

# A Comparative Study of Some Discriminative Feature Reduction Algorithms on the AURORA 2000 and the DaimlerChrysler In-Car ASR Tasks

Joan Marí Hilario and Fritz Class

DaimlerChrysler Research and Technology  
P.O. Box 2360, 89013 Ulm, Germany  
joan.mari\_hilario@daimlerchrysler.com

## Abstract

A common practice in ASR to add contextual information is to append consecutive feature frames in a single large feature vector. However, this increases the processing time in the acoustic modelling and may lead to poorly trained parameters. A possible solution is to use a Linear Discriminant Analysis (LDA) mapping to reduce the dimensionality of the feature, but this is not optimal, at least in the case where the LDA classes are HMM-states. It is shown in this paper that the feature reduction problem is essentially a problem of approximating class posterior probabilities. These can be approximated using Neural Nets (NN). Some approaches using different choices for the classes and NN topology are presented and tested on the AURORA 2000 digit task and on our in-car task. Results on AURORA show a significant performance increase compared to LDA, but none of the NN-based approaches outperforms LDA on our in-car task.

## 1. Introduction

It is well known that ASR systems can take advantage of the context of a feature frame. Appending dynamic features and/or using a large context window of 5-10 consecutive feature frames is common practice among speech researchers. Since by doing so we are increasing the dimensionality of our feature space, the processing time in the acoustic modelling stage increases. In addition, the so-called *curse of dimensionality* [1] may cause poorly trained acoustic modelling parameters.

To solve both problems, feature dimensionality reduction methods can be used. The idea is to compute a mapping from a high-dimensional into a low-dimensional space that preserves some intrinsic or extrinsic characteristic of the speech signal. Depending on the nature of the characteristic preserved, the methods to compute the mapping can fall into the one of two categories: 1) Feature Reduction for Signal Representation, e.g. PCA, or 2) Feature Reduction for Classification, e.g. LDA, also known as Discriminative Dimensionality Reduction.

Since the ASR problem is mainly a classification problem, the methods belonging to the second category are preferred. In ASR the classes used to compute the LDA matrix are usually associated to the HMM states. To find those classes a segmentation into HMM states of the training database is used in order to assign an HMM state to each feature frame in the training set. The mean vectors and covariance matrices of the classes in the LDA problem are then found by simply computing the mean and covariance over the feature vectors belonging to the same HMM state (supervised clustering). The classes obtained are usually non-gaussian and have rather different covariance matrices which contradicts the usual assumption of equal co-

variance matrices used to obtain a closed-form solution for the LDA matrix. Therefore it could be interesting to find discriminative dimensionality reduction approaches that do not make such strong assumptions on the input data. As will be seen in the next section, this can be theoretically achieved by using Neural Networks (NN).

The next section of this paper is devoted to the fundamentals of discriminative dimensionality reduction and how a NN can be used for such a purpose. In Sec. 3 we describe the different dimensionality reduction approaches we experimented with. Finally, in Sec. 4 we present and discuss the results obtained on the AURORA 2000 database and on a subset of the UKKCP database, which is a DaimlerChrysler in-car speech recognition database.

## 2. Feature Reduction for Classification

The mappings belonging to this category are found in a *discriminative* way which means that the parameters of the mapping are computed in a way to keep the class posterior probabilities similar in the original and in the transformed space. It can be demonstrated [2] that the ideal features for classification are the posterior probabilities of the relevant *classes* for our classification or pattern recognition problem.

A first key problem of feature reduction for classification is therefore to define or find those classes relevant for the ASR problem in point. As already mentioned these classes are usually associated to the HMM states of our ASR system. However, for ASR tasks with a large number of states ( $> 100$ ), the use of such a large feature vector is prohibitive for the reasons already given in the introduction. A solution to this problem is to cluster similar HMM states until a reasonable number of clusters is obtained which are then used as target classes for the mapping  $C$ . Nevertheless, this procedure is not guaranteed to be optimal because it could well be that some important discrimination information is lost during the clustering procedure.

Another possibility is to use the following theorem [2] that states a general condition for a mapping  $C$  to be *discriminative* in the sense already explained.

