IMPROVED FEATURE DECORRELATION FOR HMM-BASED SPEECH RECOGNITION

Kris Demuynck, Jacques Duchateau, Dirk Van Compernolle[*] *and Patrick Wambacq*

K.U.Leuven - ESAT - PSI, Kardinaal Mercierlaan 94, B-3001 Heverlee, Belgium
E-mail: Kris.Demuynck@esat.kuleuven.ac.be

ABSTRACT

In most HMM-based recognition systems, a mixture of diagonal covariance gaussians is used to model the observation density functions in the states. The use of *diagonal* covariance gaussians however assumes that the underlying data vectors have uncorrelated vector components: if each gaussian is replaced with its full covariant counterpart, the off-diagonal elements in the covariance matrices should be small. To that end, most recognition systems have some kind of decorrelation matrix near the end of the preprocessing. Examples are the inverse cosine transform used with cepstral coefficients, and principal component analysis (PCA) or linear discriminant analysis (LDA) of the features. However, none of these transforms is optimal if it comes to reducing the mismatch introduced by setting the off-diagonal elements in the covariance matrices to zero.

The algorithm described in this paper reduces the local correlations between feature vector components inside the gaussians with a single global linear transform at the end of the preprocessing stage. The algorithm is optimal in the sense that we calculate the linear transformation that minimises the sum of the square of all off-diagonal elements over all gaussians.

The algorithm is compared with principal component analysis, linear discriminant analysis and the recently published maximum likelihood modelling for semi-tied covariance matrices. The decorrelation method is also evaluated on two speech recognition tasks. A significant relative improvement was achieved in both cases.

1. INTRODUCTION

In many speech recognition systems, the observation density functions are modelled as mixtures of diagonal covariance gaussians. These mixtures of gaussians are however only approximations of the real distributions. One of the approximations is the assumption that the off-diagonal elements of the covariance matrices of the gaussians are close to zero. To that end, most recognition systems have some kind of parameter decorrelation near the end of the preprocessing. Examples are the inverse cosine transform used with cepstral transformations, and principal component analysis (PCA) or linear discriminant analysis (LDA) of the features. None of these transforms are however designed in an optimal way as to minimise the magnitude of the off-diagonal elements in the covariance matrices. The algorithm we propose is optimal in the sense that we calculate the linear transformation that minimises the magnitude of the off-diagonal elements in the covariance matrices over all gaussians with a least-squares method.

The remainder of the text is organised as follows. First the decorrelation algorithm is explained in detail. Next, the algorithm is compared with some existing alternatives. Finally, the method is evaluated on two speech recognition tasks, and some remarks are given.

2. ALGORITHM

As mentioned above, we search for a single linear transformation of the acoustic features that minimises the average of the square of the off-diagonal elements over a large set of covariance matrices. To compensate for a possible scaling of the axes, the off-diagonal elements are normalised with respect to the diagonal elements. Thus, what is actually minimised is a weighted sum of the square of the correlation coefficients between the parameters, and this simultaneously over all gaussians.

Let $\mu^{(m)}$ be the mean and $\Sigma^{(m)}$ the *full* covariance matrix of gaussian m with $\Sigma_{ij}^{(m)}$ the component on row i and column j. And let $N^{(m)}$ be the number of points assigned to gaussian m with $N = \sum N^{(m)}$ the total number of points in the training data and $\lambda^{(m)} = N^{(m)}/N$ the weight of the gaussian. We then have to find a transformation matrix A that minimises the following quantity:

$$\sum_m \lambda^{(m)} \sum_{i \neq j} \left(\frac{\tilde{\Sigma}_{ij}^{(m)}}{\sqrt{\tilde{\Sigma}_{ii}^{(m)} \tilde{\Sigma}_{jj}^{(m)}}} \right)^2 \qquad (1)$$

with

$$\tilde{\Sigma}^{(m)} = A \Sigma^{(m)} A^T$$

This quantity can be optimised with numerical techniques, e.g. by decomposing the transformation matrix A into a product of basic transformations of the form $(I + \delta_{ij})$ with I the identity matrix and δ_{ij} a matrix equal to zero except for element (i,j).

