A FULLY RECURRENT NEURAL NETWORK FOR RECOGNITION OF NOISY TELEPHONE SPEECH

K.Kasper, H.Reininger, D.Wolf, and H.Wüst

Institut für Angewandte Physik
Johann Wolfgang Goethe-Universität
Frankfurt am Main, FRG

ABSTRACT

For a variety of telephone applications it is sufficient to realize a speech recognition system (SRS) with a system vocabulary consisting of a few command words, digits, and connected digits. However, in the development of a SRS for application in telephone environment it has to be considered that the speech is bandpass limited and a high recognition performance has to be guaranteed under speaker independent and even adverse conditions. Furthermore, it is important that the SRS is efficiently implementable.

Fully recurrent neural networks (FRNN) provide a new approach for realizing a robust SRS with a single network. FRNN are able to perform the process of feature scoring discriminatively and independently of the length of the feature sequence. In SRS based on Hidden Markov Models (HMM), different methods have to be applied for scoring the feature vectors and for compensating the variations in phone durations.

Here we report about investigations to realize a monolithic SRS based on FRNN for telephone speech. Besides isolated word recognition, the capability of FRNN-SRS to deal with connected digit recognition is presented. Furthermore, it is shown how FRNN could be immunized against several types of additive noise.

1. INTRODUCTION

SRS are faced with two basic problems. First, contextual information between feature vectors must be exploited during the feature scoring, i.e. the assigning of likelihood values to a feature vector characterizing its belonging to the phoneme or word categories. The more contextual information about a feature vector is taken into account in the scoring process the more uniquely the likelihoods indicate a specific category. The second basic problem in speech recognition is the compensation of variations in phone durations which usually is performed by means of dynamic programming methods. These algorithms are completely different compared to that used for feature scoring and thus, lead to a SRS with a heterogeneous structure which is not well suited for efficient implementation.

One approach to get an efficiently implementable SRS is to use FRNN. FRNN are the most general type of recurrent network because all neurons are interconnected. The performance of FRNN is therefore not limited due to structural constraints. In principal FRNN represent a nonlinear dynamical system with memory. Consequently, it should be possible to classify a feature sequence independently of its length and thereby exploiting contextual information of feature vectors automatically.

2. FULLY RECURRENT NEURAL NETWORKS

Time discrete FRNN consist of fully connected neurons. The neurons have identical structure and the connection between two neurons possesses a minimum time delay of one time step. Because of the recurrent structure, FRNN are networks with dynamic behaviour and an infinite memory.

To distinguish between different types of neurons in a FRNN, the set of input neurons are denoted as $\mathcal{I}$, the set of hidden neurons as $\mathcal{U}$, and the set of output neurons as $\mathcal{O}$. With each input pattern $\underline{x}(t)$ the activities of all neurons are updated synchronously and then an output pattern is emitted. The activity of neuron j at time $t+1$ is given by

$$h_j(t+1) = \sum_{i \in \mathcal{U} \cup \mathcal{I}} w_{ij} x_i(t),$$

$$x_j(t+1) = F_j(h_j(t+1))$$

with $\mathrm{W} = \{\mathrm{w_{ij}}\}$ denoting the weight-matrix, F_j a differentiable activation function and $x_i(t)$ the activity of neuron i at time t.

The dynamic behaviour of a time discrete FRNN up to time t can be described equivalently by a multi-layer-perceptron (MLP) with t layers. Each layer of the MLP consists of a copy of the FRNN neurons and the weights in between the layers are the same. For calculating the weights w_{ij} a gradient descent algorithm called back-propagation through time (BPTT) [3] can be applied. The objective during weight training is to minimize the sum $\varepsilon(t_a,t_e)$ of the quadratic errors over the time period $(t_a, t_e]$, given by

$$\varepsilon(t_a, t_e) = \sum_{t=t_a+1}^{t_e} E(t) = \sum_{t=t_a+1}^{t_e} \frac{1}{2} \sum_{k \in \mathcal{O}} (z_k(t) - x_k(t))^2$$

where $z_k(t)$ denotes the desired output function.

