RC24458 (W0712-081) October 1, 2007
Computer Science

IBM Research Report

Adapted Extended Baum-Welch Transformations

Dimitri Kanevsky, Daniel Povey, Bhuvana Ramabhadran
IBM Research Division
Thomas J. Watson Research Center
P.O. Box 218
Yorktown Heights, NY 10598

Tara N. Sainath
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street
Cambridge, MA 02139

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

ADAPTED EXTENDED BAUM-WELCH TRANSFORMATIONS

Dimitri Kanevsky[*]*, Daniel Povey, Bhuvana Ramabhadran*

IBM T. J. Watson Research Center
Yorktown Heights
NY 10598
{kanevsky, dpovey, bhuvana}@us.ibm.com

Tara N. Sainath

MIT Computer Science and
Artificial Intelligence Laboratory
32 Vassar St. Cambridge
MA 02139
tsainath@mit.edu

ABSTRACT

The discrimination technique for estimating parameters of Gaussian mixtures that is based on the Extended Baum-Welch transformations (EBW) has had significant impact on the speech recognition community. In this paper we introduce a general definition of a family of EBW transformations that can be associated with a weighted sum of updated and initial models. We compute a gradient steepness measurement for a family of EBW transformations that are applied to functions of Gaussian mixtures and demonstrate the growth property of these transformations. We consider EBW transformations of discriminative functions in which EBW controlled parameters are adapted to a gradient steepness measurement or to the likelihood of the data given the model. We present experimental results that show that adapted EBW transformations can significantly speed up estimating parameters of Gaussian mixtures and give better decoding results.

Index Terms— MMIE training, EBW transformations, gradient steepness

1. INTRODUCTION

The EBW transformations [4] are one of variety of discriminative training techniques ([11], [12], [4], [20]) to estimate model parameters of Gaussian mixtures. These transformations involve special EBW parameters that control the growth of an objective function that is being used to estimate model parameters of Gaussian mixtures (e.g. the Maximum Mutual Information (MMI)). Changing EBW control parameters in one direction (in which training models changed less), guarantees that EBW transformations increase the value of this objective function ([6], [7], [1]), but the training process in this case requires many iterations and becomes very slow. Otherwise, if these control parameters are changed in another direction, then the training becomes unstable. Previous efforts to optimize control parameters to facilitate fast growth were successful on relatively small speech tasks [19]. But in general, for large vocabulary speech recognition tasks, unconstrained optimization of control parameters in the objective function leads to overtraining of the estimated model. One of the ways to prevent overtraining is to introduce constrained manifolds to which these control parameters belong, and to optimize these parameters subject to these constraints ([9], [10]). In this paper we suggest a method to adapt EBW control parameters during discrimination training by evaluating how well an initial model fits input data. If the initial model fits the input data better (e.g. the likelihood of the data given the initial model is high) then we impose fewer changes on the initial model via EBW transformations. This is achieved by multiplying control parameters in some representation of the EBW transformations by the likelihood (as in [16]). We demonstrate improved perfomance of the adapted EBW re-estimation over standard MMI training. We also considered the EBW gradient steepness measurements introduced in ([14], [15]) that are required to estimate the new model via the EBW transformations. The gradient steepness is flatter if the data fits the initial model better when EBW transformations are applied. Therefore one can adapt EBW transformations by multiplying EBW control parameters (in some system coordinate representation) by a factor that, for example, is inversely proportional to the square root of a gradient steepness measurement.

In the next section we define EBW transformations. In section 3 we give a novel definition of a family of EBW transformations that are convenient for introducing adapted EBW techniques. In section 4, using ([6], [7], [8]) we reproduce explicit formulas to measure the gradient steepness. Section 5 suggests adapted EBW approaches; the experiments performed are presented in section 6. Section 7 concludes the paper and discusses future work.

2. EBW TRANSFORMATIONS

Let $F(z) = F(z_{ij})$ be some function in variables $z = (z_{ij})$ and $c_{ij} = z_{ij}\frac{\delta}{\delta z_{ij}}F(z)$.