The optimisation problem is strongly simplified if the normalisation with respect to the variance is omitted. As to limit the

[*] Lernout & Hauspie Speech Products.

mismatch between the quantity that has to be minimised (function 1) and the approximation without the normalisation, some pre-compensation is done: the data space is first transformed so that the average covariance matrix $\Sigma_{\text{av}} = \sum \lambda^{(m)}\Sigma^{(m)}$ equals the identity matrix (multiplying with the transpose of the eigenvectors, followed by a proper scaling). This makes that the normalisation terms in formula 1 are close to one and thus can be omitted. To prevent scaling of axes, the rest of the transformation (on top of the pre-compensation) is limited to the class of orthonormal transformations (rotations). For the optimisation, we therefore decompose the remainder of the transformation in elementary rotations R_{ij} (Givens rotations).

$$R_{ij} = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & \cos(\theta) & & \sin(\theta) & \\ & & \ddots & & \\ & -\sin(\theta) & & \cos(\theta) & \\ & & & & \ddots \\ 0 & & & & 1 \end{bmatrix} \begin{array}{c} \\ \\ \leftarrow i \\ \\ \leftarrow j \\ \\ \\ \end{array}$$

Expanding formula 1 for an elementary Givens rotation R_{ij} (omitting the normalisation) shows that the optimal θ can be found by minimising the following quantity:

$$a\sin^2(2\theta) + 4b\cos^2(2\theta) + 2c\sin(2\theta)\cos(2\theta)$$

with

$$\begin{aligned} a &= \sum_m \lambda^{(m)}(\Sigma_{ii}^{(m)} - \Sigma_{jj}^{(m)})^2 \\ b &= \sum_m \lambda^{(m)}(\Sigma_{ij}^{(m)})^2 \\ c &= 2\sum_m \lambda^{(m)}(\Sigma_{ii}^{(m)} - \Sigma_{jj}^{(m)})\Sigma_{ij}^{(m)} \end{aligned}$$

The simplified optimisation algorithm thus consists of the following steps:

1 Do the pre-compensation, and update the covariance matrices $\Sigma^{(m)}$.

2 For every j, and for every $i < j$, determine the optimal rotation θ and update the covariance matrices $\Sigma^{(m)}$ and the transformation matrix A.

3 Repeat step 2 until convergence (e.g. less than 0.05% relative improvement on the quantity that is to be minimised).

The advantage of the simplified version is its fast convergence (5 iterations typically) at a low cost per iteration. And although the simplified version does not provide the optimal transformation, the resulting decorrelation matrix is close to the optimum and can eventually be used as starting point for a full optimisation.

The transformed set of diagonal covariance gaussians to be used in the HMM system can be easily derived from the set of full covariance matrices $\Sigma^{(m)}$ and means $\mu^{(m)}$.

It is also possible to split the set of gaussians in the HMM system in two or more logical subsets, and have a decorrelation matrix for every subset. The decorrelation matrices then can be seen as a shared full covariance matrix for the subset of gaussians, while the diagonal covariance gaussians in the subset still allow for an additional scaling of the axes. This is similar to the semi-tied covariance matrices presented in [4]. When multiple decorrelation matrices are used, a normalisation of the matrices is needed to prevent an arbitrary scaling of the likelihood of the gaussians over the different subsets. More information on semi-tied covariance matrices and on the scaling can be found in [4, 5].

3. COMPARISON WITH OTHER METHODS

3.1. Principal component analysis

Principal component analysis [1] results in a global decorrelation of the features. This means that the global covariance matrix of the features is a diagonal matrix, or after proper scaling the identity matrix. A global decorrelation however does not provide for local decorrelation of the data inside the gaussians in the system. This is clearly shown in figure 1. And although there is no perfect solution for the situation presented in figure 1, a simple redefinition of the axes can reduce the number of gaussians that cannot be modelled correctly with a diagonal covariance from 4 to 1, and that is by close approximation what the proposed decorrelation algorithm will do.