The weights are changed in direction of the error-gradient according to

$$\begin{aligned} \Delta w_{ij} &= -\eta \frac{\partial \varepsilon(t_a, t_e)}{\partial w_{ij}} \quad ; \quad \eta > 0 \\ &= \eta \delta_j(t) x_i(t-1) \end{aligned} .$$

0-7803-2431-5/95 $4.00 © 1995 IEEE

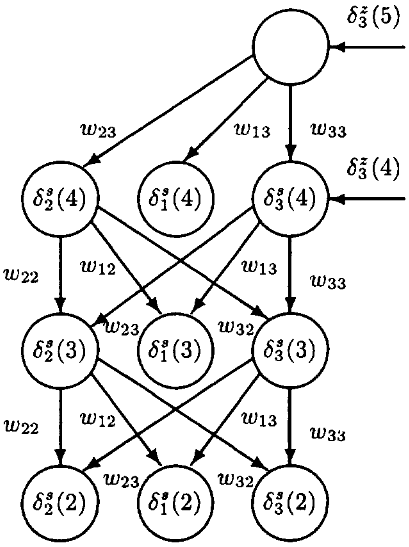

Figure 1: Truncated BPTT with $t_B = 5, t_v = 2, t_r = 3$ in a FRNN with 1 input neuron, 1 hidden neuron, and 1 output neuron

Starting at time $t = t_e$, the actual weight change can be calculated recursively for all times $t < t_e$. Applying BPTT iteratively leads to a decrease of the network-error $\varepsilon(t_a, t_e)$ till a minimum is reached.

Considering an efficient implementation of BPTT the gradient calculation is limited to a time-period t_r and BPTT is initiated only once after every $t_v < t_r$ time steps. In this case $\delta_j(t)$ is calculated according to

$$\delta_j(t) = \begin{cases} e_j(t)F_j'(h_j(t)) & ; t = t_B \\ e_j(t)F_j'(h_j(t)) + \sum\limits_{k \in \mathcal{U} \cup \mathcal{O}} w_{jk}\delta_k(t+1)F_j'(h_j(t)) & \\ & ; t_B - t_v < t < t_B \\ \sum\limits_{k \in \mathcal{U} \cup \mathcal{O}} w_{jk}\delta_k(t+1)F_j'(h_j(t)); t_B - t_r < t \leq t_B - t_v & \end{cases}$$

where t_B denotes the time when an error back-propagation is initiated and the error $e_j(t)$ is defined by

$$e_j(t) = \begin{cases} z_j(t) - x_j(t) & j \in \mathcal{O} \\ 0 & j \notin \mathcal{O} \end{cases}$$

Figure 1 illustrates the error back-propagation in this so-called truncated BPTT [4]. The gradient resulting in truncated BPTT is an approximation to the real gradient. The discrepancy depends on the parameters t_r and t_v. Therefore, t_r and t_v have to be optimized empirically for each task.

In simulation experiments the network size and the parameters η, t_v, t_r were optimized for telephone bandlimited speech recognition. For this task FRNN consisting of 52 input neurons, about 160 hidden neurons, and 11 or 23 output neurons, depending on the vocabulary, were used. The feature vectors were processed in sets of 4 vectors of dimension 13 in order to reduce the processing time. It was found that a fast and reliable weight training can be done by linearly decreasing η from 0.05 to 0.0 during 300 training epochs. Error back-propagation initiated every $t_v = 30$ time steps and considering the $t_r = 40$ preceding time steps turned out to be sufficient for learning of the relevant time dependences.

3. SPEECH DATA

In simulation experiments FRNN-SRS were realized for the recognition of isolated words as well as of connected digits. The speech signals were transmitted over telephone line and sampled with 8kHz. From these signals feature vectors were extracted every 12ms, each consisting of 12 cepstral coefficients (Cep) derived from LPC parameters. Additionally, a parameter characterizing the short-time energy of the signal was used. In order to immunize the feature vectors against telephone channel distortion the mean of each feature component measured during an utterance was subtracted.

In the case of isolated word recognition the system vocabulary consisted of the 11 German digits including the word *zwo* and 12 command words for telephone services. For computing of the SRS weight parameters, feature vectors from 100 utterances of each word, spoken by different male and female speakers, were used. Speaker independent recognition rates were measured on a disjunct set containing 100 utterances of each word by speakers not included in the training set. In the case of connected digit recognition the system vocabulary consists only of the 11 German digits from which a data corpus of 50 different digit strings were defined. The digit strings were composed of 3 digits and were arranged in order to represent a hard task for a SRS, e.g. in 13 strings a digit is followed by itself.

