SEMI-TIED COVARIANCE MATRICES

*M.J.F. Gales**

Cambridge University Engineering Department
Trumpington Street
Cambridge CB2 1PZ
England
Email: mjfg@watson.ibm.com

ABSTRACT

A standard problem in many classification tasks is how to model feature vectors whose elements are highly correlated. If multi-variate Gaussian distributions are used to model the data then they must have full covariance matrices to accurately do so. This requires a large number of parameters per distribution which restricts the number of distributions that may be robustly estimated, particularly when high dimensional feature vectors are required. This paper describes an alternative to full covariance matrices in these situations. An approximate full covariance matrix is used. The covariance matrix is now split into two elements, one full and one diagonal, which may be tied at completely separate levels. Typically, the full elements are extensively tied, resulting in only a small increase in the number of parameters compared to the diagonal case. Thus dramatically increasing the number of distributions that may be robustly estimated. Simple iterative re-estimation formulae for all the parameters within the standard EM framework are presented. On a large vocabulary speech recognition task a 10% reduction in word error rate over a standard system was achieved.

1. INTRODUCTION

In many pattern recognition applications there is a need to model data that is highly correlated. One such application is speech recognition using continuous-density HMMs and is the one considered in this paper. Multi-variate Gaussian distributions are used to model the data associated with each state. When correlations within the feature vector are explicitely modeled using full covariance matrices, a large number of parameters per component have to be estimated. This limits the number of components, hence states, that may be robustly trained. Alternatively assumptions about the correlations present in the data may be made. This allows a block-diagonal covariance matrix to be used, slightly reducing the number of parameters. Finally diagonal covariance matrices may be used. Here there are few parameters associated with the covariance matrix. However, correlations in the feature vectors cannot be modeled. To overcome this problem multiple diagonal-covariance Gaussian distributions may be used. In addition to being able to model non-Gaussian distributions they can model correlations between elements of the feature vector. Unfortunately only a limited number of components may be robustly estimated. If there were some way of effectively decorrelating the data associated with a particular state, or group of states, improved speech recognition performance should be possible. This led to the development of *semi-tied covariance* matrices[1].

Semi-tied covariance matrices may be seen as a natural extension of the state-specific rotation scheme of [7]. Instead of estimating the decorrelating transform independently of the specific components associated with it, the transform is estimated in a maximum-likelihood (ML) fashion given the current model parameters. This form of covariance modeling was first introduced in [5]. Alternatively rather than viewing them as an extension of the standard covariance matrices, they may also be viewed as a ML feature-space transformation. It is thus a technique for simultaneously optimising both the parameters and the feature-space. In contrast to other schemes that have addressed this problem, efficient re-estimation formulae are given. These may either be run in a memory or time efficient fashion. This scheme is related to a recent extension of linear discriminant analysis [6]. It is shown that the same efficient re-estimation formulae may also be used for this problem.

2. SEMI-TIED COVARIANCE MATRICES

Semi-tied covariance matrices [5] are a simple extension to the standard diagonal, block-diagonal, or full covariance matrices used with HMMs. Instead of having a distinct covariance matrix for every component in the recogniser, each covariance matrix consists of two elements, a component specific diagonal covariance element, $\mathbf{\Sigma}_{\text{diag}}^{(m)}$, and a *semi-tied* class dependent, non-diagonal matrix, $\mathbf{H}^{(r)}$. The form of the covariance matrix is

$$\mathbf{\Sigma}^{(m)} = \mathbf{H}^{(r)}\mathbf{\Sigma}_{\text{diag}}^{(m)}\mathbf{H}^{(r)T} \tag{1}$$

$\mathbf{H}^{(r)}$ may be tied over a set of components, for example all those associated with the same state of a particular context-independent phone. Possible clustering schemes are discussed in [5].

