ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Optimization-Based Control for the Extended Baum-Welch Algorithm

Janne Pylkkönen and Mikko Kurimo

Department of Information and Computer Science, Aalto University, Finland
janne.pylkkonen@aalto.fi, mikko.kurimo@aalto.fi

Abstract

The extended Baum-Welch (EBW) is the most popular algorithm for discriminative training of speech recognition acoustic models. The EBW algorithm is usually controlled with heuristic rules, which are used to determine the smoothing parameters of the algorithm. In this paper we propose a control method for EBW which is based on the optimization of an error measure over a small control set. The large vocabulary speech recognition experiments show this to have clear benefits over the heuristic control.

Index Terms: speech recognition, discriminative training, extended Baum-Welch, optimization

10.21437/Interspeech.2012-492

1. Introduction

The most common algorithm for parameter estimation when performing discriminative training of speech recognition acoustic models is the extended Baum-Welch (EBW) algorithm [1]. It can be used with a wide variety of discriminative criteria. With proper heuristics, it is applicable even with the most complex acoustic models used for large vocabulary continuous speech recognition (LVCSR).

The use of EBW algorithm requires setting its smoothing or control parameters, so called D constants. The controlling of EBW has been studied rather extensively [2, 3, 4], but the theoretical considerations have been unable to give effective rules for setting the EBW constants. Instead, heuristic rules are used. The first methods used the validity of Gaussian covariances as the criterion to set either one global constant for all the Gaussians in the acoustic model, or phone-specific constants [5] to improve the convergence rate of the algorithm. Later Woodland and Povey [1] revised the heuristics to apply Gaussian-specific EBW constants. Because of good convergence properties, it has remained as the prevailing method for controlling EBW.

This paper presents a novel method for EBW control, based on the optimization of the D constants using a held-out control set. The control method is presented and experimented in two scenarios, one with a matched condition training, and another where a small amount of mismatched target data is available for training. The proposed method is able to provide performance improvements in both cases.

2. Extended Baum-Welch algorithm

The theoretical foundation of the EBW algorithm is in the growth transformations [4]. With properly set D constants it can guarantee that the discriminative criterion under optimization increases or remains the same. For the mean ($\boldsymbol{\mu}_{ik}$) and covariance ($\boldsymbol{\Sigma}_{ik}$) of Gaussian k in mixture i, the EBW re-estimation formulae [1] are

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\mathcal{O}_{ik}^{num}(\boldsymbol{x}) - \mathcal{O}_{ik}^{den}(\boldsymbol{x}) + D\boldsymbol{\mu}_{ik}}{\mathcal{O}_{ik}^{num}(1) - \mathcal{O}_{ik}^{den}(1) + D} \quad (1)$$

$$\hat{\boldsymbol{\Sigma}}_{ik} = \frac{\mathcal{O}_{ik}^{num}(\boldsymbol{x}\boldsymbol{x}^T) - \mathcal{O}_{ik}^{den}(\boldsymbol{x}\boldsymbol{x}^T) + D[\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^T + \boldsymbol{\Sigma}_{ik}]}{\mathcal{O}_{ik}^{num}(1) - \mathcal{O}_{ik}^{den}(1) + D} - \hat{\boldsymbol{\mu}}_{ik}\hat{\boldsymbol{\mu}}_{ik}^T \quad (2)$$

where $\mathcal{O}_{ik}(1)$ represents the occupancy statistics and $\mathcal{O}_{ik}(\boldsymbol{x})$ and $\mathcal{O}_{ik}(\boldsymbol{x}\boldsymbol{x}^T)$ are the weighted sums of the observed data and squared data, respectively. Superscripts num and den are abbreviations for numerator and denominator, referring to statistics collected from the correct transcription and the alternative hypotheses, respectively. The EBW formulae can be simplified by defining $\mathcal{O}_{ik}(1) = \mathcal{O}_{ik}^{num}(1) - \mathcal{O}_{ik}^{den}(1)$ and similarly for $\mathcal{O}_{ik}(\boldsymbol{x})$ and $\mathcal{O}_{ik}(\boldsymbol{x}\boldsymbol{x}^T)$. This gives:

