NEURAL NETWORK APPROACHES TO SPEECH RECOGNITION: A GENERAL RADIAL BASIS FUNCTION NETWORK FOR SPEAKER-INDEPENDENT PHONE CLASSIFICATION

D. R. Dersch
Speech Technology Research Group
Department of Electrical Engineering
Building J03, The University of Sydney
NSW, 2006, Australia
Phone +61-2-93514509
email: dersch@speech.su.oz.au

ABSTRACT – In this paper we present neural network approaches which enable both the analysis of a high dimensional data space of phone templates and the construction of a speaker-independent isolated phone classifier based on a Generalized Radial Basis Function Network (GRBFN). Firstly, we present a codebook obtained by a neural motivated fuzzy Vector Quantization procedure. Such codebooks provide an intrinsic discretization of the data space expanded by phone templates into various phone groups, e.g. stops, fricatives, nasals and vowels of different pitch. Secondly, a codebook is used to train a three-layer GRBFN in a two-step optimization process. As a result we obtain a speaker-independent single phone recognition accuracy of 63.1% on the training set and 62.2% on the test set for 52 different phone classes. A coarse classification of five phone groups into 'stops', 'fricatives', 'nasals', 'semi-vowels' and 'vowels' yields a recognition accuracy of 87.6% on the training set and 87.0% on the test set, respectively. Phone templates are obtained from the male training corpus of the TIMIT database.

1. INTRODUCTION

In the field of automatic speech processing, neural network approaches for data analysis and classification, e.g., neural motivated Vector Quantization (VQ) methods and Radial Basis Function classifiers, have been shown to be useful tools. Both techniques and their relationship are reviewed in the following two subsections.

1.a Neural Vector Quantization procedures

Vector Quantization techniques serve to map a data space $X \subset \mathcal{R}^n$ characterized by an *a priori* probability distribution $P(\mathbf{x})$, $\mathbf{x} \in X$, onto a finite set of so-called prototypical codebook vectors $\mathbf{w}_r \in W \equiv \{\mathbf{w}_r \in \mathcal{R}^n \mid r = 1, \ldots, N\}$. The mapping is defined by a mathematical prescription which associates the data vectors $\mathbf{x} \in X$ to the codebook vectors $\mathbf{w}_r \in W$.
In the case of *hard clustering* a data vector $\mathbf{x}$ is associated to only *one* prototype $\mathbf{w}_{r'}$ which is *next* to $\mathbf{x}$. This prototype $\mathbf{w}_{r'}$ is called the "winner". The Max-Lloyd algorithm (Max, 1960 and Lloyd, 1982) implements hard clustering. In the sequential version of this algorithm a "winner takes all" learning rule performs a gradient descent on a quadratic error function. However, for non trivial data distributions the minimization of such an error function belongs to the class of NP-complete optimization problems. As is generally the case for such optimization problems, gradient descent methods are unable to find optimal or even "good" solutions, because those solutions strongly depend on the choice of the initial conditions. A second weakness of these schemes is the slow convergence due to the "winner takes all" strategy.
One tries to overcome these drawbacks by the so-called *soft competing* VQ procedures. These algorithms are characterized by *cooperative-competitive* learning rules of the type

$$\mathbf{w}_r(t+1) = \mathbf{w}_r(t) + \epsilon\, a_r[\mathbf{x}(t); W(t), \kappa]\,[\mathbf{x}(t) - \mathbf{w}_r(t)]. \tag{1}$$

Here, a data vector $\mathbf{x}$ is randomly selected according to the probability distribution $P(\mathbf{x})$. Each adaption step is scaled by a small parameter ϵ and a so-called *cooperativity function* $a_r \geq 0$.

$a_r(\mathbf{x}; W, \kappa)$ is a function of the presented data vector $\mathbf{x}(t)$ and depends parametrically on the state of the codebook $W(t)$ and the value $\kappa(t)$ of the so-called *cooperativity parameter* $\kappa \geq 0$. In the case of $\kappa \equiv 0$ the cooperativity function a_r is non vanishing only for the "winner". Here, hard clustering and a corresponding "winner takes all" strategy is recovered. Whereas for a non vanishing value of κ the "winner" *and* a set of codebook vectors are adapted. In a simulated annealing process the cooperativity parameter κ is reduced from a large value to a small or vanishing value.
A number of neural motivated VQ procedures can be considered within this framework, e.g., the Kohonenalgorithm (Kohonen, 1982), the "Neural–gas" algorithm (Martinetz and Schulten, 1991), and a fuzzy VQ scheme derived from a minimal free energy criterion (Rose et al., 1990). Soft competing VQ procedures entail better results and faster convergence than the classical hard clustering procedure. In the following we want to focus on a specific soft competing algorithm: the fuzzy VQ scheme. Here, the cooperativity function

