BMBF

Verbmobil

Verbundvorhaben

Neural Networks for Nonlinear Discriminant Analysis in Continuous Speech Recognition

W. Reichl
S. Harengel
F. Wolfertstetter
G. Ruske

Technische Universität München

Vm

Report 111
April 1996

April 1996

W. Reichl
S. Harengel
F. Wolfertstetter
G. Ruske

Forschungsgruppe Sprachverarbeitung
Lehrstuhl für Mensch-Maschine-Kommunikation
Technische Universität München
Arcisstraße 21
80290 München

Tel.: (089) 2105 - 8554
e-mail: `reichl@e-technik.tu-muenchen.de`

Gehört zum Antragsabschnitt: TP3 Spracherkennung und Sprecheradaption

Das diesem Bericht zugrundeliegende Forschungsvorhaben wurde mit Mitteln des Bundesministers für Bildung, Wissenschaft, Forschung und Technologie unter dem Förderkennzeichen 01 IV 102 C/6 gefördert. Die Verantwortung für den Inhalt dieser Arbeit liegt bei den Autoren.

NEURAL NETWORKS FOR NONLINEAR DISCRIMINANT ANALYSIS IN CONTINUOUS SPEECH RECOGNITION

W. Reichl, S. Harengel, F. Wolfertstetter and G. Ruske

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

ABSTRACT

In this paper neural networks for Nonlinear Discriminant Analysis in continuous speech recognition are presented. Multilayer Perceptrons are used to estimate a-posteriori probabilities for Hidden-Markov Model states, which are the optimal discriminant features for the separation of the HMM states. The a-posteriori probabilities are transformed by a principal component analysis to calculate the new features for semicontinuous HMMs, which are trained by the known Maximum-Likelihood training. The nonlinear discriminant transformation is used in speaker-independent phoneme recognition experiments and compared to the standard Linear Discriminant Analysis technique.

1. INTRODUCTION

In this paper a Nonlinear Discriminant Analysis (NDA) is proposed, which uses neural networks (NN) to estimate a-posteriori probabilities. The common Linear Discriminant Analysis (LDA) is a well-known method for improving discrimination properties and compressing information in statistical pattern classification [5]. It has been applied in automatic speech recognition (ASR) and was reported to improve recognition performance in combination with Hidden-Markov Models (HMM) [7]. This is mainly contributed to the additional information in the contextual part of the input vector, which is included in the new feature vector after the transformation.

On the other hand neural networks were used successfully in pattern classification tasks [3,12]. They are inherently discriminative and yield estimates for a-posteriori probabilities of the classes when trained appropriately [3,12,13]. The relation between neural networks and discriminant analysis was reported in [1,6,10,14]. In combination with the HMM framework Multilayer Perceptrons (MLP) were successfully applied to ASR in order to calculate a-posteriori probabilities or likelihoods for HMM states [2,3,11].

In our Nonlinear Discriminant Analysis neural networks are used to estimate a-posteriori probabilities, which are the optimal discriminant features with respect to classification [1,4,9]. These a-posteriori probabilities are further processed by a principal component analysis to reduce the dimension of the new features [9]. This NDA transformation was used in speaker-independent phoneme recognition experiments with semicontinuous HMMs (SCHMM), optimized by the Maximum Likelihood (ML) training. The phoneme recognition results are compared to the performance of the original features and the standard LDA method.

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 authors

2. NEURAL NETWORKS FOR NONLINEAR DISCRIMINANT ANALYSIS

Linear Discriminant Analysis techniques were successfully applied to speech recognition [7]. The LDA is used to improve discrimination and compresses information by a linear transformation A of feature vector $\vec{x}$ into the new features $\vec{y}$, such that the class separability is maximum [5]. Therefore a discriminant criterion, e.g. $J = tr(S_T^{-1} S_B)$, is maximized in the transformed space $\vec{y} = A^T(\vec{x} - \vec{m})$, where S_T and S_B are the total-class and between-class scatter matrices and $\vec{m}$ is the center of the feature space [5]. In [7] speech recognition experiments with several different definitions of classes for the LDA were reported. Sub-phone units (HMM states) resulted in the most effective improvements for a continuous speech recognition task and will be used further as classes for the discriminant analysis.

