OPTIMAL LINEAR FEATURE TRANSFORMATIONS FOR SEMI-CONTINUOUS HIDDEN MARKOV MODELS

E. Günter Schukat-Talamazzini, Joachim Hornegger, Heinrich Niemann

Lehrstuhl für Mustererkennung (Informatik 5)
Friedrich-Alexander-Universität Erlangen-Nürnberg
Martensstraße 3
D–91058 Erlangen, F.R. of Germany
E-mail: {schukat,hornegger,niemann}@informatik.uni-erlangen.de

ABSTRACT

Linear discriminant or Karhunen-Loève transforms are established techniques for mapping features into a lower dimensional subspace. This paper introduces a uniform statistical framework, where the computation of the optimal feature reduction is formalized as a Maximum-Likelihood estimation problem. The experimental evaluation of this suggested extension of linear selection methods shows a slight improvement of the recognition accuracy.

1. INTRODUCTION

It is the ultimate goal of any probabilistic approach to speech recognition to capture the entire process of word production within *one single* homogeneous model. The statistical parameters of this model then can be optimized with respect to a large training sample of speech using standard parameter estimation techniques.

A first step in this direction was the introduction of the (discrete density) *hidden Markov model* (HMM, [1]), which constitutes a simple probabilistic description of acoustical word realization down to the level of labelled speech frames. Frame labelling was performed outside the scope of the word models by a phonetic classifier or a vector quantizer.

The undesirable exclusion of the feature level from word modelling stopped with the advent of *semi-continuous* models (SCHMM) [2] which incorporated the formerly external vector quantizer into the speech production model. Thus, in a SCHMM-based speech recognizer the entire processing sequence from the feature vectors to the word level is part of the global HMM, and its free statistical parameters can be jointly optimized by the Baum-Welch algorithm.

Unfortunately, the extraction of features from the speech wave now, as before, stays outside the probabilistic framework, and the development of the signal processing component is still subject to heuristics and intuition. In order to turn, at least a part of, the feature extraction stage into a parameterized and trainable building block of the recognizer, a decomposition of the process is assumed: (1) initially, the signal is transformed into a sequence of high-dimensional short-time feature vectors $\boldsymbol{x}_t \in \mathbb{R}^D$; (2) the $\boldsymbol{x}_t$'s are mapped to vectors $\boldsymbol{y}_t$ of drastically reduced dimension $d \ll D$. In the context of speech recognition, the first stage can be thought of as computation of the log power mel-spectral coefficients, followed by the enlargement of the resulting short-time parameter vector using first and second order temporal derivatives [3], or differences, or simply the spectral coefficients of an appropriate number of neighboring speech frames. The second stage is formed by a linear mapping involving standard feature extraction activities like, for instance, the cosine transform [4] and rotations of the coordinate system (Karhunen-Loève or linear discriminant transform, [5, 6]); finally, the features belonging to the d first coordinate axes are selected to be fed into the vector quantizer.

The basic idea presented in this paper is to incorporate the feature rotation step together with the subsequent dimensionality reduction ($D \rightarrow d$) into the SCHMM formalism. Given the model structure (including the target dimension d) and a representative training sample, the particular transformation matrix will be optimized with respect to the likelihood function of the overall model, resulting in the so-called *maximum likelihood rotation* (MLR).

The rest of the paper is organized as follows: the feature transform HMM is defined in Sect. 2. In Sect. 3, we will provide the derivation of Baum-Welch training formulae for the extended model; particularly, we show that the rotation-dependent part can be separated from the conventional SCHMM part of the Kullback-Leibler function. Sect. 4 is devoted to the of numerical optimization of the MLR expression as a function of the unknown rotation parameters.

