Simplified Simultaneous Perturbation Stochastic Approximation for the Optimization of Free Decoding Parameters

Aleksei Romanenko[1], Alexander Zatvornitsky[2], and Ivan Medennikov[1,3]

[1] ITMO University, Saint-Petersburg, Russia
[2] Speech Technology Center, Saint-Petersburg, Russia
[3] Saint-Petersburg State University, Saint-Petersburg, Russia
183460@niuitmo.ru,
zatvornitskiy@speechpro.com,
ipmsbor@yandex.ru

Abstract. This paper deals with automatic optimization of free decoding parameters. We propose using a Simplified Simultaneous Perturbation Stochastic Approximation algorithm to optimize these parameters. This method provides a significant reduction in computational and labor costs. We also demonstrate that the proposed method successfully copes with the optimization of parameters for a specific target real-time factor, for all the databases we tested.

Keywords: Simplified Simultaneous Perturbation Stochastic Approximation, SPSA, decoding parameter, real-time factor, RTF, speech recognition.

1 Introduction

The balance of accuracy and speed of automatic speech recognition depends on the solution of a number of related tasks, such as:

- optimization of the acoustic model;
- optimization of the language model;
- optimization of a large set of free decoding parameters.

Optimization of both the acoustic model and the language model in automatic speech recognition for large vocabularies is a well-known task [1]. In contrast, the problem of optimizing free decoding parameters is still often solved manually or by using grid search (i.e. searching for values in a grid with a specified step). The task is complicated by the fact that each parameter can have a different impact on the accuracy of speech recognition and/or the expected decoding time. Moreover, each new domain requires searching for new optimal decoding parameters every time we change the training data. Lastly, changing hardware configuration also requires adjustment of optimal decoding parameters.

A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 402–409, 2014.

Typically, the search for optimal decoding parameters that satisfy the constraints of the real-time factor and at the same time provide high recognition accuracy is a very time-consuming task.

In this paper, we present a Simplified Simultaneous Perturbation Stochastic Approximation for optimizing free decoding parameters. The proposed method significantly reduces computational costs in compared to [2], and the reduction is even greater compared to grid search. In contrast to [3] and [4], Simplified SPSA takes into account the real-time factor, which is of vital importance for the design of an ASR system. The proposed method also requires lower computational costs than [1] and [2] for finding the optimal accuracy corresponding to a specific real-time factor. We introduce a penalty function, which is used to achieve a balance between recognition accuracy and decoding time. Then we demonstrate that this method provides robust and fast results. We present results obtained on three speech databases comprising spontaneous and read speech.

2 Simultaneous Perturbation Stochastic Approximation (SPSA)

Let us start by describing the standard form of the SPSA algorithm [5]. We denote the vector of free decoding parameters as θ. Let $\hat{\theta}_k$ denote the estimate for θ at the kth iteration. Then the algorithm has the standard form:

$$\hat{\theta}_{k+1} = \hat{\theta}_k - a_k \hat{g}_k(\hat{\theta}_k) \tag{1}$$

where $\hat{g}_k(\cdot)$ is an estimate for the gradient at the kth iteration. The gain sequence a_k satisfies certain well-known conditions [6], these conditions are necessary for the convergence of the algorithm. a_k is calculated as:

$$a_k = a/(A + k + 1)^{\alpha} \tag{2}$$

In order to determine the "simultaneous perturbation" we perturb each $\hat{\theta}_k$ with a vector of mutually independent, mean-zero random variables Δ_k satisfying the conditions given in [6]. Usually, Δ_k is taken as symmetrically Bernoulli distributed. A positive scalar is calculated as follows:

$$c_k = c/(k + 1)^{\gamma} \tag{3}$$

This positive scalar and mean-zero random variables are multiplied to obtain two new parameter tuples:

$$\hat{\theta}_k^{+} = \hat{\theta}_k + c_k \Delta_k \tag{4}$$

$$\hat{\theta}_k^{-} = \hat{\theta}_k - c_k \Delta_k \tag{5}$$

Using (2) and (3) gain sequences a_k and c_k, SPSA and Kiefer-Wolfowitz finite-difference-based SA (FDSA) [7] achieve the same level of statistical accuracy for a given number of iterations, but SPSA requires p times fewer measurements of the loss function (p is a number of free decoding parameters that are being optimized).