2.1. Nonlinear Discriminant Analysis

In extension to Linear Discriminant Analysis a Nonlinear Discriminant Analysis was presented in [1,4,9]. The NDA is based on a nonlinear transformation of the feature vectors into a new feature space with maximum discrimination between the classes. The optimal discriminant features, maximizing the discriminant criterion J, are the a-posteriori functions $p(C_i|\vec{x})$ for the K classes $C_i, i = 1, \cdots, K$ [4]. These features would be optimal for classification and are incorporated in [3,11] directly in the Viterbi decoding without further modeling assumptions. Principal component analysis is employed for dimensionality reduction and decorrelation of the a-posteriori feature space to obtain the final feature vector for the HMM-classifier. The transformation $\vec{y} = U^T(\vec{p} - \bar{\vec{p}})$ reducing the dimension of the space of the a-posteriori probabilities $\vec{p} = (p(C_1|\vec{x}), \cdots, p(C_K|\vec{x}))^T$ uses the average of the a-posteriori probabilities $\bar{\vec{p}}$ for all training data and the matrix U. This matrix contains those eigenvectors of the covariance matrix in the $\vec{p}$-space, corresponding to large eigenvalues. The average of the a-posteriori

Presented at the *Proc. EUROSPEECH-95*, Madrid, Spain, September 1995, pp. 2163-2166.

probabilities for the training data results in the a prioris of the classes $\bar{p}_i = E\{(p(C_i|\vec{x})\} = p(i)$.

The new features are computed by the nonlinear transformation $\vec{y}(\vec{x})$ and are optimal with respect to the discriminant criterion J for the classification of the HMM states. A linear approximation of the a-posteriori functions, minimizing a mean squared error, results in the same eigenequation, which is derived in the linear optimization of the discriminant criterion J [5]. This means that LDA is the linear approximation of NDA through the linear approximations of the Bayesian a-posteriori probabilities [9]. The proposed NDA results in an optimized discriminant criterion, since the nonlinear transformation capabilities of neural networks yield an improved estimation of the a-posteriori probabilities.

2.2. Neural Networks for Nonlinear Discriminant Transformation

The relations between neural networks and linear and nonlinear discriminant analysis are examined in [1,6,10,14]. It is shown that minimizing a mean squared error (MSE) criterion with a 1-out-of-K class/target coding at the output of the neural network results in maximizing a discriminant criterion similar to J in the space spanned by the outputs of the hidden units of the MLP. Furthermore, the MSE optimization leads to estimates of the a-posteriori probabilities of the classes conditioned by the acoustic vector $\vec{x}$ [12,13]. The NN output nodes $o_i(\vec{x})$ are thus used in the proposed NDA as a-posteriori probability approximations $o_i(\vec{x}) \approx p(C_i|\vec{x})$. The optimal nonlinear features $\vec{y}$ with respect to the classes are thus calculated by the following

$$\vec{y} = U^T(\vec{o}(\vec{x}) - \bar{\vec{o}}(\vec{x})). \quad (1)$$

The vector $\vec{o}(\vec{x})$ contains all estimates of a-posteriori probabilities calculated by the neural net and $\bar{\vec{o}}(\vec{x}))$ represents their average for the whole training data. This new feature vector $\vec{y}$ was used in our experiments in subsequent SCHMMs for phoneme classification and was reduced to the same dimensionality as the original features.

Hence the principal component analysis for dimension reduction is a simple linear transformation it can be integrated in the neural net by adding an additional layer of linear neurons. Since we use Multilayer Perceptrons with one hidden layer to estimate the a-posteriori probabilities, the resulting neural network consists of two hidden layers with sigmoid neurons and a final layer with linear summing neurons.

