LINEAR DISCRIMINANT ANALYSIS FOR IMPROVED LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION

R. Haeb-Umbach, H. Ney

Philips Research Laboratory Aachen
P.O. Box 1980, D-5100 Aachen, Germany

ABSTRACT

The interaction of Linear Discriminant Analysis (LDA) and a modeling approach using continuous Laplacian mixture density HMMs is studied experimentally. The largest improvements in speech recognition accuracy could be obtained when the classes for the LDA transform were defined to be sub-phone units. On a 12,000-word German recognition task with small overlap between training and test vocabulary a reduction in error rate by one fifth was achieved compared to the case without LDA. On the development set of the DARPA RM1 task the error rate was reduced by one third. For the DARPA speaker-dependent no-grammar case, the error rate averaged over 12 speakers was 9.9%. This was achieved with a recognizer employing LDA and a set of only 47 Viterbi-trained context-independent phonemes.

1. INTRODUCTION

Linear Discriminant Analysis (LDA) is a well-known technique in statistical pattern classification for improving discrimination and compressing the information contents (with respect to classification) of a feature vector by a linear transformation. LDA has been applied to automatic speech recognition tasks [1 - 6] and resulted in improved recognition performance for small-vocabulary systems [2], [3], [4].

For large vocabulary phoneme-based recognizers, the results reported do not give a clear picture. Yu et al. [5] employed LDA to transform the feature space (before vector quantization) of the BBN BYBLOS recognizer, a system based on context-dependent discrete HMMs derived from three codebooks. They defined the 50 basic phonemes as the classes to be discriminated. The discriminant analysis did not result in an overall improvement. Wood et al. [6] identified the classes with sub-phone units, so-called phonicles, which were modeled by multivariate Gaussians. They achieved improved recognition accuracy by applying the IMELDA transform [3].

The large vocabulary speech recognizer presented here is different from the ones above. We use context-independent phonemes and continuous mixture density HMMs. When employing LDA in such a framework several design alternatives have to be considered, e.g. what is the best definition of the classes to be discriminated. In Section 2 we will describe these design considerations. After a brief description of the recognition system and the data bases in Section 3 we will describe the design of the LDA based recognizer on the basis of experimental results on a 12,000-word German task (Section 4). Section 5 presents results on the DARPA task.

2. LDA FOR PHONEME-BASED RECOGNITION

Ignoring the time alignment problem for the moment, we regard speech recognition as a pattern classification task. The idea of LDA is to find a linear transformation of feature vectors X from an n-dimensional space to vectors Y in an m-dimensional space ($m < n$) such that the class separability is maximum [7]. Scatter matrices are used to formulate the optimization problem. Two matrices S_1, S_2 out of the three — W: within-class, B: between-class, T: total scatter matrix — are used where many combinations are possible. Several optimization criteria are also conceivable, the most widely used ones are to maximize

$$J_1(m) = tr(S_{2y}^{-1} S_{1y}) \qquad (1)$$

0-7803-0532-9/92 $3.00

$$J_2(m) = det(S_{2y}^{-1} S_{1y}) \quad (2)$$

where $tr(A)$ denotes the trace of A, $det(A)$ its determinant, and S_{iy} is the scatter matrix in the m-dimensional y-space. We chose $S_1 = T, S_2 = W$. That is, we consider a class-independent linear transformation of the vector space that enhances the total scatter while keeping the within-class scatter constant.

The optimization of (1) and (2) leads to the result that the input vector x has to be projected onto the subspace spanned by those m eigenvectors of $S_{2x}^{-1} S_{1x}$ which correspond to the m largest eigenvalues. Note that J_1 and J_2 lead to the same set of features and that the criteria are invariant under any nonsingular linear transformation both in the original n-dimensional space and in the resulting m-dimensional space. Even the resulting features are the same irrespective of any linear transformation in the n-dimensional space prior to the LDA transformation.

If LDA is applied as a preprocessing step in a large vocabulary continuous speech recognizer several questions arise. First, it has to be investigated how to apply LDA in the framework of mixture densities. The speech recognizer under consideration employs continuous Laplacian mixture densities which potentially render it incompatible with LDA. Second, it is not obvious what the most appropriate definition of classes is that we want to discriminate. One might argue that defining a class to be a phoneme is appropriate for our phoneme-based recognizer. However, arguments for different class definitions can also be found. Currently, our understanding of the intricate interactions between LDA and the modeling approach is such that these questions can only be answered on an experimental basis.

3. SYSTEM DESCRIPTION AND DATA BASE

The system is a phoneme-based speaker-dependent continuous speech recognizer using Laplacian mixture density HMMs. There is a set of 40 – 50 monophones each of which consists of 3 sub-phone units, which we called "phoneme segments". Each segment consists of two states which have the same emission probability density attached to them. The acoustic resolution is controlled by the number of elementary densities per mixture which is chosen to be at least 30 densities per mixture on the average. A vector of absolute deviations is pooled over all elementary densities of all mixtures. The Viterbi criterion, i.e. the most likely state sequence, is used both in training and recognition [8].