2. FEATURE-TRANSFORM MARKOV MODELS

The basic idea underlying the *feature transform HMM* (FTHMM) is illustrated in Figure 1; we want to re-

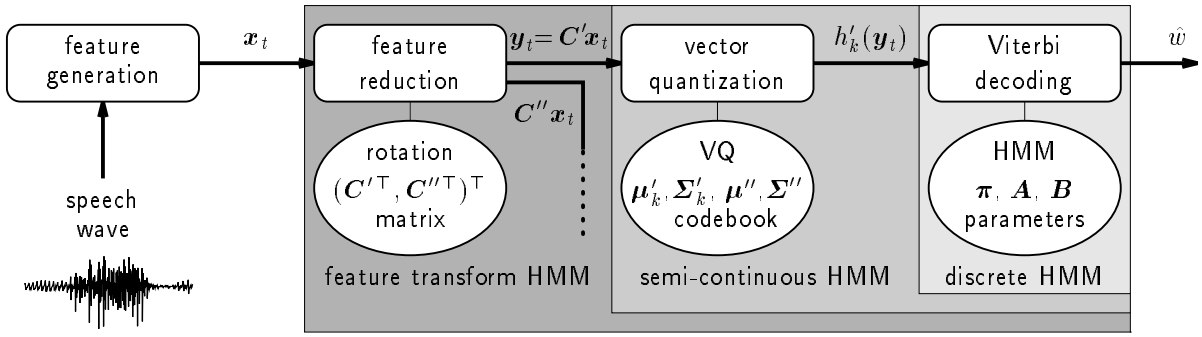

Figure 1: Different HMM-based speech recognizer architectures

place the ordinary semi-continuous HMM by the enlarged model

$$\lambda \;=\; (\pi, A, B, g, C) \quad ,$$

with initial probabilities π, transition probabilities A, mixture coefficients B, a parameterized codebook g of mixture components, and the feature transformation matrix C. The crucial issue of the FTHMM is that λ is designed to model the production of sequences of the original feature vectors $x_t \in \mathbb{R}^D$; this is opposite to the SCHMM which generates the reduced vectors $y_t \in \mathbb{R}^d$ (see Figure 1).

If the mapping $x_t \mapsto y_t$ is assumed to be singular — which is true, of course, if the dimension of the feature space is actually reduced — no inverse exists, and no probabilistic model generating the x_t can be found. This problem is circumvented by adopting a *"soft"* variant of feature space reduction.

$$b_j(x) \;=\; \sum_{k=1}^{K} b_{jk} g_k(x) \;=\; h''(C''x) \cdot \sum_{k=1}^{K} b_{jk} h'_k(C'x)$$

The k-th semi-continuous mixture component g_k of our model is factorized into

- a *codebook-class dependent* distribution h'_k operating on the most discriminative part $C'x \in \mathbb{R}^d$ of x and
- a *codebook-class independent* distribution h'' operating on the uninformative noise part $C''x \in \mathbb{R}^{D-d}$ of x.

The rows of both C' and C'' together form the *orthonormal* $D \times D$ matrix $C = (C'^\top, C''^\top)^\top$, i.e., the product $CC^\top$ equals the identity matrix I_D. The initial components $y_1, \ldots, y_d$ of the rotated vector Cx should carry the relevant information for speech recognition. The remaining coefficients $y_{d+1}, \ldots, y_D$ will not contribute to word or sentence identification since their probability $h''(C''x)$ may be factored out as a constant term in the above expression for $b_j(x)$. Consequently, the computation of $h''(C''x)$ becomes obsolete, and only the class-conditional probabilities $h'_k(C'x)$ — recall that d is usually much smaller than D — have to be provided during Viterbi decoding.

3. BAUM-WELCH REESTIMATION

Let $X = x_1, x_2, \ldots, x_T$ be a sequence of D–dimensional real valued feature vectors for estimating the parameters. In classical HMMs the EM algorithm is used for solving the incomplete data estimation problem: it is hidden which state sequence produces the observable output data. Let $q = q_1, \ldots, q_T$ be the sequence of state indices from an N-state model. Dealing with semi-continuous HMMs, we need furthermore a specification of the mixture components. Let $k = k_1, k_2, \ldots, k_T$ denote a sequence of mixture density components. The pair (q, k) represents that at time t the state s_{q_t} uses the density component g_{k_t} for generating an output vector.