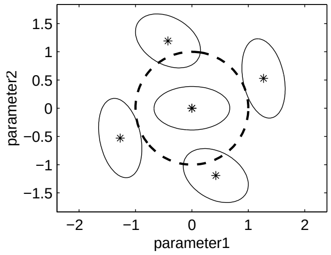

Figure 1: A global decorrelation of the parameter space is not sufficient to decorrelate the data on the gaussian level. Each ellipse is the contour-line of a full covariance gaussian. The dashed line represents the global correlation of all data.

3.2. Linear discriminant analysis

Linear discriminant analysis [3] is widely used to reduce a large set of features to a smaller set, with minimal loss in performance. One of the side effects of LDA is a 'whitening‘ (decorrelation) of the average within class covariance matrix. If the classes are defined to be the gaussians, then the 'whitening' operation will ensure that the average covariance matrix Σ_{av} as defined in section 2, has a diagonal form. However, only the *average* covariance matrix is optimised. No optimisation over all gaussians is done, which may result in a poor decorrelation after all on the gaussian level. This is shown in the left upper corner of figure 2.

3.3. Maximum likelihood optimisation

The recently developed maximum likelihood modelling for semi-tied covariance matrices [4, 5, 6] searches for the linear transformation A that maximises the likelihood of the training data (evaluated in the original, non transformed domain) when the data is modelled with diagonal covariance matrices in a transformed domain. Therefore, following expression has to be maximised:

$$|A|^N \prod_m \left|\text{diag}(A\Sigma^{(m)}A^T)\right|^{-\frac{N^{(m)}}{2}} \tag{2}$$

It is easy to show that any transformation of the form $(I + \delta_{ij})$ as defined in section 2, which decreases the amplitude of an element $\Sigma_{ij}^{(m)}$, also increases the value of $\left|\text{diag}(\Sigma^{(m)})\right|$ (note that $|I + \delta_{ij}| = 1$). The net effect of maximising function 2 is thus also one of minimising a weighted average of the amplitude of the off-diagonal elements in the covariance matrices. Main difference with the least-squares solution as presented in this paper is the cost assigned to every non-zero off-diagonal element. Figure 2 shows the effect of the maximum likelihood optimisation and the full and simplified least-squares optimisation. The slight differences between the three methods are due to the different interpretation of how bad a certain off-diagonal element is. Note that if there is a perfect solution to the problem (all off-diagonal elements equal to zero) it will be found by all three methods.

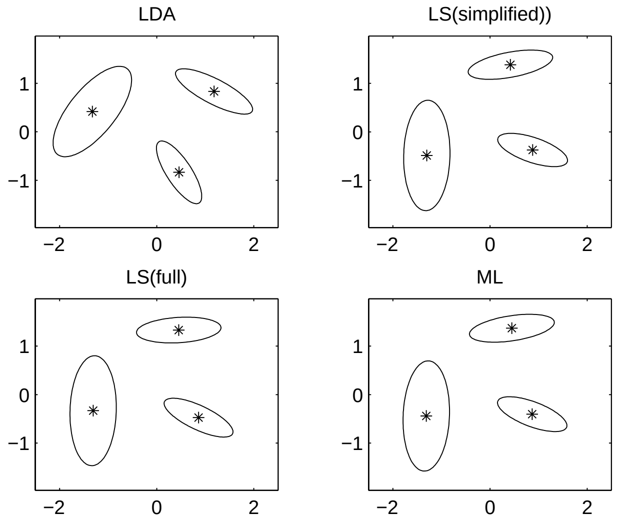

Figure 2: Comparison between three methods to decorrelate the gaussians. Starting point for all three methods is the LDA-transformation (left-upper corner).

4. EXPERIMENTS & RESULTS

A first set of experiments was conducted on the ARPA Resource Management task. We investigated the behaviour of the decorrelation in function of the number of gaussians in the system, the number of acoustic parameters and the type of preprocessing. For all situations, 46 3-state context-independent phone models were trained on the standard SI-109 training set, using our semi-continuous HMM system [2]. Next, full covariance matrices are calculated for all gaussians in the system, based on which the decorrelation matrix is calculated. The decorrelation matrix is then added at the end of the preprocessing stage and an updated set of gaussians is derived from the full covariance matrices calculated in the previous step. Finally, two extra iterations of Viterbi-training are done to adapt all parameters in the system to the new preprocessing.

