NONLINEAR DISCRIMINANT ANALYSIS FOR IMPROVED SPEECH RECOGNITION

Vincent Fontaine, Christophe Ris and Jean-Marc Boite

Faculté Polytechnique de Mons, 31, Boulevard Dolez, B-7000 Mons, BELGIUM
Tel: ++ 32 65 374176 - Fax: ++32 65 374129
Email: fontaine,ris@tcts.fpms.ac.be

ABSTRACT

Linear Discriminant Analysis (LDA) has been widely applied to speech recognition resulting in improved recognition performance and improved robustness. LDA designs a linear transformation that projects a n-dimensional space on a m-dimensional space ($m < n$) such that the class separability is maximum. This paper presents new results related to our previous work [6] on nonlinear discriminant analysis (NLDA) based on the discriminant properties of Artificial Neural Networks (ANN) and more particularly MLP. Experiments performed on the isolated word large vocabulary PHONEBOOK database show that NLDA provides a method for designing discriminant features particularly efficient as well for continuous densities HMM as for hybrid HMM/ANN recognizers.

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.

Discriminant features are often computed by applying a Linear Discriminant Analysis (LDA) on sequences of acoustic vectors. LDA extracts, from these sequences, a set of discriminant parameters maximizing the class separability by designing a linear transformation that projects a n-dimensional space on a m-dimensional space ($m < n$). Previous works show that application of LDA to speech recognition problems increases performance ([2], [3], [4]) and robustness against some types of noises [6].

In this paper, we propose a method for extracting discriminant parameters using Artificial Neural Networks (ANN) and more particularly Multilayer Perceptrons (MLP). ANN are indeed powerful tools that can be trained to solve complex nonlinear classification problems. Each hidden layer of feed-forward networks computes its outputs as a nonlinear transformation of its inputs, so that we can consider that each hidden layer 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, RASTA-PLP cepstrum, ...).

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.

LDA computes discriminant features by designing a linear transformation of vectors x (n-dimensional) into vectors y (m-dimensional, $m < n$) such that class separability is maximum. Class separability is generally defined as the trace or determinant of the product of scatter matrices [1] :

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

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

where S_1, S_2 are two scatter matrices out of three (the within scatter matrix, S_w, the between scatter matrix, S_b, and the mixture scatter matrix, S_m); $tr(A)$ denotes the trace of the matrix A and $det(A)$ its determinant.

It can be shown that the optimal linear transformation for criterion J_1 or J_2 is obtained by calculating the eigenvectors of the matrix $(S_2^{-1}S_1)$.

Though experiments demonstrate the efficiency of LDA on speech recognition tasks, 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

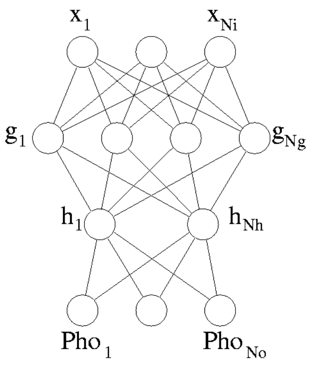

Figure 1: MLP with two hidden layers

separability when distributions are unimodal and separated by the scatter of the means. When the distributions are multimodal (as it is the case for speech) and when the means are close together, efficiency of the criterion becomes very bad. So, we are convinced that the use of neural nets for discriminant analysis can further improve 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 class overlapping problem.

Each hidden layer of a MLP performs a nonlinear discriminant analysis (NLDA) of the input features. 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.

One could worry about the possibility to train efficiently a neural network designed with a small number of hidden nodes. Practically, training such ANN will be efficient only for simple tasks. A better way to train ANN containing a bottle-neck is to introduce a second hidden layer containing a high number of neurons (as illustrated in figure 1).

The MLP is trained in a classical way (error back-propagation) to classify sequences of feature vectors (to catch the context information) in terms of phonetic classes.

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 (discrete HMM, Multi-gaussian HMM, hybrid HMM/MLP, ...) as shown in figure 2.