2.1. Diagonal Gaussian mixture densities:

Let

$$z_{ij} = z_i(\mu_j, \sigma_j) = \frac{1}{(2\pi)^{1/2}\sigma_j} e^{-(y_i-\mu_j)^2/2\sigma_j^2} \tag{1}$$

and y_i is a sample of training data. EBW transformations for diagonal mixture densities are defined as the following.

$$\hat{\mu}_j = \mu_j(C) = \frac{\sum_{i\in I} c_{ij} y_i + C\mu_j}{\sum_{i\in I} c_{ij} + C} \tag{2}$$

$$\hat{\sigma}_j^2 = \sigma_j(C)^2 = \frac{\sum_{i\in I} c_{ij} y_i^2 + C(\mu_j^2 + \sigma_j{}^2)}{\sum_{i\in I} c_{ij} + C} - \mu_j(C)^2 \tag{3}$$

[*]This work was partially supported by the Defense Advanced Research Projects Agency under contract No. HR0011-06-2-0001

2.2. Multidimensional multivariate Gaussian mixture densities:

Let

$$z_{ij} = \frac{|\Sigma_j|^{-1/2}}{(2\pi)^{n/2}} e^{-1/2(y_i-\mu_j)^T \Sigma_j^{-1}(y_i-\mu_j)} \quad (4)$$

and $y_i^T = (y_{i1}, ...y_{im})$ is a sample of training data. EBW transformations for multivariate Gaussian mixture densities are defined as the following.

$$\hat{\mu}_j = \mu_j(C) = \frac{\sum_{i\in I} c_{ij} y_i + C\mu_j}{\sum_{i\in I} c_{ij} + C} \quad (5)$$

$$\hat{\Sigma}_j = \Sigma_j(C) = \frac{\sum_{i\in I} c_{ij} y_i y_i^T + C(\mu_j \mu_j^T + \Sigma_j)}{\sum_{i\in I} c_{ij} + C} - \mu_j(C)\mu_j(C)^T \quad (6)$$

It was shown in ([6], [7]) that (2, 3, 5, 6) are growth transformations for sufficiently large C and a formula for the asymptotic growth is derived. A slight generalization of these statements is given in Section 4 below.

3. ASYMPTOTIC FAMILY OF EBW TRANSFORMATIONS

In this section we consider a different way to introduce transformations of Gaussian mixtures and show that they are in fact asymptotically equivalent to EBW transformations.

3.1. Diagonal Gaussian mixture densities:

Let $\{\tilde{\mu}_j\}$ and $\{\tilde{\sigma}_j\}$ be solutions of the following equation:

$$\sum_i z_{ij} \frac{\delta F(z)}{\delta z_{ij}} \frac{\delta \log z_i(\tilde{\mu}_j, \tilde{\sigma}_j)}{\delta \tilde{\mu}_j} = 0 \quad (7)$$

$$\sum_i z_{ij} \frac{\delta F(z)}{\delta z_{ij}} \frac{\delta \log z_i(\tilde{\mu}_j, \tilde{\sigma}_j)}{\delta \tilde{\sigma}_j} = 0 \quad (8)$$

(where z is a function of $\{\mu_j\}, \{\sigma_j\}$ as defined in 1). Let us define the following transformations for means and variances:

$$\bar{\mu}_j(\alpha_j) = \alpha_j \tilde{\mu}_j + (1-\alpha_j)\mu_j \quad (9)$$

$$\bar{\sigma}_j^2(\alpha_j) = \alpha_j \tilde{\sigma}_j^2 + (1-\alpha_j)\sigma_j^2 \quad (10)$$

where

$$\alpha_j = \frac{\sum_i c_{ij}}{C}$$

For these transformations the following statement holds.

Theorem 1 *Let $\sum_i c_{ij} \neq 0$. If $C \to \infty$ (or equivalently, $\alpha_j \to 0$) then transformations (2, 3) are asymptotically equivalent to (9, 10) in the following sense. There exist a constant d such that for all sufficiently large C the following inequalities hold.*

$$|\hat{\mu}_j(C) - \bar{\mu}_j(\alpha_j)| < d/C^2 \quad (11)$$

$$|\hat{\sigma}_j(C) - \bar{\sigma}_j(\alpha_j)| < d/C^2 \quad (12)$$

Proof First, one can easily show that

$$\tilde{\mu}_j = \frac{\sum_i c_{ij} y_i}{\sum_i c_{ij}} \quad (13)$$

and

$$\tilde{\sigma}_j^2 = \frac{\sum_i c_{ij}(y_i - \mu_j)^2}{\sum_i c_{ij}} \quad (14)$$