In the preprocessing stage of the recognizer a 30–channel FFT–based filter bank is used. These 30 log-energy intensities together with 15 first-order (30ms time span) and 15 second-order (60ms time span) time differences, the average intensity and its time differences, form a 63–component acoustic vector [8]. This analysis is carried out every 10ms. When LDA is applied, adjacent 63–component vectors may be adjoined to form an augmented vector for the subsequent LDA-transform. As the output of the LDA transform, 35 components are retained.

We carried out recognition experiments on two data bases. The first is a German business correspondence task with a 12,306–word recognition vocabulary. The training is based on 300 sentences comprising 2735 words (about 20 min of speech). The recognition tests are performed on a set of 50 sentences comprising 1099 words. There is only a small overlap between training and test vocabulary. All sentences were recorded for each of 4 speakers. The sentences were read but there was no attempt to enforce careful pronunciation and high signal-to-noise ratio. To account for variations in the recording conditions it turned out to be very important to normalize the acoustic vector with respect to an estimate of the long-term spectrum. The stochastic language model was a unigram model of test set perplexity 1831.

The second data base is the well–known DARPA resource management task RM1 which is available on CD-ROM. Details of this task can be found elsewhere, e.g. [9]. We present results on the 100–sentence development set averaged over 12 speakers.

4. LDA DESIGN AND RESULTS ON GERMAN TASK

We performed various experiments to determine the parameters of the LDA transform. In a first experiment we investigated the interaction of LDA and our modeling approach which employs mixture densities. It turned out that it is not sufficient to transform the centroids of the elementary densities, which had been determined in a training procedure based on the original non-transformed feature vectors, to obtain suitable

Table 1: *Word error rates in % (tot = del + ins + sub) for different class definitions:*
A: *Baseline system without LDA;*
B: *Class = phoneme;*
C: *Class = mixt. density (= phoneme segment);*
D: *Class = elementary density.*
In (B - D), 35 components are retained after transformation.

Speaker	A			B		
	del	ins	tot	del	ins	tot
M-21	2.6	2.4	19.4	2.5	1.6	17.7
M-22	3.2	3.2	23.1	2.4	2.4	19.7
M-24	5.4	2.5	27.8	6.3	1.5	26.8
M-25	1.6	1.8	16.3	1.8	2.1	17.2
Average			21.7			20.3
Speaker	C			D		
	del	ins	tot	del	ins	tot
M-21	2.5	1.0	15.9	2.2	1.4	16.3
M-22	2.5	2.2	17.9	2.5	2.1	19.7
M-24	5.4	2.5	27.0	4.6	1.6	25.1
M-25	1.5	2.3	15.3	1.3	2.4	15.5
Average			19.0			19.1

reference vectors in the transformed space. LDA involves a simultaneous diagonalization of two matrices which consists of a rotation plus scaling ("whitening transformation") followed by a second rotation. Because of the scaling step, the centroids resulting from a training based on non-transformed data may not be chosen optimally. However, in a new training working on the transformed input data the centroids of the elementary densities can be chosen such that they better fit the transformed data. Therefore, a 3-step training procedure is used: to obtain a segmentation for the training data, we first estimate the speech models using our standard training techniques and then segment the data automatically using the recognizer constrained to find the correct word sequence. In the second step the recognized segment boundaries are then used to assign class labels to each frame. Within-class and total scatter matrices are estimated and the LDA transformation matrix is computed. The third step is a new Viterbi based mixture density training with the transformed feature vectors at the input.

In another set of experiments we tested different class definitions to be used in LDA. Table 1 summarizes the results for each of four speakers with a unigram language model of test set perplexity 1831.

Experiment A is the baseline system without LDA preprocessing. In experiment B the classes used were the 44 basic phonemes which is the same class definition as in [5]. In C, a finer resolution was chosen. Similar to [6], a class was defined to be one of the 3*44 phoneme segments. Recall from Section 3 that each segment is modeled by a unique mixture density. Finally, a class was associated with an elementary density in experiment D. Since a mixture consists of 30 elementary densities on the average, a total of about 4000 classes results.

For each choice of class definition we achieved an improvement over the baseline system without LDA. However, identifying a class with a phoneme – which is the class definition used in [5] – yields the least improvement. From these results we decided to associate a class with a phoneme segment (i.e. method C) since it performs equally well as associating a class with an elementary density, however the estimation of the transformation matrix is less time consuming.