The estimation of the FTHMM parameters is performed iteratively using the EM Algorithm [7]. For that purpose we have to maximize the Kullback-Leibler statistics

$$Q(\lambda, \hat{\lambda}) \;=\; \sum_{q} \sum_{k} P(X, q, k \mid \lambda) \cdot \log P(X, q, k \mid \hat{\lambda})$$

with respect to $\hat{\lambda}$. After expanding the log likelihood expression of the right hand side of the above equation and reordering the summation $Q(\lambda, \hat{\lambda})$ separates into three well-known terms (see, for instance, [2, p. 160]) depending exclusively on π, A, and B, respectively. The remaining expression

$$Q_g(\lambda, \hat{\lambda}) \;=\; \sum_{k=1}^{K} \sum_{t=1}^{T} \left(\sum_{j=1}^{N} \zeta_t(j,k) \right) \cdot \log \hat{g}_k(x_t)$$

has to be maximized in order to get reestimates of the codebook parameters; the *a posteriori* expectations

$$\zeta_t(j,k) \;=\; P(q_t = j, k_t = k \mid X, \lambda)$$

are obtained through the classical forward-backward computations [2].

For a standard SCHMM, multivariate Gaussian mixture components $g_k(x) = \mathcal{N}(x | \mu_k, \Sigma_k)$ with mean vectors μ_k and covariance matrices Σ_k are assumed, and maximization of $Q_g(.,.)$ results in the well-known ML estimates

$$\hat{\mu}_k = \sum_{t=1}^{T} \frac{z_t(k)}{Z_k} \cdot x_t \;, \quad \hat{\Sigma}_k = \sum_{t=1}^{T} \frac{z_t(k)}{Z_k} \cdot (x_t - \hat{\mu}_k)(x_t - \hat{\mu}_k)^\top$$

for the semi-continuous codebook; we used the abbreviations $z_t(k) = \sum_j \zeta_t(j,k)$ and $Z_k = \sum_t z_t(k)$.

However, in the case of FTHMM mixtures

$$g_k(\boldsymbol{x}) \;=\; \mathcal{N}(\boldsymbol{C}'\boldsymbol{x} \mid \boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k) \cdot \mathcal{N}(\boldsymbol{C}''\boldsymbol{x} \mid \boldsymbol{\mu}'', \boldsymbol{\Sigma}'')$$

with the means and covariances $\boldsymbol{\mu}'_k, \boldsymbol{\Sigma}'_k$, $\boldsymbol{\mu}''$, and $\boldsymbol{\Sigma}''$ of the rotated Gaussians h'_k and h'' the separation of $Q_{\boldsymbol{g}}(.,.)$ becomes much more intricate because of the linear coupling between the partial matrices $\boldsymbol{C}'$ and $\boldsymbol{C}''$ of the underlying feature transform.

Fortunately, the Kullback-Leibler function easily decomposes into $K+1$ independent subexpressions

$$Q_{\boldsymbol{g}}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) \;=\; \sum_{k=1}^{K} Q_{h'_k}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}}) \;+\; Q_{h''}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$$

if we assume $\boldsymbol{C}'$ and $\boldsymbol{C}''$ fixed for a moment. Setting the partial derivatives of the terms $Q_{h'_k}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ to zero and solving for the unknown distribution parameters, we obtain the reestimates

$$\hat{\boldsymbol{\mu}}'_k \;=\; \boldsymbol{C}'\hat{\boldsymbol{\mu}}_k \quad \text{and} \quad \hat{\boldsymbol{\Sigma}}'_k \;=\; \boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k\boldsymbol{C}'^{\top} \;, \quad k = 1, \ldots, K,$$

where $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$ are the means and covariances of the standard SCHMM codebook distributions. Following analogous arguments, maximization of the likelihood function for the global FTHMM distribution h'' leads to the equations

$$\hat{\boldsymbol{\mu}}'' \;=\; \boldsymbol{C}''\hat{\boldsymbol{\mu}} \;, \quad \hat{\boldsymbol{\mu}} \;=\; \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t$$

as well as

$$\hat{\boldsymbol{\Sigma}}'' \;=\; \boldsymbol{C}''\hat{\boldsymbol{\Sigma}}\boldsymbol{C}''^{\top} \;, \quad \hat{\boldsymbol{\Sigma}} \;=\; \frac{1}{T}\sum_{t=1}^{T}(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_t - \hat{\boldsymbol{\mu}})^{\top}$$