Each digit string was uttered by 120 different male and female speakers. For training of the network parameters the utterances of 60 speakers were used while the performance of the SRS were measured at the set of utterances spoken by the 60 different speakers not used for training.

4. RECOGNITION EXPERIMENTS

4.1. Isolated Word Recognition

In the case of isolated word recognition FRNN are trained to estimate word likelihoods for groups of 4 consecutive feature vectors. Word hypotheses were generated by accumulating the likelihood values during the duration of an utterance. The FRNN-SRS achieved a recognition rate of R_{23}=97.1% for 23 words and R_{11}=98.2% for the digits. The FRNN-SRS outperforms SRS based on Discrete HMM (DHMM) or Continuous HMM (CHMM) using the same features. SRS based on DHMM, which uses Delta-Cep (DCep) representing dynamical information of the feature vectors explicitly, achieves with R_{23}=97.3% about the same recognition rate as the FRNN-SRS [5]. This indicates that FRNN-SRS are able to exploit automatically the information about the dynamic of the feature vectors.

4.2. Connected Digit Recognition

For the task of recognizing connected digits containing an unknown number of digits the networks were also trained to estimate word likelihoods. A digit i in a digit string is recognized if the condition

$$O_i(t) > \Theta_b \wedge O_i(t+\Delta t) < \Theta_e \quad , \Delta t \geq 2$$

is true for the activity of the corresponding output neuron O_i. The threshold Θ_b indicates the begin of a digit while Θ_e indicates the end. Via Δt a minimum word duration constraint is introduced.

First recognition experiments were done with a FRNN-SRS trained for recognizing isolated words. Only 19% of the digit strings were recognized correctly. The correctness of the single digits was 55%. The low recognition rate, compared to 98% obtained with digits spoken in isolation, is due to the drastically increased dynamics of continuously spoken digits.

In further experiments the network was trained with feature vectors extracted from the digit strings, in order to adapt the network parameters to these dynamics. The FRNN-SRS trained with continuous speech achieved a recognition rate of 75% for strings and a digit correctness of 94.4%. The rather low string recognition rate is due to word deletions which occured in 23% of the test strings.

Interestingly, the FRNN-SRS trained with digit strings recognized the isolated digit data base with an accuracy of 96.4%. This indicates that isolated digit recognition is indeed a subtask of connected digit recognition for the FRNN.

In Figure 2 the recognition process is illustrated for two digit strings whose waveforms are shown in Figure 2a. As can be seen from Figure 2b the exploitation of activities of the output neurons by the above formulated condition results in the deletion of the the word *sechs*. In order to avoid this effect, an additional ouput neuron for detecting the boundary between two digits was added to the FRNN-SRS. The derivation of the activity of this neuron is multiplied with the activities of the neurons representing the digits. If the activity of the boundary neuron, which is shown in Figure 2c, is incorporated in the digit scoring Figure 2d results and a correct segmentation of the digit string is possible. The FRNN-SRS achieves in this case a recognition rate for connected digits of 81% and on the word level an accuracy of 93.2% and a correctness of 94.5%.

The analysis of the performance of the word boundary neuron revealed that in case of a hard decision 88% of the boundaries were correct. Most of the misclassifications occuring on the digit string level are still due to a high amount of deleted digits.

In order to test the recognition performance of FRNN-SRS independently of word detection defects, a SRS based on CHMM was used to segment the digit strings for recognition with the FRNN-SRS. This simulates the task of recognizing digit strings with a known number of digits. Summing the FRNN scores in the segments belonging to a digit and classifying the unknown digit according to the neuron with maximum activity resulted in a string recognition rate of 94.2%. This rate is slightly higher than that of the CHMM-SRS which shows that the discriminative properties of a FRNN lead to an increased recognition performance.

