NONLINEAR DISCRIMINANT ANALYSIS WITH NEURAL NETWORKS FOR SPEECH RECOGNITION

Vincent Fontaine, Christophe Ris, Henri Leich

Faculté Polytechnique de Mons — TCTS
31, Bld. Dolez, B-7000 Mons, Belgium
Tel : + 32 65 374176 - Fax : + 32 65 374129
e-mail: {fontaine,ris,leich}@tcts.fpms.ac.be

ABSTRACT

Linear Discriminant Analysis (LDA) has been applied successfully to speech recognition tasks, improving accuracy and robustness against some types of noise. However, it is well known that LDA suffers from some weaknesses if the distributions are not unimodal or when the mean of the distributions are shared.
In this paper, we propose to take advantage of the nonlinear discriminant properties of the Artificial Neural Networks (ANN) in the task of reducing the dimensionality of the input space, leading to a nonlinear discriminant analysis.

1 INTRODUCTION

Speech recognition basically appears to be a statistical pattern classification problem including classical imperatives such as compression and discrimination of the speech features. Such imperatives can be satisfied by applying a so-called discriminant analysis consisting in defining a transformation of a certain signal representation into another one in order to fit the data to some phonetic classification.

Linear Discriminant Analysis (LDA, [3]) has been widely applied to speech recognition resulting in improved recognition performance [5], [1], [4] and improved robustness [7]. In this approach, the idea is to find a linear transformation that projects a n-dimensional space on a m-dimensional space ($m < n$) such that the class separability is maximum. In this paper we propose to use Artificial Neural Networks (ANN) and more particularly MLPs as feature analyzer taking advantage of their discriminant properties in classification tasks. Indeed, we can consider that each hidden layer of a MLP proposes an internal representation of the input signal that prepares the signal to the classification task. Therefore, such a representation can be seen as a nonlinear discriminant analysis (NLDA) of the input features and provides an alternative to classical speech features (MFCC, LPC-cepstrum, ...). Another MLP based feature extraction techniques has been proposed [2], such as auto-associative networks in which the inputs and the outputs were identical providing a good but non discriminant compression tool[1].

2 LINEAR DISCRIMINANT ANALYSIS

The purpose of discriminant analysis is to find parameters that are well suited for classification tasks. We know that the optimal parameters for a classification task are the *a posteriori* probabilities of the classes given the observations. Unfortunately these *a posteriori* probabilities are very hard, if not impossible, to determine. LDA provides a good alternative for computing discriminant parameters since it is based on simple criteria associated with systematic feature extraction algorithms.

Discriminant features are obtained by designing a linear transformation of vectors x (n-dimensional) into vectors y (m-dimensional, $m < n$) such that the class separability is maximum. Class separability can be defined in several ways from scatter matrices. Usually, two matrices, S_1 and S_2, out of three[2] are used to build the optimization criterion. The most widely used criteria are :

$$J_1 = tr(S_2^{-1} S_1) \tag{1}$$

$$J_2 = det(S_2^{-1} S_1) \tag{2}$$

where $tr(A)$ denotes the trace of the matrix A and $det(A)$ its determinant.

It is shown in [3] that optimization of J_1 or J_2 leads to the same linear features and that this optimization is independent of the choice of scatter matrices for S_1 and S_2. It is also shown that the optimal linear transformation, for criterion J_1 or J_2, is obtained by calculating the eigenvectors of the matrix.

We see from the above developments that LDA is very attractive for classification since it provides a simple and

[1] Moreover, it has been shown [2] that the optimal values of the weights can be deduced from classical linear algebra.

[2] The within-class scatter matrix S_w, the between-class scatter matrix S_b, and the mixture scatter matrix S_m

straightforward method for extracting discriminant features. Furthermore, the corresponding linear transformation is optimal with respect to the class separability defined as J_1 or J_2.

However, one could worry about the efficiency of the criterion itself, i.e. how accurately does $tr(S_2^{-1}S_1)$ measures the class separability ? Generally speaking, $tr(S_2^{-1}S_1)$ is a good measure of class separability when distributions are unimodal and separated by the scatter of the means. When the distributions are multimodal and when the means are close together, efficiency of the criterion becomes very bad. At the extreme, if the distributions share the same mean, the between-class scatter matrix becomes a null matrix and optimization is no more efficient.

If we observe the distributions of classical feature vectors used in speech recognition (e.g. LPC cepstrum), we can state that these distributions are far from unimodal (therefore emission probabilities of HMMs are often modeled by multi-gaussian distributions) and that the overlapping between classes is very high. In this context, a good efficiency of the optimization criterion is not guaranteed, even if LDA is applied on several feature vectors. So, we are convinced that the use of neural nets for discriminant analysis can improve further results obtained with LDA.