In [3,11] the NN outputs are used directly as likelihoods or a-posteriori probabilities for the phoneme model states in the Viterbi search to approximate the a-posteriori probabilities of the phoneme models; therefore no further Gaussian distributions are needed to calculate state probabilities. However, in our NDA approach the NNs are utilized for the nonlinear transformation of the feature vector for subsequent SCHMMs, which consists of a full soft vector-quantization with Gaussian distributions and additional mixture components. Similar hybrid NN-HMM systems were described in [2,8], however their NNs were not designed to estimate class (i.e. HMM state's) probabilities, but were constructed to serve as special feature detectors (e.g. place and manner of articulation).

3. EXPERIMENTS

For the evaluation of the NDA speaker-independent phoneme recognition experiments were carried out. Therefore a database of 100 German speakers (Phondat "Diphon"-database) was used. We applied about 7700 sentences from 67 speakers for the training of the neural network, the computation of the principal component analysis and the SCHMM training. The reported phoneme recognition results were calculated for the remaining 3300 sentences from 33 different speakers. The speech data were sampled at 16kHz and a 256-point FFT with Hamming window was calculated every 10ms. The power spectrum was combined in 20 critical (Bark-scaled) bands and normalized to sum up to one. Together with the total loudness the 21-dimensional feature vector in the the original space was constituted and in some experiments the delta-loudness spectrum was added, obtaining 41 dimensions. To compare the results of the new transformed features to the original features the reduced new feature vector was reduced to the same dimensionality (21 or 41).

3.1. Neural Network Training

The neural networks for the estimation of the a-posteriori probabilities are Multilayer Perceptrons with one hidden layer of neurons (50 or 100 neurons). The output layer consists of 169 neurons, which is the total number of phoneme model states. All neurons are using the known sigmoid transfer function, except the additional linear neurons for the dimension reduction by the principal component analysis. The input layer of the neural net is made up of a sliding window of 3 or 5 consecutive feature vectors. In some experiments 5 vectors of loudness spectrum and total loudness were used, which resulted in a total NDA transformation from 5x21 to 21 dimensions. In further experiments the delta-loudness spectrum was included in the input vector and hence this is calculated in a contextual window too, only 3 of the 41 dimensional vectors were used in the input layer. The total NDA transformation is then mapping a 3x41 dimensional feature vector into a 41 dimensional space.

The training of the NN was performed by the Backpropagation algorithm, optimizing a mean squared error objective function between the NN outputs and their targets. These were chosen to 1.0 for the neuron of the corresponding class and 0.0 for all other neurons. This target coding scheme results in approximations of a-posteriori probabilities. The accuracy of the probability estimates was determined by calculating a histogram of output value distributions. Fore this purpose the NN output values were partitioned into 100 equal sized bins between 0.0 and 1.0. For each input pattern the bin counts for all output nodes were incremented. In addition, a second histogram counting only the distribution of the output node of the correct class was computed. If the outputs of the NN were approximations of a-posteriori probabilities, the relative frequency of the correct output values to all output nodes would be expected to be close to the corresponding center of the bins. Therefore a plot of the relative frequencies for each bin versus the bin centers should indicate a diagonal. In Figure 1 these relative frequencies for the output neurons of a MLP with 100 hidden neurons are shown. The measured values for re-

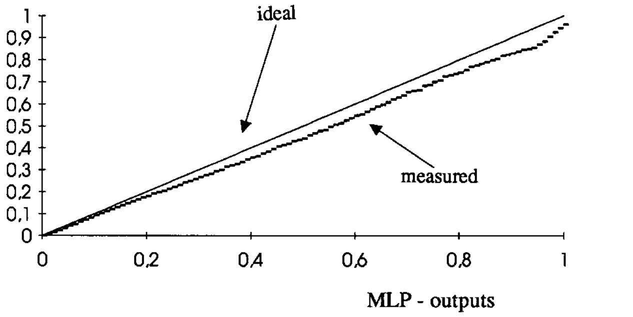

Figure 1: Relative frequency of correct class labeling by a neural network.

lative frequencies of correct output values indicate a good approximation of a-posteriori probabilities by the MLP. For higher output values the MLP is slightly overestimating the a-posteriori probabilities for the classes. For further verification NN output values must sum to one for each pattern and the averaged output of each node should equal the a priori probability of the associated class. The MLP estimates of the a-posteriori probabilities fulfill these requirements with only small averaged errors and hence they are appropriately applied in the NDA transformation.