Each component, m, has the following parameters: component weight $c^{(m)}$, component mean $\mu^{(m)}$ and the diagonal element of the semi-tied covariance matrix $\mathbf{\Sigma}_{\text{diag}}^{(m)}$. In addition it is associated with a semi-tied class, which has an associated matrix $\mathbf{H}^{(r)}$. This is used to generate the component's covariance matrix as described in equation 1. It is very complex to optimise these parameters directly so an expectation-maximisation approach is

*The author is currently at IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY10598. The majority of the work presented here was performed whilst the author was funded as a Research Fellow at Emmanuel College, Cambridge.

[1] Originally referred to as semi-tied full-covariance matrices, an even more cumbersome name!

adopted [1]. Furthermore, rather than dealing with $\mathbf{H}^{(r)}$, it is simpler to deal with its inverse, $\mathbf{A}^{(r)}$ [5], thus $\mathbf{A}^{(r)} = \mathbf{H}^{(r)-1}$. If ML estimates of all the parameters are made then the auxiliary function below must be optimised with respect to $\mathbf{A}^{(r)}$ (ignoring terms that are independent of $\mathbf{A}^{(r)}$)

$$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = \sum_{\tau, m \in M^{(r)}} \gamma_m(\tau) \log \left(\frac{|\mathbf{A}^{(r)}|^2}{|\text{diag}\left(\mathbf{A}^{(r)} \mathbf{W}^{(m)} \mathbf{A}^{(r)T}\right)|} \right) \quad (2)$$

where

$$\mathbf{W}^{(m)} = \frac{\sum_\tau \gamma_m(\tau) \left(\mathbf{o}(\tau) - \mu^{(m)}\right) \left(\mathbf{o}(\tau) - \mu^{(m)}\right)^T}{\sum_\tau \gamma_m(\tau)} \quad (3)$$

$$\mu^{(m)} = \frac{\sum_\tau \gamma_m(\tau) \mathbf{o}(\tau)}{\sum_\tau \gamma_m(\tau)} \quad (4)$$

and $\gamma_m(\tau) = p(q_m(\tau) | \mathcal{M}, \mathbf{O}_T)$ where $q_m(\tau)$ indicates component m at time τ, $\mathbf{O}_T$ is the complete set of training data and $\mathbf{o}(\tau)$ is the observation at time τ. $M^{(r)}$ is the subset of components that share the same full covariance element $\mathbf{A}^{(r)}$ or *semi-tied* class. The diagonal element of the covariance matrix is given by

$$\mathbf{\Sigma}^{(m)}_{\text{diag}} = \text{diag}\left(\mathbf{A}^{(r)} \mathbf{W}^{(m)} \mathbf{A}^{(r)T}\right) \quad (5)$$

where $m \in M^{(r)}$. The re-estimation formulae for the component weights and transition probabilities are identical to the standard HMM cases [8].

Unfortunately optimising equation 2 is non-trivial and requires numerical optimisation techniques and a full matrix, $\mathbf{W}^{(m)}$, to be stored at each component. An alternative approach is to use a variational optimisation scheme. Fortunately an exact variational optimisation scheme is possible here, with meaningful variational parameters. The set of component specific diagonal variances, given the current estimate of $\mathbf{A}^{(r)}$, $\hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}} = \left\{ \hat{\mathbf{\Sigma}}^{(m)}_{\text{diag}}, m \in M^{(r)} \right\}$, are used as the variational parameters. Now after some re-arranging and selecting a particular row of $\mathbf{A}^{(r)}$, $\mathbf{a}^{(r)}_i$, (this is a 1 by n *row vector*)

$$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}}) = \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \left\{ \log \left(\left(\mathbf{a}^{(r)}_i \mathbf{c}^T_i \right)^2 \right) - \log \left(|\hat{\mathbf{\Sigma}}^{(m)}_{\text{diag}}| \right) - \sum_j \frac{\left(\mathbf{a}^{(r)}_j \hat{\mathbf{o}}^{(m)}(\tau) \right)^2}{\sigma^{(m)2}_{\text{diag}_j}} \right\} \quad (6)$$

where $\hat{\mathbf{o}}^{(m)}(\tau) = \mathbf{o}(\tau) - \mu^{(m)}$, $\hat{\sigma}^{(m)2}_{\text{diag}_i}$ is element i of the leading diagonal of $\hat{\mathbf{\Sigma}}^{(m)}_{\text{diag}}$ and $\mathbf{c}_i$ is the i^{th} row vector of the cofactors of $\mathbf{A}^{(r)}$. It can be shown that[2]

$$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) - n\beta \geq \mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}}) \quad (7)$$