$$a_r(\mathbf{x}; W, \kappa \equiv \rho) = \frac{\exp -(\|\mathbf{w}_r - \mathbf{x}\|^2/2\rho^2)}{\mathcal{Z}} \tag{2}$$

is given by globally normalized Gaussians of width ρ. $\mathcal{Z} = \sum_r \exp -(\|\mathbf{w}_r - \mathbf{x}\|^2/2\rho^2)$ is the normalization factor and ρ is the cooperativity parameter of this model. The so-called "fuzzy range" $\rho = \kappa$ defines a length scale in data space. Due to the normalization $\sum_r a_r = 1$ and the condition $a_r \geq 0$ the cooperativity function is considered to be a conditional probability $a_r(\mathbf{x}) \equiv p(\mathbf{r}|\mathbf{x})$ of associating a given data vector $\mathbf{x}$ to codebook vector $\mathbf{r}$.
The fuzzy VQ scheme has been thoroughly studied (Rose et al., 1990, Dersch, 1996 and Dersch and Tavan, 1994, 1996). The Learning rule (1) with a_r given by (2) describes a stochastic gradient descent on a error function which is a *free energy* in a mean–field approximation whose complexity strongly increases with a decreasing value of the fuzzy range ρ.
At large values of the fuzzy range ρ one finds a completely degenerated codebook. All codebook vectors are located at the first moment of the data space and as we say there exists only *one cluster center*. At each value of the fuzzy range one finds a distinct number of cluster centers who's positions and degree of degeneration are uniquely determined by the minimality requirement of the free energy. A corresponding annealing process entails a *hierarchical, fuzzy discretization* of a data space on multiple length scales. The analysis of codebooks on different length scales defined by the fuzzy range ρ provides insight into the structure of a data space on different resolution scales. It has been shown that codebooks obtained by this scheme are characterized by well defined properties (Dersch and Tavan, 1994). These properties can be exploited to control the learning process and enhance the quality of the resulting codebooks. In the remainder of this section we show how codebooks serve to construct a classifier in a Radial Basis Function architecture.

1.b Generalized Radial Basis Function Network

Multi Layer Perceptrons (MLP), see Rosenblatt (1958), are well established classifier systems covering applications in the fields of pattern recognition and classification. Classification addresses the problem of assigning a data vector $\mathbf{x}_r \in X \equiv \{\mathbf{x}_r \in \mathcal{R}^n \mid r = 1, \ldots, N\}$ to a given class $\{\mathbf{x}_r \rightarrow c_r,$ with $c_r \in [1, \ldots, m] \mid r = 1, \ldots, N\}$. A MLP used for a classification task maps an input vector $\mathbf{x}_r \in \mathcal{R}^n$ onto an output vector $\mathbf{y}(\mathbf{x}_r) \in \mathcal{R}^m$, with $\mathbf{y}_i(\mathbf{x}_r) = 1$ for $(i = c_r)$ and $\mathbf{y}_i(\mathbf{x}_r) = 0$ for $(i \neq c_r)$.
A commonly used training algorithm for MLP is the Backpropagation algorithm (Rumelhart and McClelland, 1986). Moody and Darken (1989) proposed a promising alternative to train a three layer network. The authors propose to train a network in a two step process.
In the first step the input weights of the N-hidden neurons are obtained by a VQ procedure. Secondly, the output weights are calculated by minimizing the quadratic classification error

$$E = \langle \|\mathbf{y}(\mathbf{x}) - F\,\mathbf{A}(\mathbf{x})\|^2 \rangle_x. \tag{3}$$