The estimate of the gradient $\hat{g}_k(\cdot)$ is calculated from the values of the loss function $L(\cdot)$, as:

$$\hat{g}_k(\hat{\theta}_k) = \begin{bmatrix} L(\hat{\theta}_k^+) - L(\hat{\theta}_k^-) \Big/ 2c_k\Delta_{k1} \\ \vdots \\ L(\hat{\theta}_k^+) - L(\hat{\theta}_k^-) \Big/ 2c_k\Delta_{kp} \end{bmatrix} \tag{6}$$

The values of the non negative coefficients a, c, A, α and γ can be chosen according to the guidelines given in [6].

3 Simplified SPSA

The standard algorithm is designed so that a_k and c_k decrease with increasing k. If a_k causes a deterioration of the objective value, the optimal solution must stay at $\hat{\theta}_k$ and at the next iteration obtain the estimation of the loss function with a new a_k according to (2). Without an appropriate step size, the optimal solution will stay at $\hat{\theta}_k$ forever, which significantly slows down the rate of convergence of the algorithm [8]. This problem is illustrated in Fig. 1. The optimal solution is obviously located at $\hat{\theta}_k^+$. But the standard step size provides transition to $\hat{\theta}_{k+1}$, where we are faced with the problem described above. If we assume that the $\hat{\theta}_k^+$ point is obtained using an appropriate step size, then we can take $\hat{\theta}_k^+$ as the outcome of the current iteration.

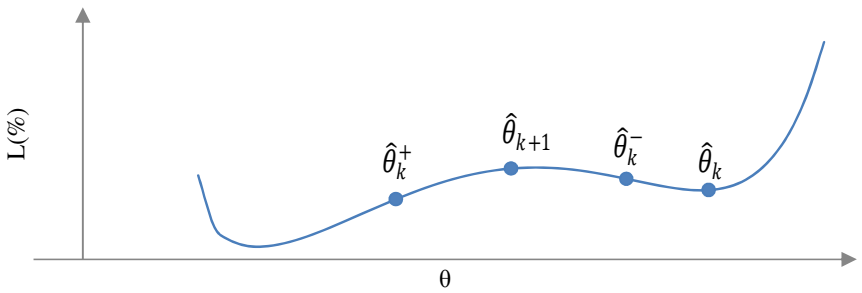

Fig. 1. The search process of the standard SPSA

According to the assumption above, SPSA takes the following form:

$$\hat{\theta}_{k+1} = \begin{cases} \hat{\theta}_k^+, if\ L(\hat{\theta}_k^+) < L(\hat{\theta}_k^-)\ and\ L(\hat{\theta}_k^+) < L(\hat{\theta}_k); \\ \hat{\theta}_k^-, if\ L(\hat{\theta}_k^-) < L(\hat{\theta}_k^+)\ and\ L(\hat{\theta}_k^-) < L(\hat{\theta}_k); \\ \hat{\theta}_k, in\ all\ other\ cases. \end{cases} \tag{7}$$

Moreover, if the parameter vector did not change at the current iteration, it means that the algorithm is close to the optimal point. In order to increase the convergence rate, it is necessary to reduce the coefficient c_k using the equation $c = c/1.5$.

The initial value of the parameter c must be chosen so that the coefficient c_k could converge to a certain minimum value in an expected number of iterations, giving the distance between the vectors of parameters $\hat{\theta}_k^+$ and $\hat{\theta}_k^-$ such that $\left|L(\hat{\theta}_k^+) - L(\hat{\theta}_k^-)\right| > 0$.

To take the decoding speed into account, we will calculate the loss function, penalizing it by the corresponding value of RTF (real-time factor). Then the loss function takes the form:

$$L(\cdot) = L(\cdot) + RTF \tag{8}$$

This function provides the tradeoff between the real-time factor and the accuracy of speech recognition.