$$\hat{\boldsymbol{\mu}}_{ik} = \frac{\mathcal{O}_{ik}(\boldsymbol{x}) + D\boldsymbol{\mu}_{ik}}{\mathcal{O}_{ik}(1) + D} \quad (3)$$

$$\hat{\boldsymbol{\Sigma}}_{ik} = \frac{\mathcal{O}_{ik}(\boldsymbol{x}\boldsymbol{x}^T) + D[\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^T + \boldsymbol{\Sigma}_{ik}]}{\mathcal{O}_{ik}(1) + D} - \hat{\boldsymbol{\mu}}_{ik}\hat{\boldsymbol{\mu}}_{ik}^T. \quad (4)$$

In the commonly used heuristic control strategy [1], the Gaussian specific constant D_{ik} is set to be the maximum of 1. denominator occupancy multiplied by two or 2. double the value necessary for the Gaussian covariance to be positive definite ($2D_{ik}^{\min}$). The first condition is needed to achieve proper convergence, whereas the second condition defines a minimum constant to guarantee feasible Gaussian covariance.

Discriminative training requires other heuristics as well, such as the use of acoustic scaling parameter [1], applying of weakened language model [2, 1] and the use of lattices [1, 6]. For some discriminative criteria such as MPE [7], the above heuristics for setting the EBW constants is not enough, and additional measures such as I-smoothing are needed in order to get well-performing models.

3. Optimizing EBW constants

This section presents a novel method for controlling EBW, where the EBW constants are numerically optimized by using a held-out set of utterances during the discriminative training. Before each parameter estimation step using the EBW re-estimation formulae, new constants[1] are obtained by a control procedure. The outline of this procedure is:

1. Collect discriminative statistics over the training data
2. Based on the current set of EBW constants and the training set statistics, estimate a new model

[1]To retain the common terminology, we call the EBW D parameters constants, even though they are optimized for each update step.

3. Compute discriminative statistics over the held-out control set, using the new model
4. Update the EBW constants to optimize the objective function (discriminative criterion over the control set)
5. Iterate steps 2.–4. until convergence or a maximum number of iterations has been reached

The performance of the models over the held-out control set could be measured with any discriminative criterion. We chose the symmetrically normalized frame error (SNFE) [8], which shows good correlation with the edit distance. Similar to MPE, this criterion gives an approximation to the phoneme error over the hypotheses encoded in the lattices.

3.1. Computing the gradients

The performance over the control set is optimized with respect to the EBW constants using derivatives of the discriminative criterion. Details about the derivatives of common discriminative criteria have been presented by Liu *et al.* [9]. The formulae used in this study, which assume diagonal covariance matrices, are:

$$\nabla\mathcal{F}(\hat{\boldsymbol{\mu}}_{ik}) = -\kappa\hat{\boldsymbol{\Sigma}}_{ik}^{-1}\left(\mathcal{O}_{ik}^{ctrl}(\boldsymbol{x}) - \mathcal{O}_{ik}^{ctrl}(1)\hat{\boldsymbol{\mu}}_{ik}\right) \tag{5}$$

$$\nabla\mathcal{F}(\hat{\boldsymbol{\Sigma}}_{ik}) = -\frac{\kappa}{2}\,\mathrm{diag}\left(\hat{\boldsymbol{\Sigma}}_{ik}^{-2}\left[\mathcal{O}_{ik}^{ctrl}(\boldsymbol{x}\boldsymbol{x}^T) - 2\mathcal{O}_{ik}^{ctrl}(\boldsymbol{x})\hat{\boldsymbol{\mu}}_{ik}^T + \mathcal{O}_{ik}^{ctrl}(1)(\hat{\boldsymbol{\mu}}_{ik}\hat{\boldsymbol{\mu}}_{ik}^T - \hat{\boldsymbol{\Sigma}}_{ik})\right]\right) \tag{6}$$

where κ is the acoustic scale (inverse of the language model scaling) and diag extracts the diagonal of the matrix. To separate the discriminative statistics computed over the training and control sets, we use superscripts $train$ and $ctrl$.