3.2. Phoneme Recognition Experiments

After the calculation of the NDA transformation the semi-continuous HMMs for the phoneme recognition experiments were trained. Therefore 41 phoneme models (including silence) with 3 to 6 states were utilized in the SCHMMs. These consist of a soft vector-quantization with 256 prototypes and Gaussian pdfs with diagonal covariance matrices. Model training is performed according to the Maximum Likelihood principle using a Viterbi training algorithm based on the most probable state sequence.

To compare the NDA derived features to the standard LDA the linear transformation matrix A was determined and SCHMM learning was run in the LDA space, too. The LDA was operating on the same dimensions as NDA (5x21 to 21 and 3x41 to 41), using the same amount of context. In Table 3.2 the phoneme recognition rates for the different feature transformations are depicted. In column 1 the dimensions of the input vectors used for the transformation are printed. Remember, the dimension of the transformed feature space is identical to the original without context window. In column 2 the results for the basis system without transformation, using the original features, are printed. The numbers specifying the NDA columns indicate the number of hidden neurons in the NN used for the NDA transformation.

In these experiments identical conditions were kept and no lexicon, language model or biphon probabilities were utilized to examine only the performance of the new features on the acoustic-phonetic decoding. The phoneme recognition rates were evaluated within an automatically determined phoneme segmentation.

input vec. dim.	feature transformation without	LDA	NDA-50	NDA-100
5x21	55.7 %	58.0 %	58.0 %	58.6 %
3x41	57.4 %	55.0 %	59.5 %	61.1 %

Table 1: SCHMM phoneme recognition rates for different feature transformations.

In line 1 the improvements by LDA and NDA compared to the original features and the influence of the number of hidden neurons in the NN can be seen. Using the extended feature vector with delta loudness spectrum in line 2 causes problems in the LDA case, because in the new feature space much irrelevant information is preserved in the 41 dimensions after the relative moderate reduction from 123 dimensions. This was verified by an additional LDA experiment with further reduction of the output space to 21 dimensions, which improves the LDA performance significantly.

The NDA transformation of the feature space shows increasing recognition rates for all cases. The usage of the larger NN improves recognition rates up to 61.1 %, which is 3.7 % higher than the baseline system using the same features and identical number of parameters.

One important difference between LDA and NDA is depicted in Figure 2, where the normalized eigenvalues of the final principal component analysis of the transformations are shown. The decline of the normalized eigenvalues for the LDA is much steeper than for the NDA. Using eigenvectors corresponding to very small eigenvalues results in the utilization of dimensions with small information for class separation, such as in the LDA experiments.

In contrary the NDA is spreading class information over all dimensions. An additional experiment using 82 dimensions of the space spanned by the 169 class probabilities further improved recognition rates to 63.5 %, indicating the importance of all a-posteriori probabilities for classification.

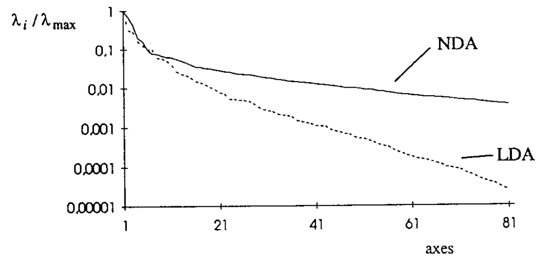

Figure 2: Normalized eigenvalues of LDA and NDA.