The algorithm obtains an optimal solution, but this solution does not satisfy the desired real-time factor. We propose increasing/decreasing the parameters stepwise to change the speed of automatic speech recognition, in order to achieve the desired real-time factor. The step size and a set of parameters are specific for each decoder.

4 Setup

For the experiments we used three databases:

- Database A: recordings of read speech prepared by a collaborating speaker, the topic is sports, 1:06h, 5257 words, maximum accuracy obtained by manual parameter tuning is 92.505 at RTF= 0.357;
- Database B: recordings of telephone conversations (spontaneous speech), 0:49h, 2828 words, maximum accuracy obtained by manual parameter tuning is 62.694 at RTF= 0.253;
- Database C: recordings of internet broadcasts, webinars and podcasts, 0:40h, 3013 words, maximum accuracy obtained by manual parameter tuning is 62.297 at RTF= 0.864.

For each of these databases we have a corresponding language and acoustic model [9,10]. We are tuning the following parameters:

- max_hyp_num – maximal number of hypotheses;
- thr_common – common threshold;
- lm_scale –factor of the weight of any edge of the graph;
- wd_add –addition to the weight of the edge of the graph.

Speech recognition was carried out by the ASR system developed at Speech Technology Center Ltd. All experiments were performed on a workstation with an Intel Core i5 Desktop Processor with 4 physical cores, and 32 GB of RAM.

5 Experiment

We performed several tests for each of the databases with different target real-time factors. Table 1 shows the results obtained by the algorithm, and further improved by selecting the parameters that affect the decoding speed.

Table 1. Accuracy and real-time factor results on all databases, for the Simplified SPSA

database	objective RTF	initial indicators		output indicators		#iterations	#runs of decoder
		Acc	RTF	Acc	RTF		
A	0.1	91.021	0.115	92.505	0.090	22	42
A	0.3			92.581	0.154	23	43
A	0.5			92.619	0.321	25	45
B	0.1	50.636	0.309	60.962	0.074	22	42
B	0.2			62.023	0.178	21	41
B	0.3			62.553	0.273	22	42
C	0.5	47.063	0.823	60.438	0.499	25	45
C	0.7			61.401	0.679	25	45
C	0.9			61.998	0.892	26	46

In all the tests for all the databases the proposed method showed high efficiency. It managed to approach to the optimal values obtained manually, and sometimes exceeded them. All the results are within the confidence interval. Figures 2 and 3 show the results for databases A and C. We can see that a considerable improvement of Acc and RTF occurs already at the early iterations. After the twentieth iteration, the algorithm is aimed at selecting a specific target real-time factor.

6 Conclusions

In this paper, we demonstrated an effective method of optimizing free decoding parameters, which enabled us to obtain the optimum for a specific target real-time factor. The method can be applied to finding the optimal parameters for a specific target factor for all the databases we tested. All our results are within the confidence interval so they can be considered optimal. In practice, this approach allows us to obtain the parameters better than by grid search, and at the same time at a lower computational cost.

We are confident that the proposed method can be used for other decoders to optimize free decoding parameters in terms of a specific target real-time factor. To do that, it is necessary to form the parameter vector, and the subset of parameters that influence the speed of speech recognition.

Acknowledgements. This work was partially financially supported by the Government of the Russian Federation, Grant 074-U01.