Three different types of preprocessing were investigated. All preprocessings start from a set of 24 mean-normalised log filter-bank outputs, and the mean-normalised log of the energy.

- **Preprocessing1** transforms the filter-bank outputs to cepstral parameters (inverse cosine transform), replaces cep[0] with the log energy and selects the first 6, 9 or 13 parameters. Finally, this set is augmented with the first and second time derivatives (eventually the last two second derivatives are omitted).
- **Preprocessing2** does an LDA-transform on the filter-bank outputs and selects the 6, 9 or 13 first parameters. Finally, this set is augmented with the first and second time derivatives (eventually the last two second derivatives are omitted).
- **Preprocessing3** does an LDA-transform on the filter-bank outputs and selects the 6, 9 or 13 first parameters. Next, 5 consecutive frames are stacked and a second LDA is performed, resulting in 16, 25 or 39 parameters.

The LDA-transforms in the described preprocessings are used to reduce the size of the feature set, not to decorrelate. So the HMM states are used as classes, not the gaussians.

In a first set of preliminary tests, the simplified optimisation was compared with the full optimisation on a small set of experiments. Both methods performed equally well on average, but since the full optimisation is conceptually more correct, the remaining experiments were done with the full optimisation.

#params.	16	25	39	16	25	39	16	25	39
	preproc. 1			preproc. 2			preproc. 3		
#gauss.	average WER without the extra decorrelation step								
3018	9.10	7.62	7.30	7.98	7.22	6.99	8.66	6.89	6.65
4527	8.41	7.02	6.82	7.55	6.34	6.61	8.15	6.74	6.16
7243	8.57	6.84	6.12	6.94	6.16	6.27	7.96	6.44	6.16
#gauss.	average WER with the extra decorrelation step								
3018	8.62	6.98	6.28	7.98	6.74	6.30	7.94	6.72	6.37
4527	8.38	6.50	6.31	7.51	6.25	5.98	7.74	6.45	5.92
7243	8.28	6.33	5.78	6.91	5.75	6.08	7.56	6.34	5.84
#gauss.	relative improvement (percentage)								
3018	5.3	8.4	14.0	0.0	6.6	9.9	8.3	2.5	4.2
4527	0.4	7.4	7.5	0.5	1.4	9.5	5.0	4.3	3.9
7243	3.4	7.5	5.6	0.4	6.7	3.0	5.0	1.6	5.2

Table 1: Results on the ARPA RM test-sets (feb89 + oct89 + feb91 + sep92) with the standard Word Pair Grammar (context-independent modelling).

Table 1 gives an overview of the results. The use of the decorrelation algorithm always results in a significant relative reduction on the error rate, so none of the proposed preprocessings does a good job in decorrelating the features on the gaussian level. The relative improvement also shows to be larger when more input parameters and/or less gaussians are used. This is in line with expectations, since larger feature-sets in general show more correlation, while the introduction of more gaussians reduces the complexity of the volume a single gaussian has to model.

A second set of experiments was conducted on the Wall Street Journal November 92 task, using context-dependent models. The baseline system used for the task is a gender-independent cross-word triphone tied-state reduced semi-continuous HMM system [2]. The HMM contains 20254 gaussians in total, with which 10436 states are modelled resulting in 33169 distinct cross-word triphones. For the preprocessing, we selected preprocessing 1 with 39 parameters. The results are shown in table 2. The decorrelation gives about 10% reduction in word error rate, irrespective of the method being used: least-squares as presented in this paper, or maximum likelihood.