Next,

$$\hat{\mu}_j = \mu_j(0)\tilde{\alpha}_j + \mu_j(1-\tilde{\alpha}_j) \quad (15)$$

where

$$\tilde{\alpha}_j = \tilde{\alpha}_j(C) = \frac{\sum_i c_{ij}}{\sum_i c_{ij} + C}$$

and $\mu_j(0)$ is defined as in (2) for $C = 0$. Statement (11) now follows from the fact that $\lim_{C\to\infty} C * \tilde{\alpha}_j(C) = 1$. Statement (12) follows from (18) in [6].

3.2. Remark

Transformations

$$\bar{\mu}_j(\alpha_j) + f(\alpha_j) \quad (16)$$

$$\bar{\sigma}_j^2(\alpha_j) + g(\alpha_j) \quad (17)$$

can be considered "asymptotically equivalent" to EBW transformations for any functions f, g such that $lim_{\alpha_j \to 0} f(\alpha_j)/\alpha_j = 0$, $lim_{\alpha_j \to 0} g(\alpha_j)/\alpha_j = 0$. The concept of asymptotically equivalent EBW transformations could also be represented in the context of weak-sense auxiliary functions [12].

4. GRADIENT STEEPNESS THEOREM FOR EBW TRANSFORMATIONS

It was shown in [4] that EBW transformations are growth transformations for sufficiently large C when F is a rational function of discrete parameters. Updated formulae (2, 3, 5, 6) for rational functions of continuous parameters were obtained through discrete probability approximation of Gaussian densities [11] and have been widely used as an alternative to direct gradient-based optimization approaches ([20], [18]). Using the linearization technique that was originally presented in our IBM Research Reports [5] and in [6] for diagonal Gaussian mixtures and in [7] for multidimensional multivariate Gaussian mixtures, we proved in [8] that transformations (2, 3, 5, 6) are growth transformations for large C if functions F obey certain smoothness constraints. In what follows, we give a slightly more general formulation of the growth transformation theorem. Let $\hat{z}_{ij} = \frac{1}{(2\pi)^{1/2}\sigma_j(D_j)} e^{-(y_i-\mu_j(C_j))^2/2\sigma_j(D_j)^2}$

Theorem 2 *Let $F(\{z_{ij}\})$, $i = 1...m$, be differentiable at μ_j, σ_j and $\frac{\delta F(\{z_{ij}\})}{\delta z_{ij}}$ exist at z_{ij}. Let either $\hat{\mu}_j \neq \mu_j$ or $\hat{\sigma}_j \neq \sigma_j$. Then for sufficiently large C_j, D_j*

$$F(\{\hat{z}_{ij}\}) - F(\{z_{ij}\}) = \sum_j (\alpha_j T_\mu^j + \beta_j T_\sigma^j) + \sum_j (o(\alpha_j) + o(\beta_j)) \quad (18)$$

Where $\alpha_j = 1/C_j, \beta_j = 1/D_j$ and

$$T_\mu^j = \frac{[\sum_i c_{ij}(y_i - \mu_j)]^2}{\sigma_j^2} > 0 \quad (19)$$

$$T_\sigma^j = \frac{\{\sum_i c_{ij}[(y_i - \mu_j)^2 - \sigma_j^2]\}^2}{2\sigma_j^4} > 0 \quad (20)$$

(Here $o(\epsilon)$ means that $o(\epsilon)/\epsilon \to 0$ if $\epsilon \to 0$). In other words, $F(\{\hat{z}_{ij}\})$ grows proportionally to $\sum_j \alpha_j T_\mu^j + \sum_j \beta_j T_\sigma^j$ for sufficiently small $\alpha_j, \beta_j > 0$.

Proof The proof is similar to one that is given in [6] for Theorem 1 (in which it was assumed that $C = D$). For details see [8].

4.1. Remark

A similar gradient steepness result, as seen in this theorem, can be formulated for multidimensional multivariate Gaussian mixtures (see [6], [8]).