The rate of change of the model parameters when D_{ik} changes can be obtained by taking a derivative of (3) and (4) with respect to D_{ik}:

$$\frac{d\hat{\boldsymbol{\mu}}_{ik}}{dD_{ik}} = \frac{\mathcal{O}_{ik}^{train}(1)\boldsymbol{\mu}_{ik} - \mathcal{O}_{ik}^{train}(\boldsymbol{x})}{\left(\mathcal{O}_{ik}^{train}(1) + D_{ik}\right)^2} \tag{7}$$

$$\frac{d\hat{\boldsymbol{\Sigma}}_{ik}}{dD_{ik}} = \frac{\mathcal{O}_{ik}^{train}(1)\left(\boldsymbol{\mu}_{ik}\boldsymbol{\mu}_{ik}^T + \boldsymbol{\Sigma}_{ik}\right) - \mathcal{O}_{ik}^{train}(\boldsymbol{x}\boldsymbol{x}^T)}{\left(\mathcal{O}_{ik}^{train}(1) + D_{ik}\right)^2} - 2\hat{\boldsymbol{\mu}}_{ik}\left(\frac{d\hat{\boldsymbol{\mu}}_{ik}}{dD_{ik}}\right)^T. \tag{8}$$

Applying the chain rule we can now combine (5)–(8) to obtain the derivative of the discriminative criterion with respect to D_{ik}:

$$\frac{d\mathcal{F}}{dD_{ik}} = \nabla\mathcal{F}(\hat{\boldsymbol{\mu}}_{ik})\cdot\frac{d\hat{\boldsymbol{\mu}}_{ik}}{dD_{ik}} + \nabla\mathcal{F}(\hat{\boldsymbol{\Sigma}}_{ik})\cdot\mathrm{diag}\left(\frac{d\hat{\boldsymbol{\Sigma}}_{ik}}{dD_{ik}}\right). \tag{9}$$

This allows the optimization of the control set discriminative criterion with respect to D_{ik}, using any general-purpose gradient-based optimization algorithm. We chose to use Quickprop [10] because of the good convergence properties and the ease of implementation.

From (7) and (8) it can be seen that the rate of change of the model parameters decreases as D_{ik} increases. To help the numerical optimization, a logarithmic parameter transformation is therefore applied. Instead of optimizing D_{ik} directly, transformed $\tilde{D}_{ik} = \log D_{ik}$ is used instead. The derivative (9) then needs to be multiplied by

$$\frac{dD_{ik}}{d\tilde{D}_{ik}} = \exp\tilde{D}_{ik} = D_{ik}. \tag{10}$$

The equations above apply to the optimization of one Gaussian specific D_{ik}. To reduce the size of the control set and avoid over-training, it might be beneficial to cluster the EBW constants. Our previous experiments had shown that already a properly set global D achieves good results in discriminative training. In the context of controlling EBW, this corresponds to a global clustering of D constants and optimization of the control set performance over a single parameter. The necessary derivative is then

$$\frac{d\mathcal{F}}{d\tilde{D}} = \sum_{i=1}^{N_m}\sum_{k=1}^{M_i}\frac{d\mathcal{F}}{d\tilde{D}_{ik}} \tag{11}$$

where N_m is the number of mixtures in the model, and M_i is the number of Gaussians in mixture i.

Another obvious clustering level is to let all Gaussians in the same mixture share the same D_i. This leads to an optimization problem with derivatives

$$\frac{d\mathcal{F}}{d\tilde{D}_i} = \sum_{k=1}^{M_i}\frac{d\mathcal{F}}{d\tilde{D}_{ik}}. \tag{12}$$

We will refer to the parameters the control procedure is optimizing as the "control parameters", which can be either clustered or direct EBW constants.

The discriminative statistics over the control set are computed over lattices. Preliminary experiments showed it was beneficial to use a full language model to rescore the control set lattices, unlike the training lattices which use only a weak language model. The segmentation of the control set lattices needs to remain the same over the course of the optimization in order to keep gradient information reliable.

3.2. Limits and priors

The requirement of the covariance matrix to be positive definite poses a minimum limit for the Gaussian specific D_{ik} parameters. This is reflected in the common heuristic method of setting D_{ik}, which states that D_{ik} may not be less than double the Gaussian specific minimum limit $D_{ik}^{\min}$. For control parameters, such hard limits cause discontinuities to the derivatives of the objective function, making the optimization problem more difficult.