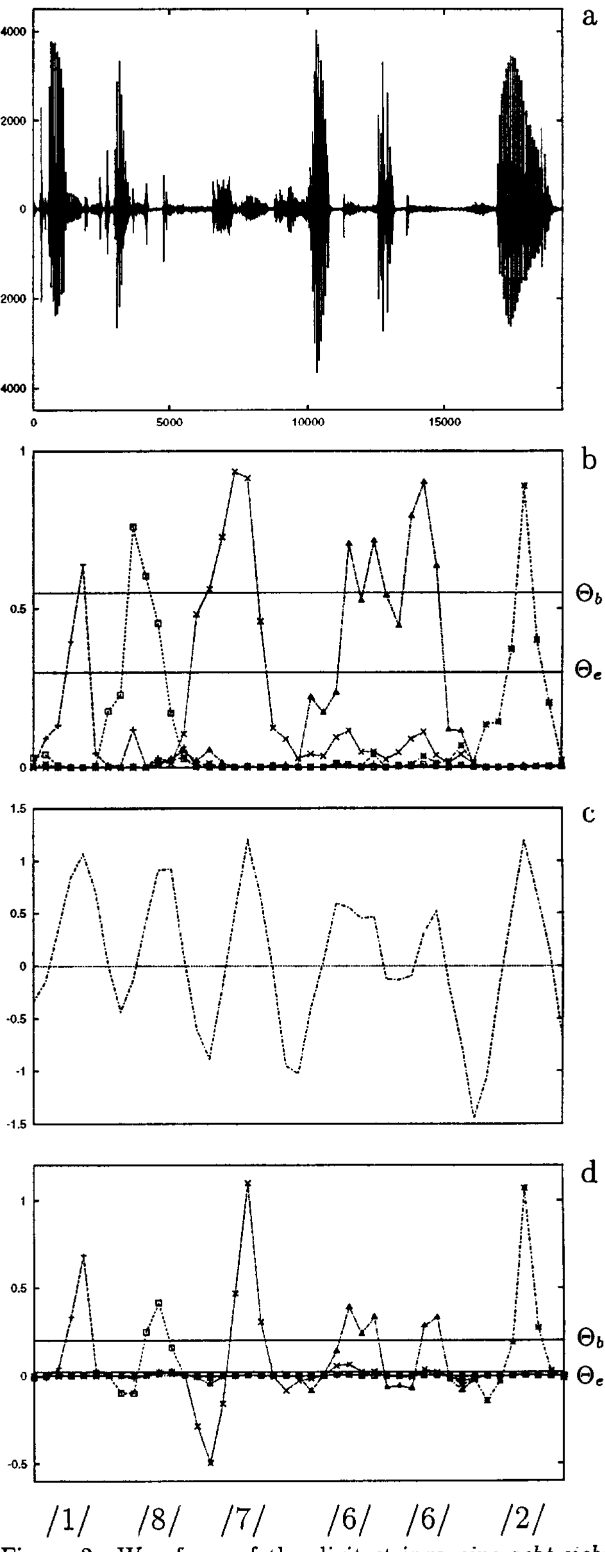

Figure 2: Waveform of the digit strings *eins-acht-sieben, sechs-sechs-zwei* (a), activities of the output neurons (b), activity of the word boundary neuron (c), and combined digit scores (d)

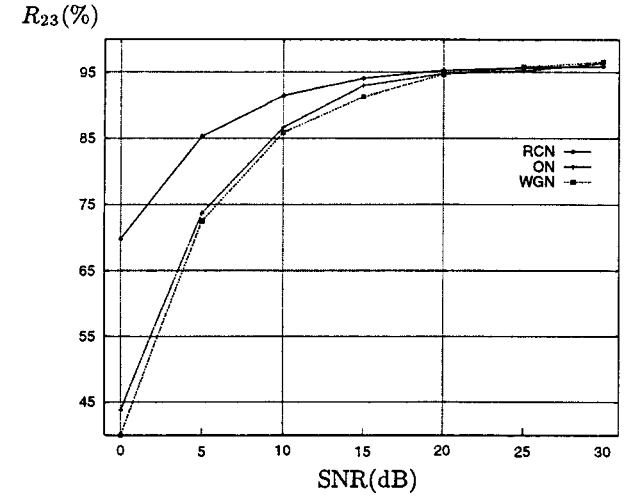

Figure 3: Recognition rates (R_{23}) of a noise immunized FRNN-SRS for speech signals contaminated with WGN, ON, and RCN at SNR levels ranging from 0dB to 30dB

4.3. Robustness of FRNN-SRS

In order to assess the robustness of FRNN-SRS, the recognition of utterances distorted by additive white gaussian noise (WGN), by office noise (ON), and by running car noise (RCN) recorded inside the cabin of a running car were considered. The noise was added to a speech signal such that a signal-to-noise ratio (SNR) between 0dB and 30dB for each word resulted. It has been reported [6, 7] that SRS based on DHMM or on MLP yield good recognition results if the model parameters were trained with speech corrupted by the same type of noise as the test signals.