**Theorem 1** Let  $C : \mathbb{R}^n \rightarrow \mathbb{R}^m$  and  $D : \mathbb{R}^m \rightarrow \mathbb{R}^k$  be two multi-dimensional mappings with  $m < k, n$ . Furthermore let  $\mathbf{x}$  and  $\mathbf{z}$  be the input vector and output vectors of the mapping  $C$  respectively, and  $\mathbf{y}$  the output vector of the mapping  $D \circ C$ . The mapping  $C$  is a discriminative mapping if and only if there is another mapping  $D$  such that

$$(D \circ C)(\mathbf{x}) = \mathbf{p}(\mathbf{x}) \quad (1)$$

with probability one, where the vector  $\mathbf{p}$  contains the posteriors of the classes  $p(C_i|\mathbf{x})$ .

The features used for classification are those at the output of the  $C$  mapping, i.e. the  $\mathbf{z}$  vector which has  $m < k$  dimensions.

To find the mapping  $C$  in practice we normally use a set of training input-output vector data pairs  $\{(\mathbf{x}^l, \mathbf{t}^l)\}$ , where  $\mathbf{t}$  is the desired output or target of the mapping  $D \circ C$ , obtained for instance by means of a forced-alignment. If we now let the target vectors  $\mathbf{t}$  adopt the 1-of- $c$  coding scheme (each component is associated to a class  $\mathcal{C}_i$  and per frame just one component of  $\mathbf{t}$  takes the value 1 and the others 0), it can be demonstrated [1] that the optimum mapping  $C$  in Theorem 1 may be theoretically found by approximating in the MSE sense the targets  $\mathbf{t}$  with the output of the mapping  $D \circ C$ .

However, the optimum solution for the mapping  $C$  can only be reached when the joint densities  $p(\mathbf{x}, \mathcal{C}_i)$  are known which is not the case in practice, since we only have a limited amount of training input-output vector data pairs [1]. Moreover, we usually do not know which is the actual functional form of the mappings so that we usually impose certain constraints on their functional form. Therefore the goodness of our approximation will also depend on the bias between the assumed functional form and the optimum mapping. A third fundamental issue is that the parameters of the assumed function must be optimized in order to converge to the appropriate minimum of the MSE cost function. That means that if an iterative algorithm is used to solve the mentioned approximation problem it must avoid local minima.

For this last reason, it has usually been preferred to make some assumptions on the joint densities  $p(\mathbf{x}, \mathcal{C}_i)$  (normality and heteroscedasticity) and on the mapping  $C$  (linear) in order to find a closed form solution, as in the LDA case. Since the heteroscedasticity assumption is rather strong some authors [3] have proposed the use of Heteroscedastic Discriminant Analysis (HDA). However those approaches are still based on the normality assumption, and the HDA does not have a closed form solution, i.e. it must be found iteratively. During the last decade, however, NN have been successfully applied in hybrid MLP/HMM ASR systems to estimate the posterior probabilities of the HMM states. Since for our problem at hand we must approximate posteriors as well, it seems logical to think that NN could also be successfully applied. Moreover, this approach does not make any assumption on the joint densities  $p(\mathbf{x}, \mathcal{C}_i)$ .

### 2.1. Neural Networks for Dimensionality Reduction

If we now represent the mappings  $C$  and  $D$  in Theorem 1 using a NN topology as shown in Fig. 1, it becomes clear that the NNs needed for our regression problem have a kind of bottle-neck topology and that they have at least one hidden layer.

To find the most suitable topology we must first decide which are the relevant classes for our ASR problem. If the number of classes is not very large, e.g. phonetic classes, the output of the NN can be directly used as a feature vector and the mapping  $D$  can therefore be removed. If on the contrary the number of classes is large, e.g. HMM state targets, both mappings must be jointly trained using for example the Error Back Propagation (EBP) algorithm. During recognition the mapping  $D$  is discarded and the output  $\mathbf{z}$  of  $C$  is used as a reduced feature vector and passed to the acoustic modelling block [4, 5]. Keeping in mind Theorem 1 we know that this output vector  $\mathbf{z}$  should theoretically be equivalent to the input vector  $\mathbf{x}$  for classification purposes.