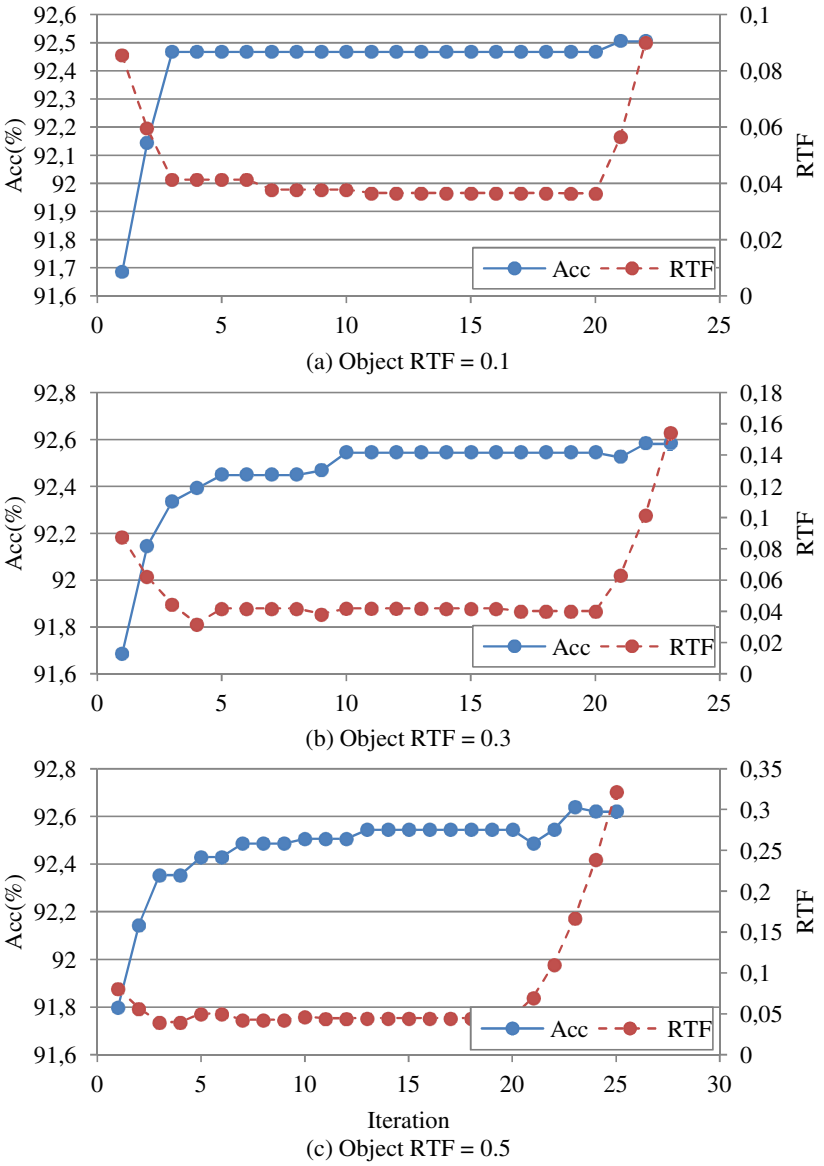

Fig. 2. Optimization runs on the database A

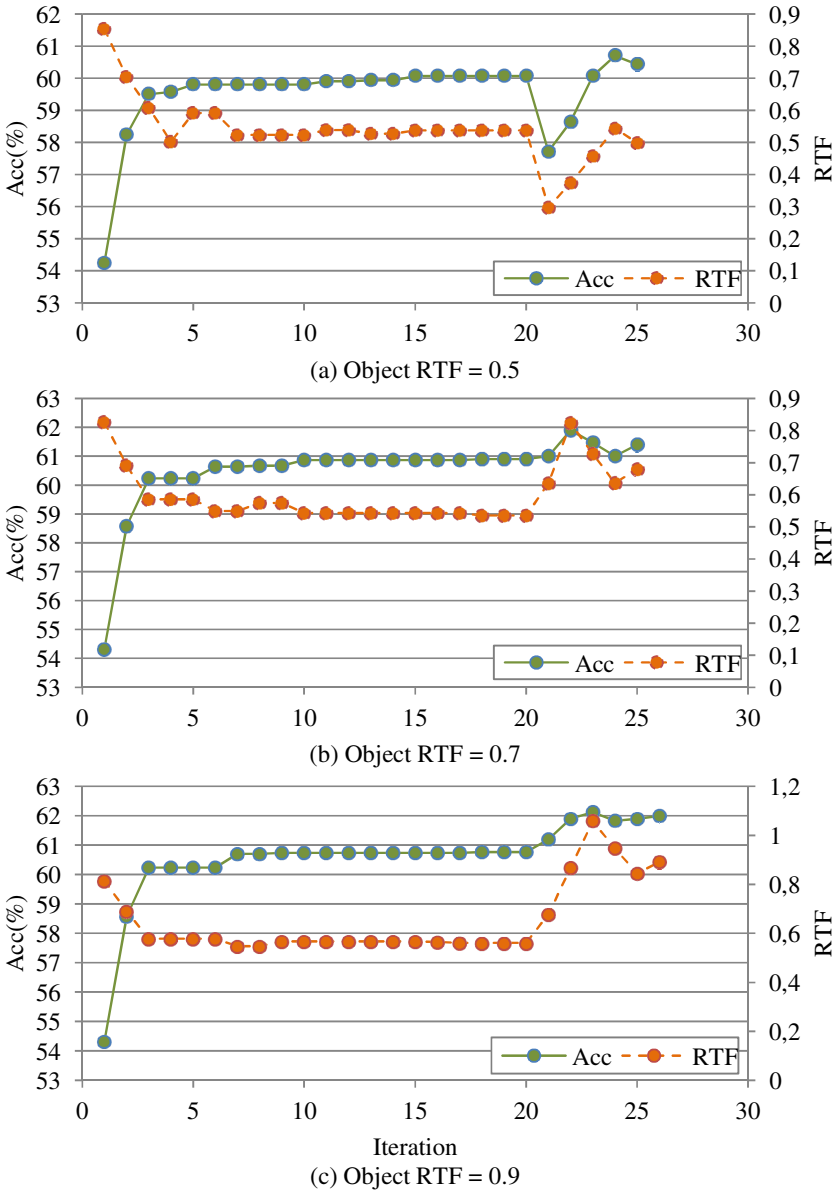

Fig. 3. Optimization runs on the database C

References

1. Stein, D., Schwenninger, J., Stadtschnitzer, M.: Simultaneous perturbation stochastic approximation for automatic speech recognition. In: Proc. of the INTERSPEECH, Lyon, France, August 25-29, pp. 622–626 (2013)
2. El Hannani, A., Hain, T.: Automatic optimization of speech decoder parameters. Signal Processing Letters 17(1), 95–98 (2010), doi:10.1109/LSP.2009.2033967
3. Mak, B., Ko, T.: Automatic estimation of decoding parameters using large-margin iterative linear programming. In: Proc. of the INTERSPEECH, Brighton, United Kingdom, September 6-10, pp. 1219–1222 (2009)
4. Kacur, J., Korosi, J.: An accuracy optimization of a dialog ASR system utilizing evolutional strategies. In: Proc. of the ISPA, Istanbul, Turkey, September 27-29, pp. 180–184 (2007)
5. Spall, J.C.: Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. IEEE Transactions on Automatic Control 37(3), 332–341 (1992), doi:10.1109/9.119632
6. Spall, J.C.: Implementation of the simultaneous perturbation algorithm for stochastic optimization. IEEE Transactions on Aerospace and Electronic Systems 34(3), 817–823 (1998), doi:10.1109/7.705889
7. Kiefer, J., Wolfowitz, J.: Stochastic Estimation of the Maximum of a Regression Function. Ann. Math. Stat. 23(3), 462–466 (1952)
8. Yue, X.: Improved Simultaneous Perturbation Stochastic Approximation and Its Application in Reinforcement Learning. In: Proc. of the International Conference on Computer Science and Software Engineering, Wuhan, Hubei, December 12-14, vol. 1, pp. 329–332 (2008)
9. Korenevsky, M., Bulusheva, A., Levin, K.: Unknown Words Modeling in Training and Using Language Models for Russian LVCSR System. In: Proc. of the SPECOM, Kazan, Russia, September 27-30, pp. 144–150 (2011)
10. Yurkov, P., Korenevsky, M., Levin, K.: An Improvement of robustness to speech loudness change for an ASR system based on LC-RC features. In: Proc. of the SPECOM, Kazan, Russia, September 27-30, pp. 62–66 (2011)