Where $\mathbf{y}(\mathbf{x})$ is the desired output vector of the network for a given class membership of a data vector. F is the $(m \times N)$ matrix of the output weights, $\mathbf{A}(\mathbf{x})$ is the N-dimensional vector of the

hidden layer activity. Here, the hidden neurons are characterized by Gaussian activation functions $A_r(\mathbf{x}) = \exp -\|\mathbf{w}_r - \mathbf{x}\|^2/2\rho^2$ of width ρ centerd at $\mathbf{w}_r$. $\langle \ldots \rangle_x$ denotes the average over the data set. In our approach we are following the two step training scheme, but apply the presented fuzzy VQ scheme, rather than the Max-Lloyd algorithm as suggested by Moody and Darken (1989) for a number of reasons mentioned above. Further, we globally normalize the activation function of the hidden layer according to Equation (2), following a scheme proposed by Girosi and Poggio (1989), instead of using unnormalized Gaussians. In this approach the hidden nodes no longer show a radial response centered at $\mathbf{w}_r$. We therefore call our network *Generalized Radial Basis Function Network*. The architecture and training of such a network of locally tuned units show a strong biological motivation (Poggio, 1990 and Marr, 1969).
The remainder of this paper is organized as follows: In the next section we present the essential processing steps of the speech signals. The following sections show the results of the VQ process and present the training and test of the GRBF-Network for speaker-independent phone classification.

2. SPEECH DATA AND PREPROCESSING

Isolated phone templates are obtained from read speech of the phonetically transcribed TIMIT database (NIST, 1990). We subdivided the speech corpus into a subset of training and test data as proposed by NIST. Only male speakers are considered.
A series of preprocessing steps are performed on the sound pressure signal. At first, a short-term Fourier transformation is performed, using a $40msec$ Hamming window and a $10msec$ step size. Secondly, the Fourier coefficients are combined into 21 channels on a nonlinear bark scale. Finally, we compress the dynamic range of the resulting power spectrum by calculating the fourth root of each coefficient and truncate noise below a certain threshold. At the center of each phonetically labelled segment phone templates comprising 21 bark channels and 9 time frames are extracted. Thus, a isolated phone is represented by a 189-dimensional phone template. Each of these phone templates is normalized to zero mean and uniform contrast.

3. INTRINSIC PHONE CLASSES

In order to gain insight in the structure of the 189-dimensional space of phone templates we performed a hierarchical fuzzy VQ for the training set of 10^5 phone templates.

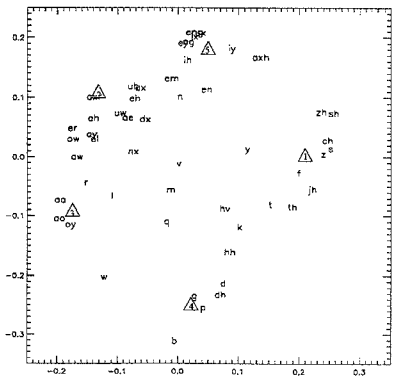

Figure 1: Multidimensional scaling of 52 prototypical phones and five cluster centers (triangles).

The corresponding annealing process undergoes a sequence of hierarchical discretizations of one, three and five cluster centers (Dersch, 1996). A closer investigation of the data space on the level of

five cluster centers, corresponding to a fuzzy range $\rho = 0.3$, is illustrated in Figure 1. Figure 1 shows the result of a reduction in dimension by multidimensional scaling (Mardia et al., 1979) of the five prototypes (triangles) and of 52 prototypical phones obtained by simple averaging over all templates of each class. Here, we made use of the labelling according to TIMIT. Figure 1 reveals a clustering of the mean values of various phone groups. Phones belonging to the phone group stops, fricatives and vowels are located at different cluster positions.

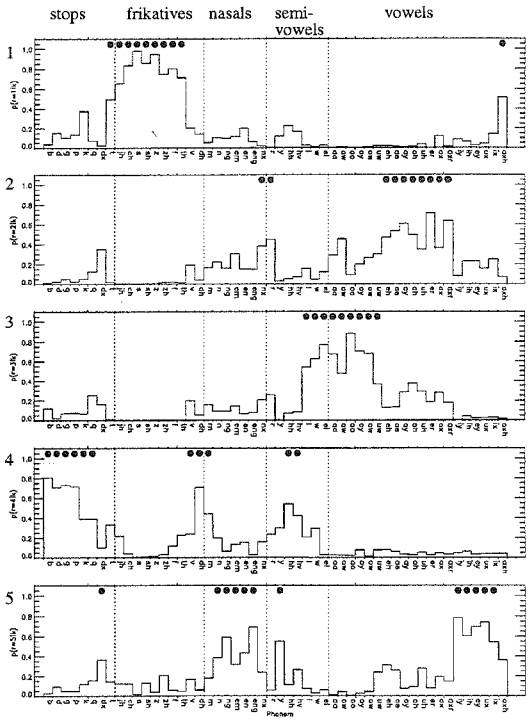

Figure 2: Conditional probabilities of associating 52 phoneme classes to the 5 prototypical templates shown in Figure 1