Since the MSE criterion is not optimal to approximate the targets  $\mathbf{t}$  coded in a 1-of- $c$  scheme [1], the Minimum Cross Entropy (MCE) criterion is used instead. It can be demon-

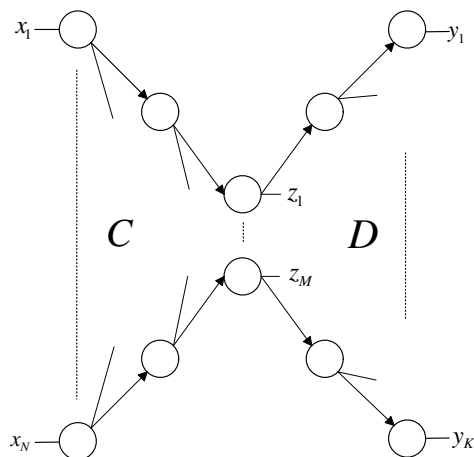


Figure 1: Representation of the mappings  $C$  and  $D$  in Theorem 1 using a Neural Network

strated [1] that the state posterior probabilities are also approximated by minimizing the MCE.

In the following section we describe some NN-based approaches with different choices for the target classes and the NN topology.

## 3. Approaches

In order to keep comparability high we used the same Feature Extraction and Acoustic Modelling for all the experiments with a database. In addition, the same baseline segmentations into HMM states of the training databases, i.e. the sets  $\{(\mathbf{x}^l, \mathbf{t}^l)\}$ , were used to train the feature reduction mapping except in the TANDEM approach where a segmentation into phonetic classes was used.

For the AURORA 2000 experiments, a 13 dimensional PLP feature vector with appended frame energy was used. The PLP features were then normalized to have zero mean and variance unity in order to have a simple initialization of the NN weights. On the AURORA 2000 task using the multi-condition training set we found in previous experiments that the results of PLP and MFCC features were not significantly different. For this reason and since the MFCC feature vectors for the UKKCP task were already pre-computed we used those features in the UKKCP experiments. The MFCC vector had 12 components plus normalized frame energy. In the experiments with NN, we also normalized the MFCC feature vector. The weights of the NN were trained to minimize the MCE criterion using the EBP algorithm implemented in the SPRACHcore package.

Our Acoustic Modelling uses Semi-Continuous HMMs with a code-book of full-covariance gaussians trained in a supervised way [6]. The HMMs used in the AURORA 2000 experiments are 11 whole-word models. A silence and a one-state pause models were added to deal with the long silences and the inter-digit pauses respectively. The total number of states was 127. By contrast the HMMs used for the UKKCP task are a combination of whole-word models (for digits, spelled letters and very frequent words), context-independent and context-dependent models. This resulted in a total of 1496 states. As in the previous task, garbage models and pause model were also

introduced.

No language modelling or grammar (not even a forced silence model at both ends of the utterances) was used in the AURORA recognition experiments.

In our experiments with NN we used the pre-nonlinearity outputs of the mapping  $C$  in order to have features more suited to the gaussian densities of our acoustic modelling [7].

### 3.1. Linear Discriminant Analysis (LDA)

In our experiments with LDA, we used a context of 9 PLP feature frames to construct a vector of 117 components which was then input to the LDA transform. The LDA transform itself was extracted from the 117 dimensional covariance matrices of each HMM-state which were obtained using the segmentation into states mentioned at the beginning of the section. A total of 32 LDA coefficients were computed for each input frame.

### 3.2. TANDEM Approach (TAN)

In the TANDEM approach [7] the targets of the NN are chosen to be the phonetic classes present in our training set. For the AURORA 2000 task this resulted in a 24 dimensional feature output vector, whereas for our UKKCP task, it resulted in a 56 dimensional feature vector. The input vector to the NN was constructed by appending a context of 9 feature frames, each constructed by appending the 1st and 2nd derivatives to the static PLP feature vector. This resulted in an input vector of 351 components. The NN used was an MLP with 351 input units, one hidden layer of 480 and an output layer with as many units as phonetic classes in the task. The units in the hidden layer were sigmoidal units whereas those in the output layer used a softmax non-linearity.

### 3.3. TANDEM Clustering of States (TAN-CoS)

In another set of experiments, we used the same configuration as the one described for the TANDEM system, but instead of using phonetic targets we used a set of clusters obtained by clustering the HMM states. Those clusters were obtained using the clustering algorithm proposed by Lee [8] to merge context-dependent HMMs. The HMM states were then clustered until a sufficiently low number of state clusters was reached. The number of clusters was 24 in the AURORA experiments and 32 in the UKKCP. The topology of the NN used was the same as in the previous approach except for the number of units in the output layer, which was set to the number of clusters.