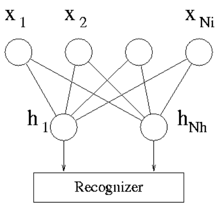

Figure 2: Configuration of the MLP as feature extractor

The so-defined features gather the following 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 global parameter optimization).
- As, under some training constraints, MLP provide estimations of *a posteriori* probabilities [7], 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

For all experiments, we used the PhoneBook database, which is completely described in [8]. PhoneBook is a phonetically rich, large vocabulary, isolated words database. The main features of PhoneBook are :

- 92,000 isolated word utterances
- 1,300 American English native speakers, each of them pronouncing 75 words in average
- 8,000 words vocabulary
- Telephone quality

The database is composed of 106 lists of 75 words (totalling 8,000 vocabulary words), each word being uttered approximately 11 times. The training set was composed of 21 lists (approximately 5 hours of speech) and 8 lists were used to design test sets(6598 utterances).

Two test sets were designed for testing the recognizers :

1. The first test set (refered to as *test set 1* in the sequel) aims to test the systems for small vocabulary tasks. Therefore, we ran eight recognition tests independently on the eight different word lists. This provided us with eight recognition rates that are averaged to obtain the global recognition rate. The words of the eight test word lists did not belong to the training vocabulary.

2. The second test set (*test set 2*) is build with the same eight word lists of 75 words but the recognition is performed with a global dictionary of 600 words ($8 * 75$). The second test set corresponds to a medium vocabulary, task independent experiment.

Tests have been conducted on both a continuous densities HMM recognizer (CDHMM) and a hybrid HMM/MLP recognizer. The feature vector was composed of 26 parameters (12 RASTA-PLP coefficients, their first derivatives, the first and second derivatives of log-energy) for the HMM/MLP recognizer and 38 parameters (including second derivatives of RASTA-PLP coefficients) for the CDHMM.

In the CDHMM based recognizers, emission probabilities of context-independent phone models (3 states/phone) were estimated by gaussian mixtures (12 mixtures/state) with diagonal covariance matrices. Neither state tying, nor mixture tying was applied in our experiments.

NLDA parameters were computed from two hidden layer MLP trained to estimate posterior probabilities of context independent phone models given nine frames of contextual information.

5. RESULTS

Recognition results are presented in table 1 for continuous densities HMM and in table 2 for the hybrid HMM/MLP recognizers. Differences of performance between CDHMM and hybrid HMM/ANN can be explained by the fact that we only trained context independent phone models and that a minimum duration of phone models was imposed for the hybrid systems and not for CDHMM.

In our experiments, we first try to extract NLDA parameters from a single hidden layer MLP (NLDA-234-38-47). Corresponding results indicate clearly that this structure is inefficient due to the reduced size of the MLP that is unable to estimate reliable posterior probabilities. Experiments with two hidden layer MLP show that improvement on the continuous densities recognizer is quite significant (about 25% reduction of the error rate) on both test sets. This could be explained by the fact that gaussians have diagonal covariance matrices which supposes that the parameters of the feature vectors are decorrelated. The MLP used for NLDA probably decorrelates the parameters to extract a maximum of information matching the assumption of diagonal covariance matrices. To verify this assumption we compared the correlation coefficients of RASTA-PLP parameters and NLDA parameters as following :

Recognizer	Rec rate (%) Test set 1	Rec rate (%) Test set 2
Baseline	6.7 %	17.7 %
NLDA 234-38-47	17.6 %	-
NLDA 234-300-26-47	5.3 %	-
NLDA 234-300-38-47	5.1 %	14.0 %
NLDA 234-300-64-47	5.8 %	-

Table 1: Recognition results using a nonlinear discriminant analysis on a continuous recognizer. NLDA $a - b(-c) - d$ stands for parameters extracted using an MLP designed with a input nodes, b hidden nodes in the first hidden layer, c hidden nodes in the second hidden layer if any and d outputs.