3 NONLINEAR DISCRIMINANT ANALYSIS

Neural networks do not suffer from the same constraints about the distributions of the classes since they are basically able to model highly complex nonlinear problems even if they can not cope with the between classes overlapping problem.

Interpretations of the output of each layer of an Artificial Neural Network should convice us about the ability of ANNs to extract pertinent information for classification. Actually, the action of a hidden layer consists in transforming its inputs (n-dimensional) into m-dimensional outputs ($m \leq n$ or $m > n$) according to a nonlinear transformation. If we now consider the last hidden layer of a MLP, its action can be viewed as generating discriminant features for the output layer.

Nonlinear discriminant analysis can then be achieved by designing a MLP where the number of nodes contained in the last hidden layer is inferior to the number of input nodes. Based on this architecture, the hidden layer will act as a bottle-neck both decreasing redundancy from the input layer and extracting relevant information for the classification.

3.1 Training Phase

The neural network is trained in a classical way (backpropagation for MLPs). Therefore, we have to fix (figure 1):

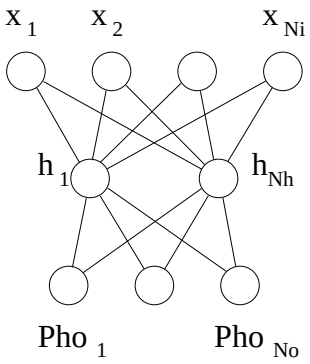

Figure 1: Configuration of the MLP for training

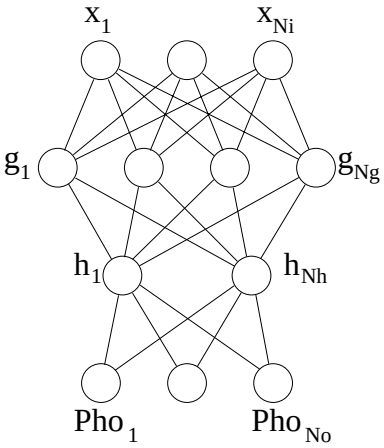

Figure 2: MLP with two hidden layers

- The kind of feature we want to transform. Actually, we were interested in the opportunity offered by the MLP to take the phonetic context of the signal into account by feeding the network with several successive frames of cepstral coefficients.

- The classes which the features will be dedicated to. Different acoustic classes can be defined, based on sub-words units as phonemes or sub-phone units, ...

One could worry about the possibility to train efficiently a neural network designed with a small number of hidden nodes. Practically, training such ANNs will be efficient only for simple tasks. A better way to train ANNs containing a bottle-neck is to introduce a second hidden layer containing a high number of neurons (figure 2). The introduction of this second hidden layer will increase the number of parameters of the MLP allowing better classification performance.

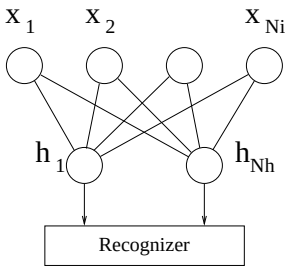

Figure 3: Configuration of the MLP as feature extractor

3.2 Recognition Phase

Once the neural network has been trained, we expect that the outputs of the last hidden layer will provide us with discriminant features that will be fed to a classical recognizer (figure 3).

The analysis we propose in this paper will act as a pre-processing of the data that could be easily inserted in any recognition chain (discrete HMM, Multi-gaussian HMM, hybrid HMM/MLP, ...). The so-defined features gather some advantages :

- We are expected to achieve a high class separability.
- We can expect that redundant information has been filtered by the MLP (as an effect of the optimization of the use of each parameter).
- As, under some training constraints, MLPs provide estimations of *a posteriori* probabilities [2], the optimization criterion we use for our discriminant analysis is directly related to *a posteriori* probabilities which is not the case for LDA.
- On the opposite to cepstral features which are widely used in speech recognition, all our parameters are relevant (indeed, only the first cepstral coefficients are used, usually 12 to 16), so that different acoustic vector dimensions could be used.
- The variances of the parameters are naturally normalized (effect of the sigmoid function) which could be of some interest when vector quantization is applied.

4 DATABASES AND RECOGNITION TASKS

Two databases have been used to assess NLDA :

1. The first results on NLDA have been obtained on the database collected in the framework of the ESPRIT P6488 Himarnnet project. This database, referred to as HER, consists in 675 speakers pronouncing each 108 isolated words over the telephone line. The database has been recorded in Switzerland and words are spoken in German. Out of the 675 speakers, 250 have been picked up for the training set and 90 were used as test set.

   The recognition task consisted in the recognition of 54 words defined for the industrial application of the Himarnnet project using discrete HMMs for computational load reasons.