4.2. Remark

If $\sum_j c_{ij} \neq 0$ then one can represent (19, 20) as

$$T'^{j}_{\mu} = \frac{(\bar{\mu}_j - \mu_j)^2}{\sigma_j^2} > 0 \quad (21)$$

$$T'^{j}_{\sigma} = \frac{(\bar{\sigma}_j^2 - \sigma_j^2)^2}{2\sigma_j^4} > 0 \quad (22)$$

and

$$F(\{\hat{z}_{ij}\}) - F(\{z_{ij}\}) = \sum_j \alpha'_j T'^{j}_{\mu} + \sum_j \beta'_j T'^{j}_{\sigma} + \sum_j (o(\alpha'_j) + o(\beta'_j)) \quad (23)$$

where $\alpha'_j = (\sum_i c_{ij})^2 / C_j, \beta'_j = (\sum_i c_{ij})^2 / D_j$. One can show that (21 and 22) approximate mean and variance "components" of the KL-divergence between two Gaussians (for updated and initial models). Therefore the gradient steepness measure can be used to evaluate the closeness of updated models to initial models. These metrics were used in ([14], [15], [16], [17]) for various speech tasks. They also can be used to avoid overtraining in adapted EBW training as discussed in the next section.

4.3. Remark

In section 3.1 let us connect models $\{\mu_j, \sigma_j^2\}$ and $\{\tilde{\mu}_j, \tilde{\sigma}_j^2\}$ with a line. Then the following cases for location of an updated model (at which F increases its value) can be considered: If $\sum_i c_{ij} > 0$ then a step $\alpha_j > 0$ and $\{\bar{\mu}_j(\alpha_j), \bar{\sigma}_j^2(\alpha_j)\}$ lies on a segment that connects $\{\mu_j, \sigma_j^2\}$ and $\{\tilde{\mu}_j, \tilde{\sigma}_j^2\}$. If $\sum_i c_{ij} < 0$ then a step $\alpha_j < 0$ and $\{\bar{\mu}_j(\alpha_j), \bar{\sigma}_j^2(\alpha_j)\}$ lies outside of the segment that connects $\{\mu_j, \sigma_j^2\}$ and $\{\tilde{\mu}_j, \tilde{\sigma}_j^2\}$. These cases correspond to cases in [10] where a sign of a step along a gradient was chosen depending on whether F has the minimum or the maximum at $\{\tilde{\mu}_j, \tilde{\sigma}_j^2\}$.

5. ADAPTED EBW TRAINING

EBW control parameters in weighted sums of updated and initial models could be required to belong to some constraint manifold (for example, one can consider constraints that guarantee the following: if speech segments were decoded correctly in one training iteration, then they are decoded correctly in next training iterations (see for details [9]). In adapted EBW training one can correlate α and β in (18) to T_μ and T_σ. The idea is that when a gradient is less steep, the initial models are closer to optimal ones and therefore higher weights should be associated with these models. For example, one can set $\alpha \sim \gamma_1 / {T_\mu}^{r_1}$ and $\beta \sim \gamma_2 / {T_\sigma}^{r_2}$. The case $r_1 = r_2 = 1/2$ was intensively studied in ([10]) where it was represented as constrained optimization. One can also consider associated with T_μ and T_σ various gradient steepness metrics from ([16], [17]). Another way to adapt EBW control parameters is to choose for each Gaussian a value C that is proportional to the likelihood of this Gaussian over training data. If this likelihood is low, then the initial model is more changed when it is exposed to data during training. We will describe experimental results using C proportional to likelihood for each Gaussian in the following section.

Iteration	Test set / Training method					
	rt03 baseline	rt03 mixed	dev04f baseline	dev04f mixed	rt04 baseline	rt04 mixed
0	13.0%		23.2%		20.5%	
1	12.6%	**12.3%**	22.4%	**21.5%**	19.9%	**18.9%**
2	**12.3%**		21.8%		19.5%	
3	12.3%		21.4%		19.1%	
4	12.2%		21.3%		18.8%	
5	12.3%	**12.0%**	21.1%	**21.1%**	18.7%	**18.3%**
7	12.1%		21.1%		18.4%	
8	**12.0%**		**21.0%**		**18.5%**	

Table 1. English WER on test sets rt03 (2:15 hours), dev04f (2:00 hours), rt04 (4:00 hours)

