriminative Learning for Minimum Error Classification*, IEEE Trans. on Signal Processing, vol. 40, no. 12, pp. 3043-3054, December 1992.

[5] R.Lippmann, E. Singer, *Hybrid Neural-Network/ HMM Approaches To Wordspotting*, IEEE Proc. 1993 Int. Conf. Acoust. Speech Signal Process., Minneapolis, pp. 565-568, April 1993.

[6] J.Moody, C.J.Darken, *Fast Learning in Networks of Locally-Tuned Processing Units*, Neural Computation, vol. 1, no. 2, pp. 281-294, 1989.

[7] W.Reichl, G.Ruske, *Syllable Segmentation of Continuous Speech with Artificial Neural Networks*, EUROSPEECH, Berlin, pp. 1771-1774, September 1993.

[8] W.Reichl, P.Caspary, G.Ruske, *A New Model-Discriminant Training Algorithm For Hybrid NN-HMM Systems*, IEEE Proc. 1994 Int. Conf. Acoust. Speech Signal Process., Adelaide, pp. 677-680, April 1994.

[9] S.Renals, R.Rohwer, *Phoneme Classification Experiments Using Radial Basis Functions*, Proc. Int. Joint Conf. on Neural Networks, Washington, D.C., vol. 1, pp. 461-467, June 1989.

[10] M.Richard, R.Lippmann, *Neural Network Classifiers Estimate Bayesian a Posteriori Probabilities*, Neural Computation, vol. 3, no. 4, pp. 461-483, 1991.

[11] E.Singer, R.Lippmann, *A Speech Recognizer Using Radial Basis Function Neural Networks in an HMM Framework*, IEEE Proc. 1992 Int. Conf. Acoust. Speech Signal Process., San Francisco, pp. 629-632, March 1992.