2. The second database we used is PhoneBook, which is completely described in [6]. The main features of PhoneBook are :

   - 92000 isolated words;
   - 1300 talkers;
   - vocabulary of about 8000 words;
   - American English;
   - Telephone quality;

   The database is decomposed in 106 lists of 75 words and each word is pronounced approximately 10 times. The training set was composed of 98 lists and the 8 remaining lists were used to design 8 independent test sets.

Recognizer	Recognition rate (%)
k-means 256 centroids	84.0 %
NLDA 60-12 + k-means	85.7 %
NLDA 60-15 + k-means	85.8 %

Table 1: Recognition results using a nonlinear discriminant analysis on a discrete recognizer.

5 EXPERIMENTAL RESULTS

5.1 HER

In order to clearly appretiate the influence of NLDA on discrete HMMs, simple discrete systems have been trained on the HER database. As feature vectors, we used only 12 liftered LPC-cepstral coefficients. A cepstral mean subtraction (CMS) was used in order to reduce the effect of the transmission channel. Vector quantization was realized by a k-means algorithm generating 256 prototypes of cepstral vectors used to train context-independent phoneme models. Results comparing the baseline discrete recognizer with the NLDA-discrete system are presented in table 1.

In the second and third experiments, a nonlinear discriminant analysis has been performed on five successive feature vectors, extracting 12 and 15 discriminant coefficients, respectively. From the results, we can observe an improvement of 1.7% and 1.8% of the recognition rate which corresponds to a reduction of 11% of the error rate. These results confirm that the NLDA is able

Recognizer	Recognition rate (%)
Baseline	84.4 %
NLDA 182-26-44	75.4 %
NLDA 182-100-26-44	84.0 %
NLDA 182-200-26-44	88.3 %

Table 2: Recognition results using a nonlinear discriminant analysis on a continuous recognizer.

to extract useful parameters for classification tasks and that the vector quantization can take benefit from the normalization of the parameters even with very small networks.

5.2 PhoneBook

The encouraging results obtained on HER motivated tests on a state-of-the-art recognizer. Therefore, we decided to test NLDA on continuous HMMs trained on PhoneBook. The baseline system has been designed with the following characteristics :

- Phonemes are context-independent.
- Each phoneme model was composed of three independent states.
- The distribution of each state was modeled by a mixture of 8 gaussian distributions.
- The 26-component feature vector was composed of 12 rasta-plp coefficients, their first derivative, the first and second derivative of the log-energy.

NLDA has been applied using networks of different sizes; the results are summarized in table 2.

We can see from table 2 that the performance of the NLDA strongly depends on the size of the MLP. For the second experiment, the MLP is clearly too small to estimate correctly *a posteriori* probabilities. As a consequence, the parameters extracted by the NLDA are not as efficient as the feature vector used for the baseline system.

The third and fourth experiments show that training MLPs with two hidden layers can overcome this problem. The biggest MLP was able to reduce the error rate by 25 %. Further reduction of the error rate could certainly be achieved by training bigger MLPs but this has not yet been tried.

6 CONCLUSIONS

After highlighting some weaknesses of the classical linear discriminant analysis, this paper presents a method for performing nonlinear discriminant analysis by taking benefit of the nonlinear discriminant properties of the MLPs.

The NLDA has been tested on discrete and continuous recognizers leading to significant reduction of the error rates when sufficiently big MLPs were used. In the case of the continuous system, the NLDA reduced the error rate by 25 %.

7 ACKNOWLEDGEMENTS

The authors would like to thank Xavier Kaufmann for his contribution to this work.

References

[1] X. Aubert, R. Haeb-Umbach, and H. Ney. Continuous mixture densities and linear discriminant analysis for context -dependent acoustic models. In *Proceedings of ICASSP93*, pages II–648 – II–651. IEEE, 1993.

[2] Hervé Bourlard and Nelson Morgan. *Connectionist Speech Recognition.* Kluwer Academic Publishers, 1994.

[3] K. Fukunaga. *Introduction to Statistical Pattern Recognition.* Academic Press, 1990.

[4] R. Haeb-Umbach, D. Geller, and H. Ney. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. In *Proceedings of ICASSP93*, pages II–239 – II–242. IEEE, 1993.

[5] R. Haeb-Umbach and H. Ney. Linear discriminant analysis for improved large vocabulary continuous speech recognition. In *Proceedings of ICASSP92*, pages I–13 – I–16. IEEE, 1992.

[6] John F. Pitrelli, Cynthia Fong, Suk H. Wong, Judith R. Spitz, and Hong C. Leung. Phonebook : A phonetically-rich isolated-word telephone-speech database. In *Proceedings of ICASSP95*, pages 101 – 104. IEEE, 1995.

[7] O. Siohan. On the robustness of linear discriminant analysis as a preprocessing step for noisy speech recognition. In *Proceedings of ICASSP95*, pages 125 – 128. IEEE, 1995.