Modular Neural Networks for Low-Complex Phoneme Recognition

Axel Glaeser

Ascom AG, Applicable Research & Technology Unit,
CH – 5506 Mägenwil, Switzerland
Email: Axel.Glaeser@ascom.ch
URL: http://www.ascom.ch/systec/ART

ABSTRACT

We present a Modular Neural Network (MNN) for phoneme recognition within the framework of a hybrid system (neural networks and HMMs) for speakerindependent single word recognition. With this approach, we are taking the computational effort into account which is used as an additional criterion for assessing the system performance.

The main idea of the proposed MNN is the distribution of the complexity for the phoneme classification task on a set of modules. Each of these modules is a single neural network which is characterized by its high degree of specialization. The number of interfaces, and therewith the possibilities for infiltering external acoustic-phonetic knowledge, increases for a modular architecture.

Moreover, after the development of a suitable topology for the MNN, each of the modules can be optimized for its specific phoneme recognition task. This is done by detecting and pruning irrelevant input parameters and leads to a more efficient system in terms of memory and computational requirements.

1. INTRODUCTION AND MOTIVATION

The dynamic nature of speech and its complex structure in both, the spectral and temporal domain, is the main problem in the field of speaker-independent speech recognition. Neural networks are especially appropriate to deal with those spectral variations, but their ability in processing the temporal structure of speech seems to be limited to short-time levels [1,2]. Therefore, the development of a hybrid system became a popular approach (e.g. [3]). It makes use of the advantages of neural networks for the phoneme recognition and the benefits of HMMs to solve the temporal alignment problem at the word and language level.

Our system consists of three major parts: A preprocessing unit for the feature extraction of the incoming signal, followed by the phoneme recognition system which is implemented as a Modular Neural Network (MNN). The final part of the system architecture is a word-recognizer for isolated, speaker-independent word recognition. The complexity of the HMM-structure can be made relatively small by using discrete HMMs with a fixed number of states.

Most of the recent approaches with neural networks for speech recognition tasks make no use of incorporating acoustic-phonetic knowledge into the system. The number of interfaces, and therewith the possibilities for infiltering this knowledge, increases for the proposed modular architecture. Moreover, the overall-complexity of the speech decision task can be distributed to a subset of modules and each of them can be optimized for the specific task. As a result of this modular concept, the determination of relevant input parameters can be done selectively for each module.

Our simulations came to two main conclusions: On the one hand, the use of MNNs can improve appreciably the overall recognition rate of the system in comparison with a single-network recognizer. On the other hand, it showed that the computational effort of the proposed approach can be reduced.

The development and optimization of the modular system is composed of two steps. First of all, the topology of the system has to be adjusted to the recognition task. In a second step, the single units of the modular structure are optimized for receiving satisfactory recognition results. Both steps will be discussed in detail in the following two sections.

2. TOPOLOGY OF THE PROPOSED MODULAR RECOGNITION SYSTEM

The starting-point for the construction of an advanced system is the simulation of a single neural network for the recognition of all phonemes. After this initial simulation, the outcoming confusion matrix is analysed. It has the desired output on the one axis and the output of the network on the other one. It is a remarkable feature of the confusion matrix, that the values of its elements are skewly distributed. Therefore, phoneme groups can be found which have mainly inner-group confusions. The filling of these phoneme groups does not necessarily has a phonetic background. The new approach of constructing a modular recognition system was motivated by these results.

The algorithm for its development can be divided into three steps:

1. First of all, the phoneme classes without essential confusion with other classes are separated in a first module. This is a neural network with a relatively moderate complexity which is simulated and trained on this specific classification task.

2. After the initial differentiation in main classes, subsequent modules are trained. Each of them has to deal with only a fraction of the original variation of phonemes. But the complexity of the recognition task of a following module is