To avoid optimization difficulties, the minimum $D_{ik}^{\min}$ values are taken into account by a continuous transform. The Gaussian specific D_{ik} is determined by a function $g_{ik}(D)$ of the control parameter, defined as

$$g_{ik}(D) = \begin{cases} D & \text{if } D \geq 2D_{ik}^{\min} \\ 1.5D_{ik}^{\min} + \alpha\exp\left(\frac{D-2D_{ik}^{\min}}{\alpha}\right) & \text{if } D < 2D_{ik}^{\min} \end{cases} \tag{13}$$

where $\alpha = 0.5D_{ik}^{\min}$. This way the Gaussian specific D_{ik} can reach the usual minimum limit $2D_{ik}^{\min}$, but if the control parameter is below that, the realized D_{ik} becomes restricted, approaching $1.5D_{ik}^{\min}$ when $D \to -\infty$. To account for the transformation in the gradient-based optimization, the derivative of

$g_{ik}(D)$ is needed:

$$g'_{ik}(D) = \begin{cases} 1 & \text{if } D \geq 2D_{ik}^{\min} \\ \exp\left(\frac{D-2D_{ik}^{\min}}{\alpha}\right) & \text{if } D < 2D_{ik}^{\min}. \end{cases} \quad (14)$$

Note that the derivative is continuous at $D = 2D_{ik}^{\min}$.

The control procedure optimizes the EBW constants for one update step at a time. Unfortunately this is not an optimal strategy when considering discriminative training with several EBW iterations. To avoid local minima of model parameters and excessive changes in control parameters, two couteracting steps are taken:

- The control parameters are initialized from the optimization results of the previous EBW step and each parameter is limited to $0.5\mu \leq D \leq 2\mu$, where μ is the mean of the control parameters in the previous step
- A prior is added to favor small control parameter values, corresponding to larger step sizes

The prior is implemented by adding a small positive constant to the derivative $\frac{d\mathcal{F}}{dD_{ik}}$. It corresponds to the addition of

$$r(D_{ik}) = \frac{\log(D_{ik}) - \log(0.5\mu)}{\log(2.0\mu) - \log(0.5\mu)}\beta \quad (15)$$

to the objective function, where μ is the mean of the control parameters in the previous EBW step and β is an adjustable bias parameter.

3.3. Training data division

For the control set, a part of the training data is needed. Excluding that part from the training statistics used for the EBW re-estimation could have negative impact to the acoustic models due to reduced amount of training data. On the other hand, if the control set is naively included as a part of the training set, severe over-training will occur. This problem can be solved by first collecting the training set statistics without the control set and using that in the control procedure. After the EBW constants have been optimized, the control set statistics (now with weakened language model lattices) are computed and added to the training set statistics. This way the EBW update can use the full training set for parameter re-estimation.

For the selection of the control set we considered two scenarios. The first one is a conventional training scenario with matching training and testing data. In this case, it is possible to vary the control set between different EBW iterations, enabling better coverage of the training data while keeping the control set size sufficiently small. For each EBW step before the control procedure, a random set of speakers from the training set is selected as the control set, generating a new control set for every EBW iteration.

Another scenario we considered the controlling method for is when there is a mismatch between the training data and the target task, but there is a small sample of the target data available for training. In addition to incorporating this small sample to the existing training set, we can emphasize its role by using that set as the control set. Compared to the first scenario, this may reduce the amount of data for the control optimization, and hence increase the risk of over-training.

4. Experiments and results

To evaluate the proposed EBW control method, LVCSR experiments with matched and mismatched tasks using MMI and SNFE discriminative training criteria were performed. The speech recognition system developed at the Aalto University [11] was used with standard acoustic model settings, i.e. three-state hidden Markov models and Gaussian mixture emission probabilities with diagonal covariances. The models utilized MFCC features with first and second differentials, having 39 dimensions in total. The ML models were first trained to create initial models and unigram lattices for the discriminative training. Discriminative training consisted of 15 iterations.