Next we adjoined adjacent input frames to an augmented vector prior to the transformation. We could improve our results with a one-frame (63-component) vector by a three-frame splicing: 3 adjacent frames were used to form a (3*63)-component vector. Because of our choice of time differences (see Section 3) such an augmented vector covers a 90ms time window. After the transformation, we retained again 35 components. Table 2 shows that we could improve the error rate from 19.0% to 17.9%.

Table 2: *Word error rate (in %) for 3-frame splicing.*

Speaker	del	ins	tot
M-21	1.8	1.4	15.0
M-22	2.4	2.3	18.1
M-24	4.5	2.0	25.0
M-25	1.4	1.6	13.5
Average			17.9

5. RESULTS ON DARPA RESOURCE MANAGEMENT TASK

Unlike the German data base used in the last section, the DARPA RM1 task is characterized by a large vocabulary overlap between training and test data. This circumstance favors the use of highly specialized models which are able to memorize fine details of the training material but which might not have great generalization capability. Thus very good results had been obtained by using both intra-word and inter-word context-dependent phones, see e.g. [10].

The type of acoustic modeling used here is different. We keep the 47 basic context-independent phonemes and increase the acoustic resolution by allowing for more elementary densities per mixture. This method does neither increase the complexity of the recognition algorithm nor the cost of the search for the correct word sequence. It only increases the effort for the computation of the local log-likelihood of an acoustic event.

Table 3 presents the results for the speaker-dependent development test section of RM1 which comprises 12 speakers who spoke 100 sentences each, amounting to 10,242 words in total. The results given are for the no-grammar case.

Table 3: *Word error rate (in %) for speaker dependent DARPA RM1 task, no-grammar case. (3*63)-component input vector has been reduced by LDA to 35 components. Numbers are average over 12 speakers.*

	densities/ mixture	densities (total)	del	ins	tot
No LDA	30	4000	2.8	2.4	18.6
LDA	30	4000	1.6	1.7	13.2
LDA	150	20500	1.2	1.1	9.9

With the baseline system without LDA and an acoustic resolution of on the average 30 elementary densities per mixture the error rate was 18.6%. Employing LDA reduced the word error rate to 13.2%. Further significant improvement could be achieved by increasing the acoustic resolution to about 150 densities per mixture. An average error rate of 9.9% was obtained. This result is only 1.5% worse than the results published in [10], and it was achieved with a recognizer which does not use any explicit context-dependent modeling, neither within words nor across words.

6. CONCLUSIONS

Defining sub-phone units as classes to be discriminated in the LDA transform proved most effective for a continuous mixture density based speech recognizer. On a 12,000-word German task with small vocabulary overlap between training and test, the error rate was reduced by 18% by applying LDA. On the DARPA task with its closed vocabulary a reduction by one third was achieved. The currently best result using no context-dependent models is an average error rate of 9.9% for the speaker-dependent no-grammar case. Currently we are trying to improve our results by employing context-dependent phone models.

References

[1] P.F. Brown, *The acoustic-modeling problem in automatic speech recognition*, PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 1987.

[2] G.R. Doddington, *Phonetically sensitive discriminants for improved speech recognition*, Proc. ICASSP, pp. 556-559, Glasgow, Scotland, May 1989.

[3] M.J. Hunt, S.M. Richardson, D.C. Bateman, A. Piau, *An Investigation of PLP and IMELDA acoustic representations and of their potential for combination*, Proc. ICASSP, pp. 881-884, Toronto, Canada, May 1991.

[4] S.A. Zahorian, D. Qian, A.J. Jagharghi, *Acoustic-phonetic transformations for improved speaker-independent isolated word recognition*, Proc. ICASSP, pp. 561-564, Toronto, Canada, May 1991.

[5] G. Yu, W. Russell, R. Schwartz, J. Makhoul, *Discriminant analysis and supervised vector quantization for continuous speech recognition*, Proc. ICASSP, pp. 685-688, Albuquerque, NM, April 1990.

[6] L. Wood, D. Pearce, F. Novello, *Improved vocabulary-independent sub-word HMM modelling*, Proc. ICASSP, pp. 181-184, Toronto, Canada, May 1991.

[7] K. Fukunaga, *Introduction to statistical pattern recognition*, 2nd ed., Academic Press, 1990.

[8] H. Ney, *Acoustic modelling of phoneme units for continuous speech recognition*, Signal Processing V: Theories and Applications; L. Torres, E. Masgrau, M.A. Laguas (eds). Elsevier Science Publishers, pp. 65 - 72, 1990.

[9] P. Price, W. Fischer, J. Bernstein, D. Pallet, *A database for continuous speech recognition in a 1000-word domain*, Proc. ICASSP, pp.651-654, New York, April 1988.

[10] D. B. Paul, *The Lincoln tied-mixture HMM continuous speech recognizer*, Proc. ICASSP, pp. 329-332, Toronto, Canada, May 1991.