[2] This uses the equality that at the ML estimate of the mean and diagonal variance the following *minimum* value is obtained

$$\sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) \sum_i \frac{\left(\mathbf{a}^{(r)}_i \hat{\mathbf{o}}^{(m)}(\tau) \right)^2}{\sigma^{(m)2}_{\text{diag}_i}} = n \sum_{m \in M^{(r)}, \tau} \gamma_m(\tau) = n\beta$$

with equality when diagonal elements of the covariance matrix are given by equation 5. Optimising $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}})$ is itself non-trivial, however an efficient iterative solution is possible [4] (an alternative scheme is given in [5]). It is shown that

$$\mathbf{a}^{(r)}_i = \mathbf{c}_i \mathbf{G}^{(ri)-1} \sqrt{\left(\frac{\sum_{m \in M^{(r)}} \sum_\tau \gamma_m(\tau)}{\mathbf{c}_i \mathbf{G}^{(ri)-1} \mathbf{c}^T_i} \right)} \quad (8)$$

where

$$\mathbf{G}^{(ri)} = \sum_{m \in M^{(r)}} \frac{1}{\hat{\sigma}^{(m)2}_{\text{diag}_i}} \mathbf{W}^{(m)} \sum_\tau \gamma_m(\tau) \quad (9)$$

This optimisation is iterative on a row by row basis, since each row is related to the other rows by the cofactors. Thus each row is optimised given the current estimate of all the other rows. However the sufficient statistics for this optimisation are very simple, namely $\mathbf{G}^{(ri)}$ and the semi-tied class occupancy count. Thus once these are collected the optimisation is a function of the number of semi-tied classes and the dimensionality, *not* the number of components in the system.

$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}})$ can now be optimised. All the component specific diagonal covariance elements may then be updated, yielding a new set $\hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}}$ and process repeated. At each iteration the likelihood is guaranteed to increase. This optimisation process may be run in one of two distinct modes.

1. **Time efficient**: At each component, the occupancy, vector sum and $\mathbf{W}^{(m)}$ is stored[3]. It is then possible to optimise $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}})$ iteratively without having to examine the data again. At each optimisation of $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}})$, $\hat{\mathbf{\Sigma}}^{(m)}_{\text{diag}}$ is estimated using the current estimate of $\mathbf{A}^{(r)}$ (computational cost $\mathcal{O}(n^3)$ per component). Then $\mathbf{G}^{(ri)}$ is found (computational cost $\mathcal{O}(n^3)$ per component) and finally the transform estimated (computational cost $\mathcal{O}(n^4)$ per semi-tied class). In terms of computational cost this may be contrasted with standard numerical optimisation schemes. These will usually require the calculation of the current gradient, an operation costing at least $\mathcal{O}(n^3)$ per component for every iteration in the optimisation. Though the variational scheme may be more expensive per iteration in practice it converges after very few iterations (< 10). In contrast the numerical optimisation scheme may take an order of magnitude more iterations. Furthermore each iteration is guaranteed to increase the likelihood for the variational scheme. Hence there are no stability problems.
2. **Memory efficient**: For many large vocabulary speech recognition tasks it is not practical to store $\mathbf{W}^{(m)}$ for every component. To get around this problem the model parameters are estimated in two separate runs through the data. On the first run occupancy counts and vector sum for each component, and $\mathbf{G}^{(ri)}$ for each semi-tied class estimated using the current values of the diagonal elements of the covariance matrix. Given that in many applications there will be very few (< 100) semi-tied classes compared to the number of components (> 10000) the memory cost of storing $\mathbf{G}^{(ri)}$ is very small. It is then possible to optimise $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}; \hat{\mathbf{\Sigma}}^{(r)}_{\text{diag}})$. On the next pass through the data the