All the speech data used in the experiments was from the Finnish Speecon corpus. The training set for the matched scenario contained only clean sentences from 310 speakers. About 15h of speech was used for training. The resulting acoustic models had 24587 Gaussians in 1170 states/mixtures. The performance of the models were evaluated over independent development and evaluation sets, both of which contained 40 speakers, about 1100 utterances and 1.9h of speech. The control set for the matched scenario was on each iteration a random set of about 1100 utterances, about 12% of the training data.

For the mismatched scenario, parallel recordings from the development and evaluation sets were utilized. Whereas the clean sets had been recorded using a headset microphone, medium distance microphone recordings of the same sessions were used as the mismatched data. Due to varying recording environments several types of background noises were present in the data. The noisy development set was split into two parts, 20 speakers each. Only one of the halves was used for development purposes, the other half played the target set and was incorporated in the clean training set for discriminative training. It was also the control set of the mismatched case.

The language model was a high-order morph-based N-gram model [11]. As the number of morphs that constitute a word is not limited, the resulting vocabulary is unlimited. Finnish words are commonly rather long, so to improve the resolution of the error measurements, letter error rate (LER) instead of word error rate (WER) was used for evaluation. For each evaluation, the corresponding development set was used to pick the best model among the 15 discriminative training iterations.

The baseline control method for EBW was the common heuristics, I-smoothing was used for the SNFE ($\tau = 50$). I-smoothing was implemented as smoothing to the previous model [12]. For the first EBW iteration, the control parameters of the proposed method were initialized as the median of the Gaussian-specific D values of the heuristic method. In addition to the proposed control method, a simple EBW control which uses a fixed D over all the EBW iterations was evaluated in the experiments. The fixed D was set to the same initial value as the control method initialization. Compared to some previously applied global D approaches with EBW, the one applied here differs by incorporating $D_{ik}^{\min}$ similar to the common heuristic method. That is, $2D_{ik}^{\min}$ is used as the EBW constant for the Gaussian if the fixed D was below that.

The control methods used maximum of 25 Quickprop iterations for optimizing the control parameters in each EBW step. This roughly doubled the computations required for the discriminative training. Quickprop parameters were set during preliminary experiments to achieve good convergence in the control set optimization. The logarithmic parameter transform caused the learning rate values to be small, value 0.005 was used for the globally clustered case and 0.02 for other cases.

For the control bias parameter, $\beta = 5/N_g$ was used for all cases, where N_g is the number of Gaussians in the model. In the clean training case and a global control parameter, this can be interpreted as an addition of about 0.005 to the SNFE error

Table 1: *Speecon development set results (LER). Corresponding ML model results are 3.0%/10.1% for the clean/noisy task.*

EBW control method	Clean		Noisy target	
	MMI	SNFE	MMI	SNFE
Heuristic	2.9%	**2.6%**	8.5%	7.9%
Fixed D	**2.8%**	**2.6%**	6.7%	7.6%
Global clustering	**2.8%**	**2.6%**	**6.4%**	7.4%
Mixture clustering	**2.8%**	**2.6%**	**6.4%**	**7.3%**
Gaussian level	**2.8%**	2.7%	**6.4%**	7.6%

Table 2: *Speecon evaluation set results (LER). Corresponding ML model results are 3.3%/10.0% for the clean/noisy task.*

EBW control method	Clean		Noisy target	
	MMI	SNFE	MMI	SNFE
Heuristic	3.2%	3.0%	7.8%	6.9%
Fixed D	**3.0%**	**2.9%**	6.3%	6.5%
Global clustering	3.1%	3.0%	**6.0%**	6.4%
Mixture clustering	**3.0%**	**2.9%**	6.3%	**6.3%**
Gaussian level	**3.0%**	**2.9%**	6.3%	**6.3%**

measure for each control sentence if the maximum parameters were chosen instead of the minimum ones. In the mismatched scenario with a smaller control set, similar figure is the addition of 0.01 to the SNFE error. Preliminary experiments showed the need for the bias as otherwise only the global clustering resulted in good optimization of the control set SNFE performance. The bias parameter was picked heuristically, optimal value for it was not pursued.