In order to solve the estimation problem for the rotation parameters $\boldsymbol{C}'$ and $\boldsymbol{C}''$ we have to plug in the estimates derived above into our target function $Q_{\boldsymbol{g}}(\boldsymbol{\lambda}, \hat{\boldsymbol{\lambda}})$ which yields the expression

$$\sum_{k=1}^{K} \underbrace{\left(\sum_{t=1}^{T} z_t(k) \cdot \log \mathcal{N}(\boldsymbol{C}'\boldsymbol{x}_t \mid \boldsymbol{C}'\hat{\boldsymbol{\mu}}_k, \boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k\boldsymbol{C}'^{\top}) \right)}_{E_k}$$

$$+ \sum_{t=1}^{T} \log \mathcal{N}(\boldsymbol{C}''\boldsymbol{x}_t \mid \boldsymbol{C}''\hat{\boldsymbol{\mu}}, \boldsymbol{C}''\hat{\boldsymbol{\Sigma}}\boldsymbol{C}''^{\top}) \;.$$

Since the weights $z_t(k)$ of the kth expectation term E_k are identical to the weights involved in the estimation of $\hat{\boldsymbol{\mu}}_k$ and $\hat{\boldsymbol{\Sigma}}_k$, the equation

$$E_k \;=\; -\frac{Z_k}{2} \cdot \left(\log |2\pi \boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k\boldsymbol{C}'^{\top}| - d \right)$$

can be shown to be valid. A quite similar argument holds for the right hand side of $Q_{\boldsymbol{g}}(.,.)$. After division by a factor $-T/2$ and elimination of additive terms irrelevant to the optimization problem at hand we are left with the expression

$$\ell(\boldsymbol{C}', \boldsymbol{C}'') \;=\; \sum_{k=1}^{K} \left(P_k \cdot \log |\boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k\boldsymbol{C}'^{\top}| \right) + \log |\boldsymbol{C}''\hat{\boldsymbol{\Sigma}}\boldsymbol{C}''^{\top}|$$

where the weights P_k denote the ratio Z_k/T. We conclude that in order to obtain the desired MLR matrix $\boldsymbol{C} = (\boldsymbol{C}'^{\top}, \boldsymbol{C}''^{\top})^{\top}$ the log likelihood function $\ell(\boldsymbol{C}) = \ell(\boldsymbol{C}', \boldsymbol{C}'')$ has to be *minimized* subject to the orthonormality restrictions $\boldsymbol{C}\boldsymbol{C}^{\top} = \boldsymbol{I}_D$.

4. OPTIMIZATION OF THE MLR MATRIX

Unfortunately, the function ℓ presents a delicate structure, including log determinants of non-quadratic matrix products. Moreover, the situation is complicated by the implicit coupling of $\boldsymbol{C}'$ and $\boldsymbol{C}''$ based on the orthonormality constraint of $\boldsymbol{C}$. This fact prevents us from applying a straightforward gradient descent algorithm, controlled by the partial derivatives

$$\boldsymbol{H}' \;=\; \nabla_{\boldsymbol{C}'}\, \ell(\boldsymbol{C}) \;=\; \sum_{k=1}^{K} \left[P_k \cdot \left(\boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k\boldsymbol{C}'^{\top} \right)^{-1} \boldsymbol{C}'\hat{\boldsymbol{\Sigma}}_k \right]$$

and

$$\boldsymbol{H}'' \;=\; \nabla_{\boldsymbol{C}''}\, \ell(\boldsymbol{C}) \;=\; \left(\boldsymbol{C}''\hat{\boldsymbol{\Sigma}}\boldsymbol{C}''^{\top} \right)^{-1} \boldsymbol{C}''\hat{\boldsymbol{\Sigma}}$$

