Approaches to Parameter Tuning for ASR Systems: LVCSR and Expert Grammars*

◎ Josef Novak Nobuaki Minematsu Keikichi Hirose
The University of Tokyo, Tokyo, Japan
{novakj,mine,hirose}@gavo.t.u-tokyo.ac.jp

1 Abstract

In this work we look at the impact and appeal of different approaches to parameter tuning for Automatic Speech Recognition systems. This is a significant issue in the development and comparison of practical as well as research-oriented ASR systems which is nonetheless rarely afforded attention commensurate with its importance. Here we provide a brief discussion of several popular automated approaches to the tuning process.

2 Introduction

Modern Automatic Speech Recognition (ASR) systems support a large number of tunable parameters. The CMU-based decoder Sphinx3 supports over 100, while other popular decoders such as HDecode, HVite, BBN, etc. support between 20 and 65. These parameters can take a variety different forms including real-valued functions, integer values and boolean states, etc. Parameters also affect different aspects of the decoding process including pruning, model weights, GMM processing, speaker adaptation and cache sizes.

Unfortunately parameters are often interdependent and their sheer number thus makes it impractical to compute optimal settings for each new LVCSR application using brute-force grid search. This is made more difficult in a mixed environment of LVCSR and expert grammar systems as these typically have very different operating points. There has been some growing interest in finding efficient, automated solutions to the tuning problem but there is still little in the way of a consensus as to what the most principled approach might be.

3 Approaches to Parameter Tuning

There are several different iterative approaches to automated tuning which have appeared in the literature. The most pragmatic of these focus on the relationship between Word Error Rate (WER) and Real-Time Factor (RTF).

In [1], the problems is framed as the joint optimization of a 2-D objective function which comprises two components, WER $R(\theta)$ and RTF $W(\theta)$, based on the parameter vector θ. This approach depends on a cost function, which is constructed on the basis of the dependent RTF, and takes the form,

$$C(t) = \frac{W(\hat{\theta}_{opt}(t-1) - \alpha k(t-1)) - W(\hat{\theta}_{opt}(t-1))}{R(\hat{\theta}_{opt}(t-1) - \alpha k(t-1)) - R(\hat{\theta}_{opt}(t-1))}. \quad (1)$$

Here $\alpha(t) = \alpha(0)e^{(-\lambda t)}$ defines the step-size as an exponential decay function, proportional to the RTF and α and λ are experimental constants. In experiments using HTK this approach converged in a reasonable average of 30 iterations. However it focused on only 6 parameters.

In [2] the problem is limited to two parameters: the language model factor, K_{gf}, and the word insertion penalty, K_{wip}, and framed as a discriminative training problem where the solution focuses on a linear programming (LP) approach. This approach is appealing in that experiments showed that only 5-6 iterations were required for convergence, and LP solvers are readily available to perform the necessary optimization computations. Nevertheless, computational complexity for mixed integer linear programming, or non-linear programming problems such as this typically grow exponentially in the number of variables, and it is not clear how well this approach would scale to the full set of variables available to many modern ASR systems.

The approach in [3] focuses on a genetic programming solution utilizing a stochastic optimization strategy. Here the parameter vector θ is mutated at each iteration according to a stochastic process, $\theta' = \theta + N(0, \delta)$, where N represents a Gaussian distribution with 0 mean and δ variance. Fitness of an entity is determined by WER, RTF, or a combination thereof. This strategy produces a high qual-

*音声認識システムのパラメーターチューニングへのアプローチ：LVCSR や記述文法
ノバックジョセフ、峯松信明、広瀬啓吉 (東大)

ity θ, but does not scale well to a large number of parameters.

In [4], a more heuristic algorithm is introduced which, like [1] aims to strike an explicit balance between RTF and WER, but unlike the approaches above was put into practice with a very large set of parameters. This approach takes advantage of the fact the WER and RTF are almost always monotonic in all the tunable parameters, implying that the problem has only a small number of local optimums. The solution that the authors propose focuses on the theory of Lagrange multipliers which says that, assuming the functions $WER(\theta)$ and $RTF(\theta)$ are both smooth, and WER is minimized at $RTF(\theta) = S$, then any local parameter vector θ should yield a λ such that,

$$\frac{\partial WER}{\partial \theta_i} - \lambda \frac{\partial RTF}{\partial \theta_i} = 0 \qquad (2)$$

The optimization algorithm then consists in iteratively perturbing individual parameter values, C_i in the parameter vector, θ by small increments, ϵ_i, such that with the performance metric set to $PR(\lambda, \theta) = -WER(\theta) - \lambda RTF(\theta)$ then,

$$PR(\lambda, \theta') - PR(\lambda, \theta) \qquad (3)$$

is negative for all $\theta' = \theta \pm \epsilon_i e^i$ with e^i the vector where only i-th parameter is non-zero. In the event that the perturbation results in a performance improvement, then the baseline vector θ is updated and the algorithm then proceeds to the next iteration.

Experiments reported on this approach covered 63 parameters for the BBN LVCSR system, and resulted in significant improvements in terms of WER for specific values of S. The maximum number of iterations was fairly high at 304 decoding runs, however this compares quite favorably with the other solutions discussed earlier given the number of parameters in play. Furthermore the approach facilitates parallelization and affords a simple and straightforward implementation.

One final optimization approach which is worthy of note is that proposed recently in [5], which focuses on search error risk minimization for viterbi beam search in the WFST framework. This optimization strategy seeks to optimize the hypothesis pruning step by introducing a more precise data-driven pruning function that utilizes the rich features extracted from hypotheses during an additional tuning phase. The method employs a batch-style algorithm and gradient descent to iteratively update pruning function parameters based on lattice output. This method was shown to be quite effective at pulling-forward the RTF vs. Word-Accuracy (WACC) curve, reducing the minimum RTF for a given WER, yet did not affect the overall best accuracy.

4 Conclusion

This paper summarized several different approaches to parameter tuning for ASR systems. It is difficult to compare these methods directly as they frame the optimization problem with a variety of different mathematical models, however it is possible to make a several general observations. One important variable is the ratio of iterations required for convergence to the number of parameters. It is also reasonable to expect that the LP-based approaches will be infeasible for a large number of parameters, do in part to the uptick in computational complexity. The heuristic method of [4] demonstrably wins out on both these points, and because the solution supports arbitrary parameters - real-valued, boolean, integer, etc., and affords a simple implementation there is a strong argument for promoting this approach in practice. The technique of [5] however, also promises to tighten the results of any general parameter tuning result, which suggests that an optimal strategy, particularly in the case of an on-the-fly WFST-decoder, might involve a combination of these two approaches. Although there was not sufficient time to implement this combination, we intend to use this analysis as a basis to implement and empirically test this in the near future.

参考文献

[1] Hannani *et al.*, "Automatic Optimization of Speech Decoder Parameters," IEEE Proc. Letters, 95-98, 2010.

[2] Mak *et al.* "Min-max Discriminative Training of Decoding Parameters Using Iterative Linear Programming," Proc. Interspeech 2008, 915-918, 2008.

[3] Kacur *et al.*, "Accuracy Optimization of a Dialog ASR System Utilizing Evolutional Strategies," Proc. ISPA, 180-184, 2007.

[4] Colthurst *et al.*, "Parameter Tuning for Fast Speech Recognition," Proc. Interspeech 2007, 1477-1480, 2007.

[5] Hori *et al.*, "Search Error Risk Minimiation in Viterbi Beam Search for Speech Recognition," Proc. ICASSP 2010, 4934-4937, 2010.