[3] This ignores the transition probability updates.

standard HMM estimation statistics (though the variance is estimated after applying $\mathbf{A}^{(r)}$) are stored and the diagonal elements of the covariance matrix may be updated. Thus many runs through the data are required, at each run $\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}})$ is not maximised, however the likelihood is always guaranteed to increase. This form of memory efficient optimisation is not possible with standard numerical optimisation schemes.

One of the major advantages of semi-tied covariances over component specific full and block-diagonal covariance matrices is their computational efficiency during recognition. The likelihood calculation is based on

$$\mathcal{L}\left(\mathbf{o}(\tau); \mu^{(m)}, \mathbf{\Sigma}^{(m)}\right) = \mathcal{N}\left(\mathbf{o}^{(r)}(\tau); \mathbf{A}^{(r)}\mu^{(m)}, \mathbf{\Sigma}^{(m)}_{\text{diag}}\right) + \frac{1}{2}\log\left(|\mathbf{A}^{(r)}|^2\right) \quad (10)$$

where $m \in M^{(r)}$. Thus by storing $\mathbf{A}^{(r)}\mu^{(m)}$ instead of $\mu^{(m)}$ the cost of calculating the likelihoods associated with semi-tied covariance matrices is that of one matrix vector multiplication per-semi-tied class and an addition. If only one semi-tied class is used then $\log\left(|\mathbf{A}^{(r)}|^2\right)$ does not discriminate between the models so may be ignored. Also, this addition may be removed if desired by appropriately scaling $\mathbf{A}^{(r)}$ and $\mathbf{\Sigma}^{(m)}_{\text{diag}}$ [6].

3. RELATIONSHIP TO HLDA AND ML VARIANCE ADAPTATION

Though the re-estimation formulae presented here are applied to a specific problem, similar approaches may be used in other estimation problems.

Heteroscedastic linear discriminant analysis (HLDA) [6] is related to semi-tied covariance matrices. HLDA is a generalisation of the standard linear discriminant analysis (LDA) scheme [2], which relaxes the assumption that all the within class covariance matrices are the same. The transform is again required to reduce the dimensionality from an initial n-dimensional space to a p-dimensional space in a ML fashion. The objective function optimised is[4] [6]

$$\mathcal{Q}(\mathcal{M}, \hat{\mathcal{M}}) = \sum_{\tau, m \in M} \gamma_m(\tau)\left\{-\log\left(\left|\text{diag}\left(\mathbf{A}_p\mathbf{W}^{(m)}\mathbf{A}_p^T\right)\right|\right) + \log\left(|\mathbf{A}|^2\right) - \log\left(\left|\text{diag}\left(\mathbf{A}_{n-p}\mathbf{T}\mathbf{A}_{n-p}^T\right)\right|\right)\right\} \quad (11)$$

where

$$\mathbf{T} = \frac{1}{T}\sum_{\tau}\left(\mathbf{o}(\tau) - \mu^{(g)}\right)\left(\mathbf{o}(\tau) - \mu^{(g)}\right)^T \quad (12)$$

$\mu^{(g)}$ is the global mean of the data, $\mathbf{A}_p$ is the first p rows of $\mathbf{A}$ and $\mathbf{A}_{n-p}$ are the remaining $n-p$ rows[5]. An identical variational transform is used as equation 6, except now a modified set of variational parameters are used

$$\hat{\sigma}^{(m)2}_{\text{diag}_j} = \begin{cases} \mathbf{a}_j\mathbf{W}^{(m)}\mathbf{a}_j^T & (j \le p) \\ \mathbf{a}_j\mathbf{T}\mathbf{a}_j^T & (j > p) \end{cases} \quad (13)$$

[4] The dependence on the semi-tied class has been dropped as there is typically only one semi-tied class.