Here we investigated the immunization capability of FRNN-SRS. While the recognition performance of a FRNN-SRS trained with clean speech dropped dramatically for noisy speech, the noise-adapted FRNN-SRS showed significantly improved recognition results for noisy speech while the recognition rates for clean speech decreased only slightly. Furthermore, we investigated whether a single FRNN-SRS could be immunized against all three noise types at different SNR levels simultanously. For this task the network was trained with feature vectors derived from noise-free and from RCN, WGN, and ON contaminated speech signals with SNR-values of 0dB, 10dB, and 20dB. In order to test the generalization of the extracted information, recognition rates were also measured on vectors derived from speech signals of the test sequence which were contaminated with noise at SNR-levels not used for training. As can be seen in Figure 3, the multi noise immunized FRNN-SRS achieves significantly improved recognition results for all noise types and SNR levels. At a high SNR above 20dB the recognition rates of the immunized FRNN-SRS are comparable to the rate obtained with clean speech independent of the noise type. In the range between 20dB and 10dB the decrease of the recognition rates are different for the noise types but still above 85%. For SNR values below 10dB the robustness of the immunized FRNN-SRS is not sufficient. A comparison of these results with that reported in [6] show that the immunized FRNN-SRS outperforms the DHMM-SRS.

5. CONCLUSIONS AND PERSPECTIVES

With FRNN it is possible to develop a speech recognizer for speaker independent recognition of telephone command words and isolated as well as connected digits which can be efficiently implemented.

The results of the simulation experiments demonstrate that FRNN are able to extract information relevant for speech recognition from noise contaminated speech and thus achieve a robust recognition performance. While the recognition of isolated words can be performed with very high recognition rate, the recognition of connected digits is a rather hard task for the FRNN-SRS. The main problem is to avoid word deletions in the task of recognizing digit strings with unknown length. By introducing a word boundary neuron the amount of deletions decreased but still remains the main source of incorrectly recognized digit strings.

Current research is focused on a more sophisticated exploitation of the information delivered by the word boundary neuron. Furthermore, it will be investigated whether the feature extraction of speech signals can also be incorporated in the FRNN.

This work is supported by the Deutsche Forschungsgemeinschaft in the research program "System- und Schaltungstechnik für hochgradige Parallelverarbeitung".

6. References

[1] Waibel,A., Hanazawa,T., Hinton,G., Shiano,K., and Lang,K., "Phoneme Recognition using Time-Delay Neural Networks", IEEE Trans. on Acoust., Speech, and Signal Processing, vol.37(3) 1989, pp. 328-339.

[2] Kasper,K., Reininger,H., and Wolf,D., "Phoneme Based Isolated Word Recognition Using Neural Networks for Prediction and Classification", Proc. EUSIPCO-92, Brussels, pp. 427-430.

[3] Werbos,J.P., "Backpropagation Through Time: What It Does and How to Do It", Proc. IEEE, vol.78 no.10, pp.1550-1560, 1990.

[4] Williams,R.J., and Zipser, D., "An Learning Algorithm for Continually Running Fully Recurrent Neural Networks", Neural Computation 1, pp. 271-280, 1989.

[5] Kasper,K., Reininger,H., Wolf,D., Wüst, H., "A Monolithic Speech Recognizer Based on Fully Recurrent Neural Networks", Neural Networks for Signal Processing IV, Ermioni 1994, pp. 335-344.

[6] Nicol,N., et.al., "Improving the Robustness of Automatic Speech Recognizers Using State Duration Information", Proc. Speech Processing in Adverse Conditions, Cannes 1992, pp. 183-186.

[7] Sankar,R., and Patravali, S., "Noise Immunization Using Neural Net for Speech Recognition", Proc. IEEE Int. Conf. on ASSP, Adelaide 1994, vol.2, pp. 685-688.