### 3.4. TANDEM with State Targets and PCA (TAN-PCA)

As suggested in [9], an alternative to clustering could be to train the NN using the HMM states as targets and then apply a linear feature reduction transform (PCA) to the high-dimensional output of the NN. However, this method is certainly not optimum in a theoretical sense. Nevertheless, we tried this approach on the AURORA 2000 database since in this case the training of the NN is not very time consuming. The NN topology was the same as in the previous points except for the number of output units which was set to the number of HMM states. The targets of the NN were the 127 HMM states coded using the 1-of-c scheme. After training the NN in the same way as in the previous cases, the data in the training set was input to the net to generate a set of 127 dimensional vectors. This data set was then used to compute an PCA matrix to reduce the dimensionality to 32 components. Afterwards this 32 dimensional vector was passed to the acoustic modelling.

Approach	test set			mean
	testa	testb	testc	
LDA32	10.4	15.6	14.3	13.3
TAN-PCA24	8.2	12.9	11.1	10.7
TAN-CoS24	8.3	12.3	11.2	10.5
TAN24	8.2	12.5	10.4	10.3
NLDA24	7.8	12.8	11.5	10.5

Table 1: Mean percent Word Error Rate (WER) results over 4 SNR levels (0dB, 10dB, 20dB and clean) and all noise types for the three tests sets of AURORA

Approach	SNR				mean
	clean	20dB	10dB	0dB	
LDA32	2.4	3.6	7.8	39.2	13.3
TAN-PCA24	1.4	2	5.9	33.4	10.7
TAN-CoS24	1.4	2.3	5.7	32.6	10.5
TAN24	1.3	1.7	5.1	33.2	10.3
NLDA24	1.1	2.4	6.1	32.5	10.5

Table 2: Mean percent WER results over all noise types for different SNR levels of AURORA

### 3.5. Non-linear Discriminant Analysis (NLDA)

Finally, the last method used a NN with two hidden layers in a bottle-neck topology as used by Asoh [4] and by Fontaine [5]. As already mentioned in Sec. 2.1, the idea is to use a NN with 2 hidden layers to approximate the state posterior probabilities and to use the output of the 2nd hidden layer as NLDA features. The NN targets to train the weights were chosen to be the HMM states coded in the usual 1-of-c scheme. The topology of the NN used had 351 input units and 480 sigmoidal units in the 1st hidden layer. In the 2nd hidden layer, 24 sigmoidal units were used for the AURORA experiments and 32 in the UKKCP experiments. The number of softmax units in the output layer was equal to the number of states in the acoustic modelling.

## 4. Experiments

### 4.1. AURORA 2000 Experiments

The AURORA 2000 task is a noisy digit recognition task in English which is derived from the TI-DIGITS database. Noise in this database were artificially added to the speech signal. The 8440 files in the multi-condition set of the AURORA 2000 database were used to train the neural nets and HMMs of the approaches described in the previous section. The total number of training vectors was approximately 1,500,000. To test the approaches, only 4 of the 7 possible SNR levels (0dB, 10dB, 20dB and clean) were used which resulted in a total of 16016 test files over the three test sets of the database.

In Table 1 we can compare the mean performance of the described approaches on each of the test sets of the AURORA database. As expected, all the approaches using a non-linear mapping to reduce the feature dimensionality are significantly better than the linear mapping (LDA). This difference is similar for all test sets which seems to suggest that our NN-based approaches are quite robust to mismatches between training and test conditions. Nevertheless, the performances of the different NN-based approaches are quite similar.

In Table 2 the results at different SNR levels are displayed. As in the previous table, we see that the NN-based approaches have similar performances over different SNR levels. In ad-

Approach	test set			Mean
	Digits	Spelling	Cities	
LDA32	3.6	12.0	16.5	6.3
TAN-CoS32	3.8	14.2	21.4	7.1
TAN56	4.6	13.8	21.8	7.6
NLDA32	4.6	13.7	21.2	7.5

Table 3: Mean percent WER for all the test sets of the UKKCP database

dition, the NN-based approaches are superior to the LDA approach for all the SNR levels.

Although these results agree with our theoretical expectations, they probably cannot be solely attributed to a better match between the statistics of the classes and the feature reduction mapping. One can argue that the context of the input vector in the NN approaches is larger than in the LDA case because the 1st and 2nd derivatives had been appended to each frame. This may have contributed to the overall performance improvement, but not as much as the better match between mapping and class statistics.