For a closer investigation of the data space we approximately calculate the *conditional probabilities* $p(\mathbf{r}|k)$ that template $\mathbf{x}_k$ from class k is associated to prototype $\mathbf{w}_r$ by the relative frequencies,

$$p(\mathbf{r}|k) = 1/N(k) \sum_{\{\mathbf{x} \in k\}} a_r(\mathbf{x}; W, \rho = 0.3).$$

Here, $N(k)$ is the total number of data vectors $\mathbf{x}$ belonging to phoneme class k. $a_r(\mathbf{x}; W, \kappa \equiv \rho)$ is given by (2).

Figure 2 shows the conditional probabilities for each prototype 1–5 as a function of the 52 phoneme classes. The horizontal lines indicate the borders between five phoneme groups: stops, fricatives, nasals, semivowels and vowels. The dots mark the prototype $\mathbf{r}$, where $p(\mathbf{r}|k)$ reaches the maximum value for a given phone class k. Figure 2 confirms the results of Figure 1. The five prototypes

provide a rough classification of the various phoneme groups. Fricatives are associated to prototype 1, whereas stops and nasals are mainly associated to prototype 4 and 5, respectively. Interestingly, semivowels and vowels are not associated to a certain prototype. But a closer investigation of Figure 2.2, 2.3 and 2.5 reveals a coarse seperation of semivowels and vowels into groups of low, middle and high pitched voice, respectively.

4. CLASSIFICATION RESULTS

To train the GRBFN, we first perform a hierarchical VQ procedure on the training set of 10^5 phone templates using a codebook of 700 codebook vectors. We then calculate the output weights connecting the hidden layer with the 52 output neurons by solving Equation (3). An optimal value of the width ρ is obtained by maximizing the overall recognition accuracy on the training set.

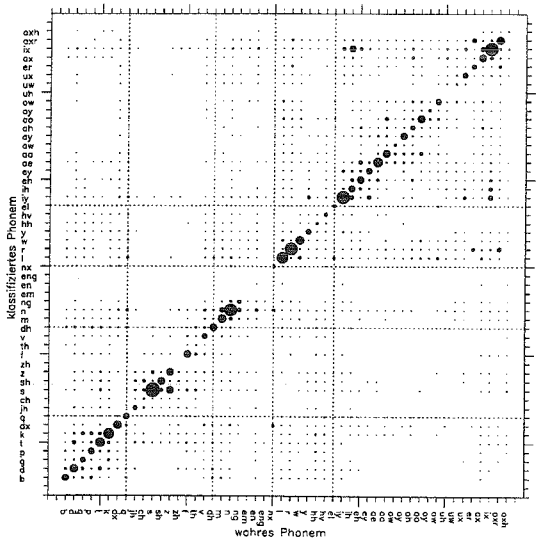

Figure 3: Classification results of the GRBF-Network. Confusion matrix for 52 phoneme classes.

The 52×52-confusion matrix in Figure 3 illustrates the performance of the network on the set of $3.5\,10^4$ test phones. The rows show the true membership to each phone class, whereas the lines show the membership classified by the network. The area of each dot in Figure 3 is proportional to the number of data vectors. The dashed horizontal and vertical lines mark the boarders between various phone groups. A perfect classifier should show up in a confusion matrix with a diagonal structure. However, due to misclassification there also exist off-diagonal elements. From the confusion matrix we calculate the overall performance for speaker-independent phone classification. Table 1 summarizes the results on the training and the test set.

classes	52	39	5 groups	top-3
train	63.1 %	67.2 %	87.6 %	87.6 %
test	62.2 %	66.5 %	87.0 %	86.7 %

Table 1: Classification results on the training and the test set.