(see [8, p.568]). In order to overcome the latter complication, our optimization task has to be transformed towards an unrestricted problem. Our crucial device for that enterprise is the observation that every rotation $\boldsymbol{C}$ of $\mathbb{R}^D$ may be decomposed into a product

$$\boldsymbol{C}(\boldsymbol{\phi}) \;=\; \boldsymbol{C}(\phi_1, \ldots, \phi_R) \;=\; \prod_{r=1}^{R} \boldsymbol{U}_{p_r q_r}(\phi_r)$$

of *elementary rotations* of the special form

$$\boldsymbol{U}_{pq}(\phi) = \boldsymbol{I}_D + \cos\phi(\boldsymbol{e}_p\boldsymbol{e}_p^{\top} + \boldsymbol{e}_q\boldsymbol{e}_q^{\top}) + \sin\phi(\boldsymbol{e}_p\boldsymbol{e}_q^{\top} - \boldsymbol{e}_q\boldsymbol{e}_p^{\top})$$

where $\boldsymbol{e}_p$ denotes the pth unit vector and the index pairs (p_r, q_r) range over all combinations $1 \le p < q \le D$, i.e., $R = (D-1) \cdot D/2$. From the geometrical point of view, $\boldsymbol{U}_{pq}(\phi)$ describes a rotation of $\mathbb{R}^D$ in the (p,q)-plane by the angle ϕ. Henceforth, two basically different approaches are possible, according to whether the free parameters ϕ_r shall be improved independently from another or not.

In the former case we proceed quite similar to the Jacobian coordinate descent algorithm for eigenvector computation. Iteratively, a declining sequence $\boldsymbol{C}_0, \boldsymbol{C}_1, \ldots$ of feature transformations

$$\boldsymbol{C}_0 \;=\; \boldsymbol{I}_D \;, \quad \boldsymbol{C}_s \;=\; \boldsymbol{C}_{s-1} \cdot \boldsymbol{U}_{p_s q_s}(\phi_s)$$

is created subject to the requirement $\ell(\boldsymbol{C}_s) \le \ell(\boldsymbol{C}_{s-1})$. The task of generating an improved rotation in step s, then, reduces to a one-dimensional minimization problem in the variable ϕ_s. For the control of some standard descent algorithm computing the partial derivative of $\ell(\boldsymbol{C}_s)$ with respect ϕ_s is required. The total derivative results in

$$\frac{\partial \ell(\boldsymbol{C}_s)}{\partial \phi_s} \;=\; \sum_{i=1}^{D}\sum_{j=1}^{D} \frac{\partial \ell(\boldsymbol{C}_s)}{\partial (C_s)_{ij}} \cdot \frac{\partial (C_s)_{ij}}{\partial \phi_s} \;.$$

Due to the fact that the rotations depend on each other

the required number of iteratively applied elementary rotations will generally exceed R.

If, however, the entries of the parameter vector $\boldsymbol{\phi}$ representing the transformation matrix are to be jointly optimized, i.e., a successor $\boldsymbol{\phi}^+$ of simultaneously improved angles is sought for in each iteration step, the problem stays multidimensional. Similar to the Jacobian method the gradient vector $\nabla_{\boldsymbol{\phi}}\, \ell(\boldsymbol{C}(\boldsymbol{\phi}))$ for this multivariate function is given by

$$\left(\frac{\partial \ell(\boldsymbol{C}(\boldsymbol{\phi}))}{\partial \phi_r}\right)_{1\le r\le R} = \left(\sum_{i=1}^{D}\sum_{j=1}^{D} \frac{\partial \ell(\boldsymbol{C}(\boldsymbol{\phi}))}{\partial C_{ij}} \cdot \frac{\partial C_{ij}}{\partial \phi_r}\right)_{1\le r\le R}.$$

The left hand side derivatives inside the summation can be drawn from the matrices $\boldsymbol{H}'$, $\boldsymbol{H}''$ as defined above. The rightmost expression involves the calculation of $\partial \boldsymbol{C}/\partial \boldsymbol{\phi}_r$ for each r. It is easily shown that $\boldsymbol{C}_r(\boldsymbol{\phi})$ shares the product form with the original matrix $\boldsymbol{C}$ except for the rth factor which amounts to $\partial \boldsymbol{U}_{p_r q_r}(\phi_r)/\partial \phi_r$ and is available through elementary calculus.