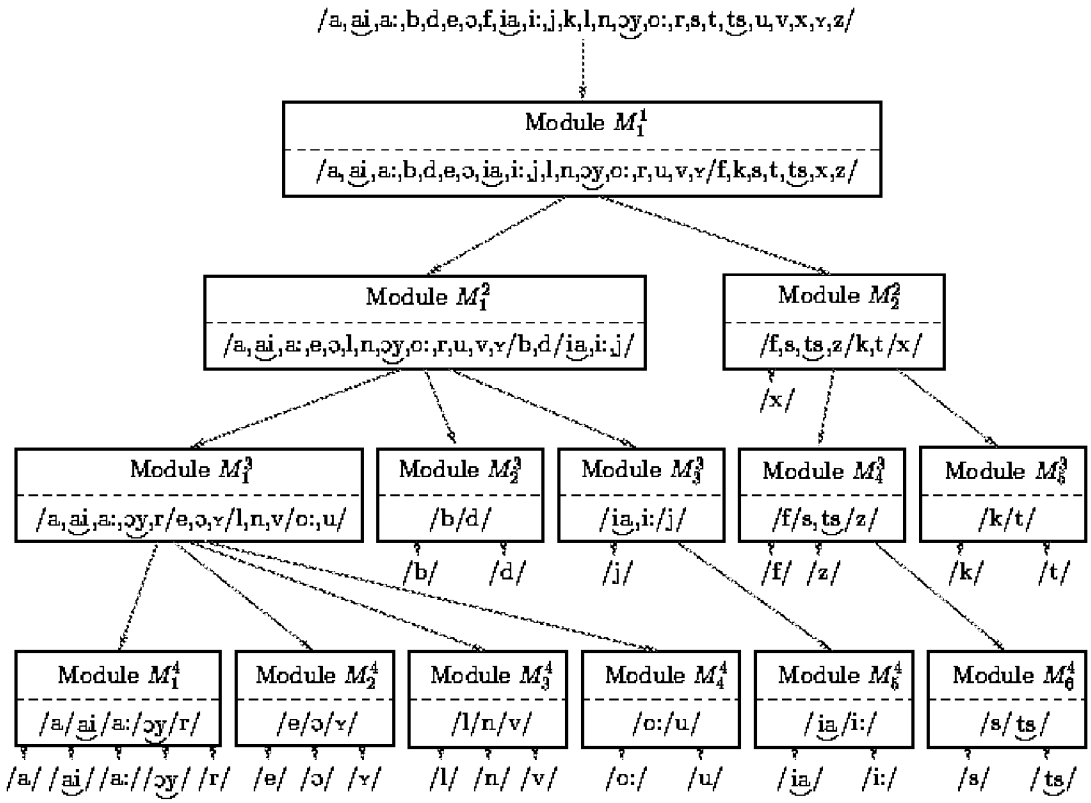

Figure 1: Topology of the proposed Modular Neural Network (MNN) for 25 german phonemes.

mainly constant because of the increasing degree of confusion inside the remaining phonemes groups.

3. The described process of building phoneme classes and training the required neural networks is a recursive procedure which ends when the outcoming classes consist only of single phonemes.

This approach leads to a system which is characterized by a tree-structured topology (see figure 1).

Beginning with the module in the first layer, each frame of a phoneme is put forward through a maximum of one module per layer until the frame is related to a single phoneme instead of a phoneme group. Obviously, the main drawback of this topology is the inability of the system to correct a unique error done in a preceding unit. Beyond that, the final recognition propability of a phoneme is the multiplication of the probabilities of the modules it has passed. In order to minimize these errors, we experimented unsuccessfully with recurrent networks with different implementations for error detection and correction algorithms.

3. IMPROVED RECOGNITION USING RELEVANT INPUT PARAMETERS

This chapter describes the types of neural networks utilized in the proposed system. Moreover, an algorithm is presented which determines the relevance of input parameters. This procedure leads to a Compact Modular Neural Network (CMNN) for phoneme recognition.

In principle, all variations of neural networks can be combined in the presented modular system because the training algorithm is restricted only to each module instead of the whole system. We are using conventionel Multilayer Perceptrons (MLPs) and Time Delay Neural Networks (TDNNs) [4] with Backpropagation-Algorithm for the learning phase.

As mentioned in the previous chapter, every module in our speech recognition system has a moderate complexity, compared with the case of a single, non-modular recognizer. Our examinations showed that these comparatively low-complex neural networks are especially appropriate for the determination of the relevance of the input parameters. Afterwards, the input vector size can be pruned iteratively, based on the estimated relevance of its elements.