6. EXPERIMENTAL RESULTS

In this section we report results on a speaker independent English broadcast news system. Discriminative baseline for training is done with an acoustic weight of 0.1 and language model weight of 1.0, and $E = 2.0$ for the baseline MMI (for backoff). The acoustic model for the English system is trained on 450 hours of speech comprising the 1996 and 1997 English Broadcast News Speech collections and the English broadcast audio from TDT-4. Lightly-supervised training was performed on the TDT-4 audio because only closed captions were available. The recognition features are 40-d vectors computed via an LDA+MLLT projection of 9 spliced frames of 19-d PLP features. Utterance-based cepstral mean subtraction is used, but no speaker adaptation. The model has 6000 quinphone context dependent states and 250K Gaussians. The language model used to build the decoding graph is trained on a 192M word corpus comprising the 1996 and 1997 English Broadcast News Transcripts, the 1996 CSR Hub4 Language Model data, the EARS BN-03 English closed captions, the English portion of TDT-4, and the GALE Y1Q1 and Y1Q2 English closed captions. The final language model is 4-gram, Kneser-Ney smoothed and has 3.2M n-grams. The vocabulary has 77K words with 1.08 variants per word. We use the test sets rt03, dev04f and rt04 as defined for the English portion of the EARS program, which after silence removal have lengths of 2:15, 2:00 and 4:00 hours respectively. Table 1 describes experiments on test sets in which we iterated baseline training with adapted EBW training for 1st and 5th iterations. Columns that are labeled as *test name/baseline* (for example, *rt04/baseline*) contain results for baseline MMI training for 8 iterations (starting from a ML baseline). Each column labeled as *test name/mixed* represents two results. In a line *iteration 1* result of the application of an adapted EBW method to the ML baseline is presented. These results show, for example, a single iteration in *rt04 mixed* achieved a WER of 18.9%, which is similar to the WER of *rt04 baseline* that was achieved at the 4th iteration. In other words, the first iteration of the adapted EBW training allows to achieve decoding results that require 2-4 iterations of the baseline traning. Next, in the line *iteration 5* application of the adapted EBW training to an output of a 4th baseline training are presented. These results are slightly better or the same that were obtained with 8 iterations of the baseline training. We also tested what improvement this adapted EBW method gives if it is applied consequently in two iterations. The results are shown in table 2. As expected, a second adapted EBW iteration on top of a first adapted EBW iteration produces less improvement in decoding accuracy. This can be explained as follows: an adapted EBW transformation involves a Gaussian specific control parameter $C \sim const + loglikelihood$

Iteration	Test set Training method		
	rt03 adapted EBW	dev04f adapted EBW	rt04 adapted EBW
0	13.0%	23.2%	20.5%
1	12.3%	21.5%	18.9%
2	12.1%	21.3%	18.7%

Table 2. English WER on test sets rt03 (2:15 hours), dev04f (2:00 hours), rt04 (4:00 hours); adapted EWB training

where $const = max|loglikelihood| - \epsilon$. Here maximum is computed for all Gaussians and all training data and ϵ is some small integer number. Therefore C is relatively small. This causes overtraining if this small C is applied in consequent iterations. In order to avoid overtraining in a consequent interation, one needs to mutilpy C by some large constant. We run some preliminary experiments for consequent iterations multiplying $C \sim const + loglikelihood$ by a factor that increases with each iteration and this helped to boost the overall perfomance. Recently, a novel modified form of the MMI objective function which gives improved results for discriminative training was suggested in [13]. The modification includes (among other things) implementation of the Extended Baum-Welch update equations for MMI that cancels any shared part of the numerator and denominator statistics on each frame. We have begun testing our adapted EWB training with this modifed MMI training. Initial experimental results confirmed that the adapted EBW training can help speed up MMI training and improve the overall performance. For example, the modified MMI training in [13] achieves 17.3% WER on the rt04 test set after the 4th iteration . But if we run only 3 iterations of the modified MMI training and then apply the adapted EBW transformations, we get WER 16.9%. Similarly, if we run the adapted EWB in the first iteration in which it is applied to denominator statistics that are generated by the modified MMI then we get WER 17.8% . This result is by 0.6% better absolutely than WER that is obtained by using only the modified MMI training in the first iteration.