train-set	2-gram 5k	2-gram 20k	3-gram 5k	3-gram 20k
	WER without the extra decorrelation step			
WSJ0	6.61	13.34	4.09	11.13
WSJ0+1	4.99	11.39	3.25	9.16
	WER with the extra decorrelation step			
WSJ0	5.70	12.62	3.38	10.07
WSJ0+1	4.50	10.46	2.62	8.54
	relative improvement (percentage)			
WSJ0	13.8	5.4	17.4	9.5
WSJ0+1	9.8	8.2	19.4	6.8
	WER with maximum likelihood modelling [4]			
WSJ0+1	4.58	10.56	2.82	8.52
	WER with the simplified decorrelation algorithm			
WSJ0+1	4.46	10.63	2.71	8.51

Table 2: Results on the ARPA WSJ-nov92-nvp test-sets, using the official bigram and trigram language models (context-dependent modelling).

A comparison between the decorrelation matrices for the WSJ-task shows that the maximum likelihood optimisation and full least-squares optimisation result in almost identical matrices: the one matrix can be mapped on the other by multiplying with an almost identity matrix (a 0.01 perturbation on the elements on average). The difference between the matrices obtained with the simplified and the full least-square optimisation is about seven times bigger. This is still a remarkable small difference given the fact that the simplified version only has half the amount of parameters to optimise compared to the full optimisation.

The actual structure of the decorrelation matrix is shown in figure 3. The block structure indicates that the static, delta and delta-delta features are almost uncorrelated on the gaussian level. Experiments where the time derivatives were replaced with another inverse cosine transform (a two dimensional IDCT-transform on the log filter bank outputs) showed far more correlation between static and dynamic features. The use of derivatives is thus close to optimal if it comes to decorrelating the features.

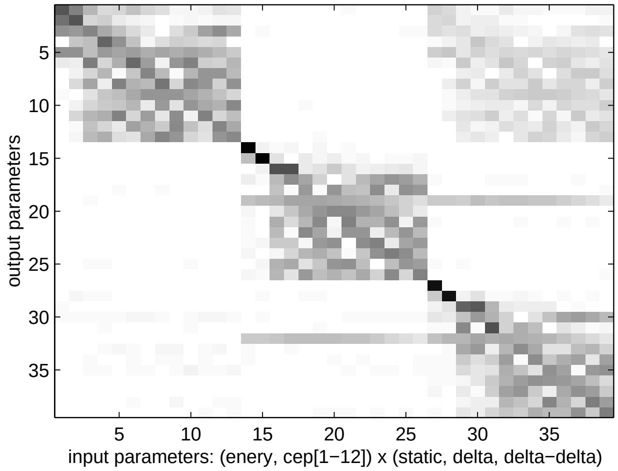

Figure 3: The magnitude of the elements in the transformation matrix (WSJ0+1). White corresponds to zero.

5. CONCLUSIONS

This paper showed that an extra decorrelation of the input features on the gaussian level can result in substantially better acoustic models when mixtures of diagonal covariance gaussians are used to model the observation density functions. It also showed the equivalency between the recently published maximum likelihood modelling for semi-tied covariance matrices and the least-squares decorrelation approach used in this paper. We believe that parameter normalisations like the one presented in this paper will become more and more important in the near future when more experimental feature sets like acoustic parameters are to be used.

6. REFERENCES

1. N. Ahmed and K. R. Rao. *Orthogonal Transforms for Digital Signal Processing*, pages 200–205. Springer-Verslag, 1975.
2. J. Duchateau, K. Demuynck, and D. Van Compernolle. Fast and accurate acoustic modelling with semi-continuous HMMs. *Speech Communication*, 24(1):5–17, July 1998.
3. R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*, pages 114–121. John Wiley & Sons, 1973.
4. M. Gales. Semi-tied covariance matrices. In *Proc. ICASSP*, volume II, pages 657–660, Seattle, May 1998.
5. R. Gopinath. Maximum likelihood modeling with gaussian distributions for classification. In *Proc. ICASSP*, volume II, pages 661–664, Seattle, May 1998.
6. N. Kumar. *Investigation of Silicon-Auditory Models and Generalization of LDA for Improved Speech Recognition*. PhD thesis, John Hopkins Univ., 1997.