Let v be the complete feature vector (including derivatives if any). For each HMM state, we computed the correlation matrix between the coefficients of the feature vectors :

$$\rho = [\rho_{ij}] = [\frac{\sigma_{ij}}{\sigma_i \sigma_j}] \qquad (3)$$

where $\sigma_{ij} = E[(v(i) - \mu(i))(v(j) - \mu(j))]$ and $\sigma_i = \sqrt{\sigma ii}$. To facilitate the comparison of correlation matrices, we computed the value r related to the correlation matrix by :

$$r = \sum_{i=1}^{N} \sum_{j=1}^{N} \rho_{ij}^2 \qquad (4)$$

where N is the feature vector dimension. This value gives us an idea of the global correlation between the coefficients of a feature vectors. In figure 3 we computed the global correlation of the feature vector coefficients (RASTA-PLP for solid line and NLDA for dotted line) corresponding to each phoneme model. This figure shows that global correlations for RASTA-PLP and NLDA parameters are almost the same. This is quite interesting since NLDA parameters are extracted from several context frames. This indicates that NLDA is able to extract parameters incorporating context information while keeping global correlation at the same level as for one RASTA-PLP feature vector.

In a second set of experiments, we trained neural network directly on NLDA parameters, again with some context frames (9 frames except for the 64 components NLDA vector where we used 5 context frames). Results with these hybrid recognizers also show improved performance especially for the second test set. However

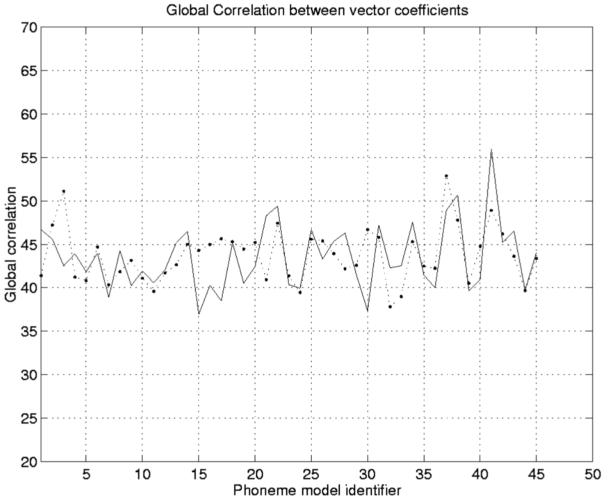

Figure 3: MLP with two hidden layers

Recognizer	Rec. rate (%) Test set 1	Rec. rate (%) Test set 2
Baseline	2.1 %	8.0 %
NLDA 234-300-26-47 Rec 234-600-47	1.8 %	6.5 %
NLDA 234-300-38-47 Rec 342-600-47	1.9 %	6.7 %
NLDA 234-300-64-47 Rec 320-600-47	1.8 %	6.7 %

Table 2: Recognition results using a hybrid HMM/MLP recognizer. NLDA $a-b-c-d$ gives the topolgy of the MLP used to extract the NLDA parameters while Rec $a-b-c$ gives the topology of the MLP used for probability estimation.

improvements are not so important as for the CDHMM probably because both MLP (used for NLDA, and for probability estimation) are trained with the same criterion. Improvements could result from a better modelization of the context since the probability estimator accounts for some context of the discriminant features that are themselves extracted from some acoustic context. It is interesting to note that increasing the context (more than 90 ms) for the baseline recognizer never led to better performance. Also increasing the number of hidden nodes (to 1,000) for the baseline did not decrease the error rate.

It is interesting to note that improvements generated by NLDA parameters are quite independent of the size of feature vectors : quite similar results have been achieved for 26, 38 and 64 components vectors.

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 MLP.

The NLDA has been tested on small and medium vocabulary, task independent recognition experiments. Recognition results show that the nonlinear discriminant analysis offers an efficient way of computing discriminant parameters leading to reduction of the relative error rate up to 25 % for the CDHMM and up to 20 % for hybrid HMM/ANN systems.

7. REFERENCES

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

[2] 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.

[3] 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.

[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] 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.

[6] V. Fontaine, C. Ris, and H. Leich. Nonlinear discriminant analysis with neural networks for speech recognition. In *Proceedings of EUSIPCO96*, pages 1583 – 1586. EURASIP, 1996.

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

[8] 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.