7. CONCLUSION AND FUTURE WORK

In the paper we considered a family of transformations that can be associated with weighted sums of updated and initial models. We showed that this family of transformations is asymptotically equivalent to EWB transformations. This observation allowed us to compute gradient steepness for this family of transformations. We introduced discriminative training in which EBW controlled parameters are adapted to gradient steepness or to the likelihood of the data given the initial model. We presented experimental results that show that adapted EBW transformations can significantly speed up training and improve the overall performance across diferent speech tasks. We plan to continue to study EBW based training in which EBW control parameters are correlated to gradient steepness along "mean and variance directions."

8. REFERENCES

[1] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative Training of Subspace Constrained GMMs for Speech Recognition," to be submitted to IEEE Transactions on Speech and Audio Processing.

[2] L.E.Baum and J.A. Eagon, "An inequality with applications to statistical prediction for functions of Markov processes and to a model of ecology," *Bull. Amer. Math. Soc.*, vol. 73, pp.360-363, 1967.

[3] A. Gunawardana and W. Byrne, "Discriminative Speaker Adaptation with Conditional Maximum Likelihood Linear Regression," ICASSP, 2002.

[4] P.S. Gopalakrishnan, D. Kanevsky, D. Nahamoo and A. Nadas, "An inequality for rational functions with applications to some statistical estimation problems", IEEE Trans. Information Theory, Vol. 37, No.1 January 1991

[5] D. Kanevsky, "Growth Transformations for General Functions", RC22919 (W0309-163), September 25, 2003.

[6] D. Kanevsky, "Extended Baum transformations for general functions", in Proc. ICASSP, 2004.

[7] D. Kanevsky, "Extended Baum Transformations for General Functions, II", tech. Rep. RC23645(W0506-120), Human Language technologies, IBM , 2005, http://cogprints.org/5058/01/rc23645.pdf

[8] D. Kanevsky, "Extended Baum-Welch Transformations for General Functions, III", ech. Rep. RC, Human Language technologies, IBM , 2007

[9] D. Kanevsky, "Constrained corrective training for continuous parameter system", US patent 6,044,344, March 28, 2000.

[10] Cong Liu Peng Liu Hui Jiang Soong, F. Ren-Hua Wang, "A Constrained Line Search Optimization for Discriminative Training in Speech Recognition", in Proc. ICASSP, 2007.

[11] Y. Normandin, "An improved MMIE Training Algorithm for Speaker Independent, Small Vocabulary, Continuous Speech Recognition", Proc. ICASSP'91, pp. 537-540, 1991.

[12] Daniel Povey, "Discriminative Training for Large Vocabulary Speech Recognition", PhD Thesis March 1, 2003.

[13] Daniel Povey, Dimitri Kanevsky, Brian Kingsbury, Bhuvana Ramabhadran, George Saon and Karthik Visweswariah "Boosted MMI for model and feature-space discriminative training", submitted for ICASSP'08.

[14] Tara N. Sainath, Dimitri Kanevsky, Giridharan Iyengar,"Unsupervised Audio Segmentation Using Extended Baum-Welch Transformations, in Proc. ICASSP, 2007.

[15] Tara N. Sainath, Victor Zue, Dimitri Kanevsky, "Audio-Classification using Extended Baum-Welch Transformations" , in Proc. Interspeech 2007.

[16] Tara N. Sainath, Dimitri Kanevsky, Bhuvana Ramabhadran "Broad Phoentic Recognition in a Hidden Markov Model Framework Using Extended Baum-Welch Transformations " , to appear in Proc. ASRU 2007.

[17] Tara N. Sainath, Dimitri Kanevsky, Bhuvana Ramabhadran "Gradient Steepness Metrics Using Extended Baum-Welch Transformations for Universal Pattern Recognition Tasks" , submitted for ICASSP 2008.

[18] R. Schluter, W. Macherey, B. Muler and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition", Speech Communication, Vol. 34, pp.287-310, 2001.

[19] V. Goel, Personal communication, 2000.

[20] V. Valtchev , P.C. Woodland and S. J. Young, "Lattice-based Discriminative Training for Large Vocabulary Speech Recognition Systems", Speech Communication, Vol. 22, pp. 303-314, 1996.