#### 4.2. UKKCP database experiments

The UKKCP speech database was recorded in real car environments, and therefore noise has not been artificially added to the speech signal. The speakers were male and female native Britons who uttered a few long sentences, digit strings, spelled words, city names and commands useful for in-car applications, e.g. "radio on". A total 32107 speech files were used for training which once parameterized resulted in nearly 10.000.000 training vectors. The number of training vectors is larger than in the previous task in order to avoid undertraining of the NN and HMM parameters. The test set consisted of 3894 utterances spoken by 16 women and 19 men. The test lexicon contained 2827 words and was comprised of digit strings, spelled words and city names.

In contrast to the AURORA 2000 experiments none of the NN-based feature reduction approaches was better than the LDA system (Table 3). The difference was especially large in the "Cities" test set where context-dependent acoustic modelling was used. This is quite disappointing given the excellent results obtained on the previous database. The reasons may be different for each of the NN approaches tested. In the TANDEM with Clustering of States (TAN-CoS) case, it could well be that the clustering procedure (mentioned in Sec. 3.3) of the 1496 states removed some important classification information, since very different HMM states may have been clustered together. The TAN56 approach may have failed for the same reasons given already in [9], since the 56 phonetic targets probably increase the overlap between the context-dependent states of the acoustic modelling. Although we hoped that the NLDA approach would overcome this problems, our obtained results do not show any significant improvement compared to the TANDEM approach. However, it may be that in the NLDA approach the EBP gets stuck in a local minimum owing to the high-dimensional output space (1496 outputs). Alternatively, it could be that the topology with 2 hidden layers is not well suited to this problem, since we assume that there is a linear mapping between state posteriors and the reduced feature vector. As already mentioned in Sec. 2, this assumption may introduce a large bias between the actual and the optimum output of the NN. A possible solution would be to introduce a 3rd hidden layer between

the output layer and the layer that generates the reduced features. Another cause of the disappointing results could be due to the the measure used during cross-validation to avoid overfitting of the NN. This measure was the number of frames correctly classified which is suitable for problems where the net is used as a classifier but not for a more complex regression problem such as ours [2]. Whatever the causes are, it is clear that it is not possible to simply transfer the successful set-up of the AURORA NN-based experiments to more complex ASR tasks where context-dependent acoustic modelling is used.

## 5. Conclusions

We have shown that the discriminative feature reduction problem can be viewed as a regression problem, whereby the HMM state or class posterior probabilities are approximated. Some different neural net based approaches have been proposed to estimate these posteriors. Tests on the AURORA 2000 task show significant improvement compared to the classical LDA approach. Unfortunately, none of the approaches was able to improve the performance of LDA in the experiments with our in-car speech database UKKCP.

## 6. Acknowledgements

This work was funded by the EU in the framework of the SP-HEAR and RESPITE projects.

## 7. References

- [1] Christopher M. Bishop. *Neural networks for pattern recognition*. Oxford University Press, 1996.
- [2] L. Devroye, L. Györfy, and G. Lugosi. *A Statistical Theory of Pattern Recognition*. Springer Verlag, 1996.
- [3] G. Saon, M. Padmanabhan, R. Gopinath, and S. Chen. Maximum likelihood discriminant feature spaces. In *Proc. ICASSP'2000*, Istanbul.
- [4] T. Kurita, H. Asoh, and N. Otsu. Nonlinear discriminant features constructed by using outputs of multilayer perceptron. In *Proceedings of the IEEE International Symposium on Speech, Image Processing and Neural Networks*, pages 417–420, April 1994.
- [5] V. Fontaine, C. Ris, and J.M. Boite. Nonlinear discriminant analysis for improved speech recognition. In *Proc. EUROSPEECH'97*, volume 4, pages 2071–2074, Rhodes, 1997.
- [6] F. Class, A. Kaltenmeier, and P. Regel-Brietzmann. Optimization of an hmm-based continuous speech recognizer. In *Proc. EUROSPEECH'93*, 1993.
- [7] S. Kajarekar P. Jain S. Sharma, D. Ellis and H. Hermansky. Feature extraction using non-linear transformation for robust speech recognition on the aurora database. In *Proc. ICASSP'2000*, Istanbul.
- [8] K.F. Lee. Context-dependent phonetic hidden markov models for speaker independent continuous speech. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 1990.
- [9] D.P.W. Ellis, R. Singh, and S. Sivasdas. Tandem acoustic modelling in large-vocabulary recognition. In *Proc. ICASSP'2001*, Salt Lake City.