We compared different existing algorithms for pruning nodes in neural networks. All of them are limited on the hidden nodes

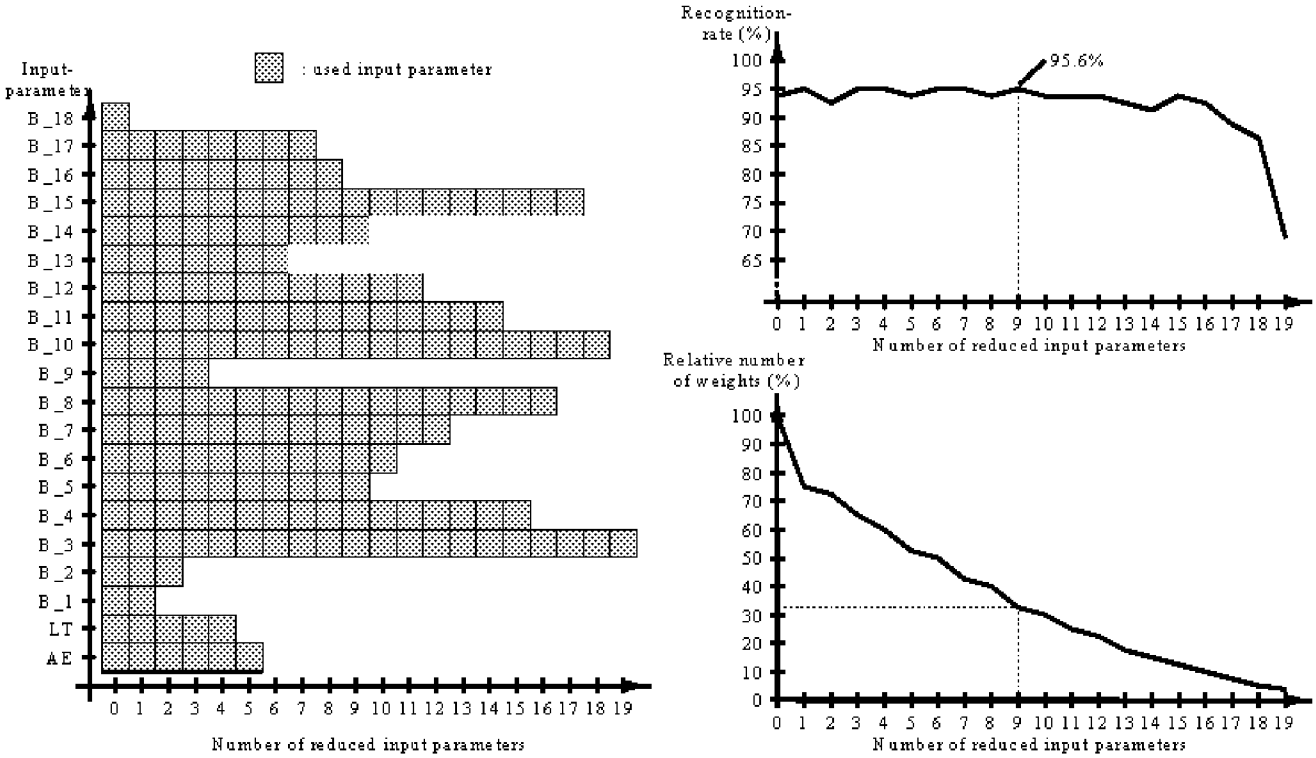

Figure 2: This example demonstrates the detection and pruning of irrelevant input parameters. Moreover, the influence on the recognition rate and the number of required weights is shown.

of layered neural networks because in conventionel approaches, the dimensions of input and output layers remain constant. So, the algorithms were adapted in a way that they can deal even with input nodes.

We achieved the best results by calculating and implementing an adapted version of a pruning algorithm, introduced by B. Appoloni and G. Ronchini [5]. It starts with the calculation of the mean value and standard deviation of all weights connected with the first input node. This procedure is done for all input nodes. With these results, we can finally draw up a ranking list of all input nodes depending on the characteristics of their corresponding weights. Assuming that all input parameters are normalized on the same range of values, the calculated *sensitivity* for the input nodes is proportional to the relevance of their associated input parameters for a specific recognition task.