5. FIRST EXPERIMENTAL RESULTS

The MLR approach was evaluated running a simple frame-by-frame phone recognition task. For that purpose, a multivariate Gaussian classifier was designed using ten hours of speech for training and one hour for testing. The recognition rates achieved when 44 different phonetic labels had to be distinguished are shown in Table 1. Three different orthonormal mappings were employed in order to reduce the 36-dimensional space (12 cepstral coefficients taken from 3 neighboring speech frames) to the $\mathbb{R}^{12}$, decreasing the number of distribution parameters as well as computational complexity by roughly one order of magnitude.

rotation matrix	**KL**	**LD**	**MLR**
recognition rate	52.7%	53.2%	54.1%

Table 1: Labeling accuracy and feature rotations

As indicated in Table 1, the MLR mapping allows a slight improvement of recognition accuracy over the Karhunen-Loève (KL) or linear discriminant (LD) transforms. Whilst the MLR approach certainly presents a new target function which is superior to the variance criteria of KL and LD, the ML rotation matrix is undoubtedly much harder to optimize. The gradient descent methods tested so far proved quite impractical due to the enormous expense in computing the partial derivatives. Moreover, local optimization is a questionable strategy when confronted with an objective function as rugged as ℓ. Thus we moved to global minimizers without reference to derivatives such as the simplex algorithm [9] or combinatorial optimization procedures; the above result relates to the great deluge algorithm [10].

6. CONCLUSION

The FTHMM outlined above is an extension of the SCHMM formalism which incorporates a feature rotation matrix $\boldsymbol{C}$ as a Maximum-Likelihood (ML) trainable component of the probabilistic model. The ML parameter estimation can be done straightforwardly by applying the EM algorithm for the computation of the free transform parameters, too. The MLR generalizes work on linear feature selection, for instance the KL and LD transformations.

Nevertheless, in contrast to classical linear transformation methods the resulting parameter estimation problem forces the use of numerical or combinatorial procedures for global optimization. Thus, the actual slight increase of recognition accuracy correlates with the solution of a more complex optimization within the off-line training stage.

7. REFERENCES

[1] S.E. Levinson, L.R. Rabiner, and M.M. Sondhi. An Introduction to the Application of the Theory of Probabilistic Functions of a Markov Process to Automatic Speech Recognition. *Bell Systems Technical Journal*, 62(4):1035–1074, 1983.

[2] X.D. Huang, Y. Ariki, and M.A. Jack. *Hidden Markov Models for Speech Recognition*. Number 7 in Information Technology Series. Edinburgh University Press, Edinburgh, 1990.

[3] S. Furui. Speaker-Independent Isolated Word Recognition Using Dynamic Features of Speech Spectrum. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 34(1):52–59, 1986.

[4] S.B. Davis and P. Mermelstein. Comparison of Parametric Representation for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 28(4):357–366, 1980.

[5] J.T. Tou and R.P. Heydorn. Some Approaches to Optimum Feature Extraction. In J.T. Tou, editor, *Computer and Information Sciences*, volume 2, pages 57–89. Academic Press, New York, 1967.

[6] H.P. Friedman and J. Rudin. On Some Invariant Criteria for Grouping Data. *American Statistical Association Journal*, 12:1159–1178, 1967.

[7] A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. Royal Statist. Soc. Ser. B*, 39(1):1–22, 1977.

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

[9] H. A. Eiselt, G. Pederzoli, and C.-L. Sandblom. *Continuous Optimization Models*. Walter de Gruyter, Berlin, 1987.

[10] G. Dueck, T. Scheuer, and H. Wallmeier. Tolerance Threshold and Great Deluge: New Ideas for Optimization (in German). *Spektrum der Wissenschaft*, pages 42–51, März 1993.