[5] It is worth emphasising that this is not the ML projection down to p dimensions due to the final term in equation 11. If the ML projection is to be used then a similar technique may be applied, however the second term in equation 11 is replaced by $\log\left(|\mathbf{A}_p\mathbf{A}_p^T|\right)$, requiring some modification to the re-estimation formulae.

Since the variational optimisation scheme optimises the rows of $\mathbf{A}$ one at a time it is simple to see that the same variational optimisation scheme may be directly applied to optmising this case. Solutions to the case when full covariance matrices are used is also possible [3].

Another closely related problem is ML linear transformations of the variances for speaker and environmental adaptation [4]. Here a transformation, typically tied over many components, is required to adapt the means and variances. When adapted in an *unconstrained* fashion [4] the variance may have the form

$$\hat{\mathbf{\Sigma}}^{(m)} = \mathbf{H}\mathbf{\Sigma}^{(m)}\mathbf{H}^T \quad (14)$$

When $\mathbf{H}$ is estimated in a ML fashion it results in an identical expression to the variational function, equation 6, other than the mean estimate which is now typically based on a linear transform of the mean parameters. It can therefore be optimised in the same fashion without the need to update the variational parameters as these are usually fixed in the adaptation task[6]. At recognition time the same efficient decoding as the semi-tied models may be used.

4. EXPERIMENTS AND RESULTS

An initial investigation of the use of semi-tied covariance matrices was carried out on a large-vocabulary speaker-independent continuous-speech recognition task. All recognition experiments were carried out on the 1994 ARPA Hub 1 data. The H1 task is an unlimited vocabulary task with approximately 15 sentences per speaker. This data was recorded in a clean[7] environment.

4.1. Recognition System

The baseline system used for the recognition task was a gender-independent cross-word-triphone mixture-Gaussian tied-state HMM system. This was the same as the "HMM-1" model set used in the HTK 1994 ARPA evaluation system [9]. The speech was parameterised into 12 MFCCs, C_1 to C_{12}, along with normalised log-energy and the first and second differentials of these parameters. This yielded a 39-dimensional feature vector, to which cepstral mean normalisation was applied. The acoustic training data consisted of 36493 sentences from the SI-284 WSJ0 and WSJ1 sets, and the LIMSI 1993 WSJ lexicon and phone set were used. The standard HTK system was trained using decision-tree-based state clustering to define 6399 speech states. The number of components per-state was increased using *mixing-up* [10] until there were 12 components in each speech state. This standard model-set will be referred to as *Standard*. For the H1 task a 65k word list and dictionary was used with the trigram language model described in [9]. All decoding used a dynamic-network decoder.

Two semi-tied covariance systems were investigated. For both systems $\mathbf{A}^{(r)}$ was constrained to be block-diagonal, with separate blocks for the static, delta and delta-delta elements of the feature vector. All components of the same context-independent phone were clustered together into the same semi-tied class, approximately an additional 25 thousand parameters in a system of 6 million parameters for the 12-component per state case. Both systems had 12-components per speech state with the same state clustering as the standard system. Due to memory constraints the memory-efficient estimation scheme was used in each case.

[6] These linear adaptation schemes have been successfully applied to adapting semi-tied models [3]

[7] Here the term "clean" refers to the training and test conditions being from the same microphone type with a high signal-to-noise ratio.

1. **System 1**: This was built from scratch using mixing-up. The procedure was as follows starting with the single component standard system: perform two iterations of Baum-Welch updating the means and diagonal covariance elements; update the full covariance element; perform two iterations of Baum-Welch updating the means and diagonal covariance elements; mix-up and repeat as required. This scheme allows full advantage of the semi-tied full-covariance matrices to be made. This model set will be referred to as *Semi-Tied (1)*.
2. **System 2**: Here the standard 12-component system is used as the initial model. Using this model set the full element of the covariance matrix is estimated and then two iterations of Baum-Welch updating the means and diagonal covariance elements was performed. This model set will be referred to as *Semi-Tied (2)*.