We obtain a recognition accuracy of 63.1% on the training set and 62.2% on the test set for 52 different phone classes. Combining phone classes by a scheme first proposed by Lee and Hohn (1989)

we obtain for the remaining 39 phone classes a recognition accuracy of 67.2% on the training set and 66.5% on the test set. A coarse classification of five phone groups in 'stops', 'fricatives', 'nasals', 'semi-vowels' and 'vowels' yields a recognition accuracy of 87.6% on the training set and 87.0% on the test set, respectively. The last row in Table 1 shows the "top-3" recognition accuracy on the set of 52 different phone classes. Here, a phone is considered to be "correctly classified", if the output value of the true phone class is amongst the three largest values. The top-3 recognition accuracy is 87.6% on the training set and 86.7% on the test set. Note, the classifier has a very good generalization. The performance on the test set shows a negligible short-cut of less than one percent as compared to the training set. The presented speaker-independent single phone classifier clearly outperforms other neural network approaches (Anderson, 1989) and is reasonable as compared to classical Hidden Markov Model approaches (Lee and Hohn, 1989, Rathinavelu and Deng, 1996 and Ruxin and Jamieson, 1996).

Acknowledgments

The author is grateful to Prof. Paul Tavan, Dr. Axel Wismüller and Sebastian Albrecht for many useful discussions and Karsten Kumpf for constructive comments on this paper. The work was sponsored by the BMBF in the frame of the project "Spracherkennung in Neuronaler Architektur (SPINA)" grant no. 01 IN 108 C/8.

5. REFERENCES

Anderson T.R. (1994) *Auditory models with Kohonen SOFM and LVQ for speaker independent phoneme recognition*, Proceedings of IEEE International Conference Neural Networks ICNN 94, 4466–4469.

Dersch D. R.and Tavan P. (1994) *Control of annealing in minimal free energy vector quantization* Proceedings of the IEEE International Conference on Neural Networks ICNN 94, 698-703. Dersch D. R.and Tavan P. (1996) *Annealing in minimal free energy vector quantization* Proceedings of HELNET '94, 11-22, UV University press Amsterdam, 1996

D. R. Dersch. (1996) *Eigenschaften neuronaler Vektorquantisierer und ihre Anwendung in der Sprachverarbeitun*, Harri Deutsch, Frankfurt am Main.

Girosi F. and Poggio T. (1989) *Networks and the best approximation property*, A. I. Memo 1164, Massachusetts Institute of Technology 10.

Kohonen T. (1982) *Self-organized formation of topologically correct feature maps* Biol. Cybern, 43:59–69.

Lee K.F. and Hon H.W. (1989) *Speaker-independent phone recognition using hidden markov models*, IEEE Trans. ASSP, 37:1641–1648.

Lloyd S. P. (1982) *Least squares quantization in PCM* IEEE Trans. Inform. Theory, 28:129–137.

Mardia K. V., Kent J. T. and Bibby J. M. (1979) *Multivariate Analysis* Academic Press, London.

Marr D. (1969) *A theory of cerebellar cortex*, J. Physiol 202:437-470.

Martinetz T. M. and Schulten K. J. (1991) *A 'neural gas' network learns topologies* Proceedings of the International Conference on Artificial Neural Networks, ICANN 91, 397–402, Elsevier Science Publishers, Amsterdam, 1991.

Max J. (1960) *Quantizing for minimum distortion* IRE Trans. Inform. Theory, 3:7–12.

Moody J. and Darken C. (1989) *Fast learning in networks of locally-tuned processing units*, Neural Computation, 1:281–294.

NIST (1990) Nat. Inst. of Stand. & Tech. *TIMIT acoustic-phonetic continuous speech corpus* NIST Speech Disc 1-1.1, Oct. 1990.

Poggio T. (1990) *A theory of how the brain might work*, Cold Spring Harbor Laboratory Press, 899–910.

Rathinavelu C. and Deng, L. (1996) *HMM-Based Speech Recognition Using State-Dependent, Discriminatively Derived Transforms on Mel-Warped DFT Features*, Proceedings ICASSP 96, 9-12

Rose K., Gurewitz E., and Fox G. (1990) *Statistical mechanics and phase transitions in clustering*, Phys. Rev. Lett., 65:945-948.

Rosenblatt F. (1958) *The Perceptron: The probabilistic model for information storage and organization in the brain*, Psychological Review, 65:386–408.

Rumelhart D. E. and McClelland J. L. (1986) *Learning internal representations by error propagation*, Parallel Distributed Processing, volume I. M.I.T. Press, Cambridge, MA.

Ruxin C. and Jamieson, L. H. (1996) *Explicit Modeling of Coartigulation in a Statistical Speech Recognizer*, Proceedings ICASSP 96, 463-466