To improve phoneme recognition rate, biphon probabilities exploiting the statistical dependencies between the phonemes can be employed in the phonetic decoding. The incorporation of these phoneme transition probabilities, estimated on the training data, results in 70.8 % correct phoneme recognition without transformation and 72.8 % using

NDA transformation and the same quite simple SCHMMs with 256 prototype vectors. Better results would be expected for this task by more sophisticated HMMs (e.g. continuous density HMMs). Furthermore the NDA is not restricted to phoneme recognition, and by the usage of a lexicon and a language model improvements in word recognition rates will be achieved.

4. CONCLUSIONS

A Nonlinear Discriminant Analysis was presented, which is based on a-posteriori probabilities, estimated by neural networks. We compared this nonlinear approach to the common linear transformation and showed improvements in phoneme recognition experiments of 3.7 points to 61.1 % without biphon probabilities and of 2.0 points to 72.8 % incorporating biphon probabilities and the new nonlinearly derived features. The amount of improvement is dependent on the accuracy of the estimates of the Bayesian probabilities, which was examined for the utilized neural networks. Increasing the approximation capabilities of the NN will lead to further improvements in recognition results and will be examined together with a joint optimization of NN and HMM by discriminant learning techniques.

5. REFERENCES

[1] H.Asoh, N.Otsu, *Nonlinear Data Analysis and Multilayer Perceptrons,* Proc. Int. Joint Conference on Neural Networks, pp. 411-415, 1989.

[2] Y.Bengio, R.DeMori, G.Flammia, R.Kompe, *Global Optimization of a Neural Network-Hidden Markov Model Hybrid,* IEEE Trans. on Neural Networks, vol. 3, no. 2, pp. 252-259, March 1992.

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

[4] K.Fukunaga, S.Andro, *The Optimum Nonlinear Features for a Scatter Criterion in Discriminant Analysis,* IEEE Trans. on Information Theory, vol. 23, no. 4, pp. 453-459, July 1977.

[5] K.Fukunaga, *Introduction to Statistical Pattern Recognition,* 2nd ed., Academic Press, 1990.

[6] P.Gallinari, S.Thira, F.Badran, F.Fogelman-Soulie, *On the Relations Between Discriminant Analysis and Multilayer Perceptrons,* Neural Networks, vol. 2, pp. 349-360, 1991.

[7] R.Haeb-Umbach, H. Ney, *Linear Discriminant Analysis For Improved Large Vocabulary Continuous Speech Recognition,* IEEE Proc. 1992 Int. Conf. Acoust. Speech Signal Process., San Francisco, pp. 13-16, March 1992.

[8] F.T.Johansen, M.H.Johnsen, *Non-Linear Input Transformation For Discriminative HMMs,* IEEE Proc. 1994 Int. Conf. Acoust. Speech Signal Process., Adelaide, pp. 225-228, April 1994.

[9] T.Kurita, H.Asoh, N.Otsu, *Nonlinear Discriminant Features Constructed by Using Outputs of Multilayer Perceptron,* III Int. Symp. on Speech, Image Process. and Neural Networks, Hong Kong, pp. 417-420, April 1994.

[10] D.Lowe, A.Webb, *Optimized Feature Extraction and the Bayes Decision in Feed-Forward Classifier Networks,* IEEE Trans. on Pattern Analysis and Machine Intelligence, vol. 13, no. 4, pp. 355-364, April 1991.

[11] 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.

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

[13] D.Ruck, S.Rogers, M.Kabrisky, M.Oxley, B.Suter, *The Multilayer Perceptron as an Approximation to a Bayes Optimal Discriminant Function,* IEEE Trans. on Neural Networks, vol. 1, no. 4, pp. 296-298, December 1990.

[14] A.Webb, D.Lowe, *The Optimized Internal Representation of Multilayer Classifier Networks Perform Nonlinear Discriminant Analysis,* Neural Networks, vol. 3, pp. 367-375, 1990.