4.2. Results

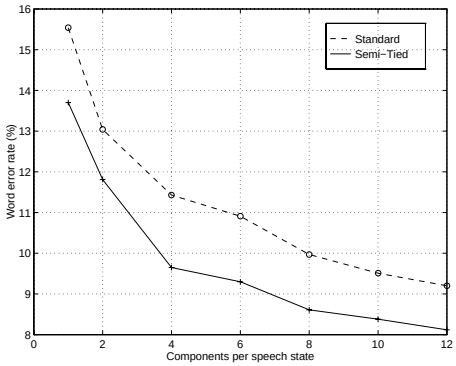

Figure 1: Performance of *Semi-Tied (1)* on the H1 Evaluation data

Figure 1 shows the recognition performance of *Semi-Tied (1)* on the H1 Evaluation data at the various stages of the mixing-up process. At all stages the use of semi-tied matrices, though introducing very few additional parameters, shows a marked improved in recognition performance. In particular the performance of the standard 12-component system was achieved with around half the number of parameters. These results may be compared with state-specific rotations [7]. A separate rotation was generated for each context-independent state, three times the number of semi-tied transforms. Slight gains were observed when the system had few components. Using the single component and 2-component systems error rates of 14.25% and 12.85% respectively were obtained compared to 15.54% and 13.04% for the standard system. However, as the number of components per state increased the performance of the two systems became about the same.

Table 1 shows the baseline performance of the three systems. As previously noted [5], semi-tied systems can give around 10% reduction in word error rate when trained from scratch, *Semi-Tied (1)*, and around 5% when trained as a "second" pass, *Semi-Tied (2)*.

System	Error Rate (%)		
	H1 Dev	H1 Eval	Average
Standard	9.57	9.20	9.38
Semi-Tied (1)	8.62	8.12	8.36
Semi-Tied (2)	9.00	8.59	8.78

Table 1: Baseline and semi-tied covariance matrices results on H1 development and evaluation data [5]

5. CONCLUSIONS

This paper has presented an extension to the standard covariance matrices used with HMMs, or more generally any Gaussian classifier. Simple and efficient re-estimation formulae are presented, which may be run in either a memory or time efficient fashion depending on the nature of the model being used. In terms of recognition performance on large vocabulary speech recognition task a 10% reduction in word error rate may be obtained with minimal increase in the number of model parameters or recognition time. Furthermore the same recognition performance as the standard system may be obtained with about half the number of model parameters.

6. REFERENCES

[1] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[2] K Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, 1972.

[3] M J F Gales. Adapting semi-tied full-covariance matrix hmms. Technical Report CUED/F-INFENG/TR298, Cambridge University, 1997. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[4] M J F Gales. Maximum likelihood linear transformations for HMM-based speech recognition. Technical Report CUED/F-INFENG/TR291, Cambridge University, 1997. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[5] M J F Gales. Semi-tied full-covariance matrices for hidden Markov models. Technical Report CUED/F-INFENG/TR287, Cambridge University, 1997. Available via anonymous ftp from: svr-ftp.eng.cam.ac.uk.

[6] N Kumar. *Investigation of Silicon-Auditory Models and Generalization of Linear Discriminant Analysis for Improved Speech Recognition*. PhD thesis, John Hopkins University, 1997.

[7] A Ljolje. The importance of cepstral parameter correlations in speech recognition. *Computer Speech and Language*, 8:223–232, 1994.

[8] L R Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, February 1989.

[9] P C Woodland, J J Odell, V Valtchev, and S J Young. The development of the 1994 HTK large vocabulary speech recognition system. In *Proceedings ARPA Workshop on Spoken Language Systems Technology*, pages 104–109, 1995.

[10] S J Young, J Jansen, J Odell, D Ollason, and P Woodland. *The HTK Book (for HTK Version 2.0)*. Cambridge University, 1996.