Table 1 shows the development set results for the different speech recognition tasks. Only the result of the best performing model among the 15 discriminative iterations is shown for each task/method. In the clean case the performance difference between the control methods is small. The noisy target task shows bigger differences. Already the fixed D method improves the results compared to the heuristics. EBW constant optimization is able to improve results even further. The slightly lower performance of the Gaussian level EBW constant optimization in SNFE may be due to too small a control set and over-training.

Table 2 shows the evaluation set results. For the clean MMI case, the 6% relative decrease in LER compared to the heuristic method observed with the Fixed D, Mixture clustering and Gaussian level control methods are statistically significant according to the Wilcoxon signed rank test ($\alpha = 5\%$). For the noisy task it is clearly beneficial to avoid using the usual EBW heuristics. For MMI, the EBW contant optimization with global parameter clustering shows 25% decrease in LER compared to the heuristic method, and 5% decrease compared to the fixed D method, both improvements being statistically significant. All the alternative EBW control methods are statistically significantly better than the heuristic method also for the mismatched SNFE task. The mixture clustered control method achieves the best performance, with 9% relative reduction in LER compared to the heuristic method.

5. Conclusions

Controlling the EBW algorithm by optimization of the D constants over a held-out set provides a well-founded method to a problem often solved by heuristic rules. The large vocabulary speech recognition experiments showed that in conventional clean model training and matched testing the proposed control method can provide small improvements over the heuristic setting of EBW constants. In the case of mismatched testing conditions and the use of a small amount of target data for discriminative training, even bigger benefits were observed. However, already using a fixed D instead of the Gaussian-specific heuristic values could give clear improvements in these cases.

Future research will focus on finding the optimal clustering and constraint settings, and investigating how to lower the computations required for the control set optimization.

6. Acknowledgements

This work was supported by the Academy of Finland in the projects numbered 135003 and 251170, and by Tekes in Perso project.

7. References

[1] P. C. Woodland and D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," *Comp. Speech and Lang.*, vol. 16, pp. 25–47, 2002.

[2] R. Schlüter, W. Macherey, B. Müller, and H. Ney, "Comparison of discriminative training criteria and optimization methods for speech recognition," *Speech Communication*, vol. 34, pp. 287–310, 2001.

[3] S. Axelrod, V. Goel, R. Gopinath, P. Olsen, and K. Visweswariah, "Discriminative estimation of subspace constrained gaussian mixture models for speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 172–189, 2007.

[4] X. He, L. Deng, and W. Chou, "Discriminative learning in sequential pattern recognition," *IEEE Signal Process. Mag.*, vol. 25, no. 5, pp. 14–36, 2008.

[5] V. Valtchev, J. Odell, P. C. Woodland, and S. J. Young, "MMIE training of large vocabulary recognition systems," *Speech Communication*, vol. 22, pp. 303–314, 1997.

[6] J. Pylkkönen, "Investigations on discriminative training in large scale acoustic model estimation," in *Proc. Interspeech*, 2009, pp. 220–223.

[7] D. Povey and P. C. Woodland, "Minimum phone error and I-smoothing for improved discriminative training," in *Proc. ICASSP*, 2002, pp. 105–108.

[8] M. Gibson and T. Hain, "Error approximation and minimum phone error acoustic model estimation," *IEEE Trans. Audio, Speech, Lang. Process.*, no. 6, pp. 1269–1279, 2010.

[9] P. Liu, C. Liu, H. Jiang, F. Soong, and R.-H. Wang, "A constrained line search optimization method for discriminative training of HMMs," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 16, no. 5, pp. 900–909, 2008.

[10] E. McDermott, T. J. Hazen, J. L. Roux, A. Nakamura, and S. Katagiri, "Discriminative training for large-vocabulary speech recognition using minimum classification error," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 15, no. 1, pp. 203–223, 2007.

[11] T. Hirsimäki, J. Pylkkönen, and M. Kurimo, "Importance of high-order N-gram models in morph-based speech recognition," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 4, pp. 724–732, 2009.

[12] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, and K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," in *Proc. ICASSP*, 2008, pp. 4057–4060.