The example in figure 2 shows this novel approach for the optimization of the first module of the MNN. It does a voiced/unvoiced-classification for 25 german phonemes using a TDNN. At the beginning, a neural network with 20 input parameters was simulated. (These parameters are: absolute energy (AE), ratio of energies (LT) and 18 melscaled spectral coefficients from 0 .. 4kHz). After the initial simulation, the input parameter with the smallest value of sensitivity is pruned and a new simulation with only 19 remaining input parameters is started. This iterative process is continued until a single input parameter is left.

For the special case of a voiced/unvoiced recognition, our acoustic-phonetic knowledge tells us, that the lower frequency bands are more relevant than the upper ones and that is what the described algorithm also detects (grey fields in the left diagram in figure 2). But the proposed pruning method also works for modules where the phonetic background of the classification task is not as obvious as in the example mentioned before.

The relation between the computational effort during the forward phase (proportional to the number of weights) and the recognition rate is documented in the other two diagrams of figure 2. It is an essential result of our examinations that the number of weights can be reduced clearly (77% in the given example) without any loss of recognition accuracy.

The described way of detecting and pruning irrelevant input parameters was applied separately to each of the 14 modules of the MNN. This proceeding forms the novel Compact Modular Neural Network (CMNN) with the same tree-structured topology as the MNN but with considerably improved performance in terms of computational effort and memory requirements. This qualitative statement will be extended in a quantitative way in the next chapters.

4. SPEECH DATABASE

The speech data were extracted from the database TELEROM 1 of the Forschungsinstitut der Deutschen Bundespost

Telekom. It contains data from 100 speakers (female and male). The preprocessing is characterized by a sampling frequency of 8 kHz, resolution of 12 bit, frame length of 16ms, frame shift of 10ms and the use of a Hamming window. As mentioned before, each frame is represented by a 20-element input vector containing mainly melscaled spectral coefficients.

The phonemes of the speech data were labeled and were divided into three independent classes:

- *Training data:* They are used for training the neural networks. The synaptic weights of the modules are changed only in order to minimize the resulting square error of the training data.
- *Testing data:* The quality of trained neural networks and their acceptance for the modular system is based on the results obtained by these data.
- *Reference data:* Both, training data and testing data, have a direct influence on the topology and internal structure of the system and the modules. So, only this third speech data class represents really unknown speakers for testing and assessing the final recognition system.

5. RESULTS

An experimental word recognition system based on the described CMNN with an additional HMM for word recognition was implemented. Due to the fact, that the vocabulary of the speech database is quite limited, the output of the word recognition system consists of only 13 german words which requires 25 different phonemes in the CMNN. We obtained the following performance (tables 1 and 2).

	non-modular	MNN	CMNN
training data	74.4%	89.2%	84.5%
test data	69.2%	82.5%	81.3%
reference data	71.2%	80.4%	79.5%
number of weights (memory requirements)	8605	24512	12526
average number of weights used per frame (computational effort)	8605	7360	4598

Table 1: System performance I: Phoneme recognition results and the influence on the memory and computational effort.

	non-modular	MNN	CMNN
training data	91.3%	92.5%	93.8%
test data	87.6%	90.6%	91.2%
reference data	79.1%	84.7%	86.9%

Table 2: System performance II: Word recognition results.

The recognition rates for the modular system increase especially for the test and reference data which can be explained as a better generalization of the MNN and CMNN.

In addition to our simulations, we developed a stand-alone signal-processor system based on a single ADSP-21020-processor where the proposed word recognition system, including the CMNN is implemented. The fact, that the system works in real time with a reliable recognition performance, justifies our statement that a CMNN is appropriate to improve the recognition rate for speaker-independent phoneme recognition and, at the same time, it reduces the necessary effort for simulating the system after the initial learning phase.

6. REFERENCES

1. Lippmann, R.P. "Review of Neural Networks for Speech Recognition", *Neural Computation 1, 1989.*
2. Morgan, D. and Scofield C. "Neural Networks and Speech Processing", *Kluwer Academic Publishers, 1991.*
3. Devillers, L. and Dugast, C. "Hybrid system combining expert-TDNNs and HMMs for continous speech recognition", *ICASSP 1994, p. 165-168.*
4. Waibel, A. et al. "Phoneme recognition using time-delay neural networks", *IEEE ASSP, Vol. 37, No. 3, 1989.*
5. Apolloni, B. and Ronchini, G. "Dynamic sizing of multilayer perceptrons", *Biologic Cybernetics, Vol. 71, p. 49-63, 1994.*