

基于词图重估的语音解码参数优化方法

尹明明, 屈丹, 李弼程, 黄山奇

(解放军信息工程大学信息工程学院, 郑州 450002)

摘要: 在大词汇连续语音识别系统中, 语言模型权值和插入代价等语音解码参数对系统的识别率有较大的影响, 而在实际应用中常通过实验手动调整其值寻求最佳识别结果。为此, 提出一种利用二元文法进行词图重估的方法, 自动优化语音解码参数。在重估的参数空间搜索过程中采用线性搜索与模拟退火搜索相结合的方法, 使优化参数具有全局最优和对初值稳定性强的优点。实验结果表明, 相比凭经验设置的参数, 该方法估计出的参数值能大幅降低识别词错误率, 与经典的 N-best 优化相比, 其优化速度有较大提升。

关键词: 词图重估; 语言模型权值; 解码; 插入代价

Speech Decoding Parameters Optimization Method Based on Word Graph Rescoring

YIN Ming-ming, QU Dan, LI Bi-cheng, HUANG Shan-qi

(Institute of Information Engineering, PLA Information Engineering University, Zhengzhou 450002, China)

【Abstract】 In Large Vocabulary Continuous Speech Recognition(LVCSR) system, speech decoding parameters——Language Model(LM) weight and insertion cost can greatly affects the recognition performance. But in practice, they are usually hand-tuned through experiment to obtain best recognition performance. This paper proposes the rescoring based method that uses bi-gram LM to optimize the parameters automatically, meanwhile the method of combine line search and Simulated Annealing(SA) search in parameters search space of rescoring, which is globally optimal and is insensitive to initial value of parameter. Experimental results show that the method can dramatically reduce the word error compared with empirical parameter setting method, and gains much faster optimization speed than classical N-best optimization.

【Key words】 word graph rescoring; Language Model(LM) weight; decoding; insertion cost

DOI: 10.3969/j.issn.1000-3428.2011.16.054

1 概述

连续语音识别系统目前主要通过统计模式识别进行搜索解码, 这是一种基于贝氏网络并结合声学模型和语言模型的识别框架, 语音识别即是在这种框架下寻求最佳的词序列。声学模型目前主要由连续密度的隐马尔科夫模型(Continuous Density Hidden Markov Model, CDHMM)表示, 语言模型主要由 N-gram 的语言模型表示。由于识别系统中实际产生的描述声学语言的数学模型不够准确, 并且声学得分与语言模型得分的动态范围相差较大, 因此常需要采用权衡因子(LM Scale)对声学得分和语言模型得分进行均衡; 同时为了减少词插入错误, 引入了词插入代价的参数。

这 2 种参数在实际中常通过实验中手动调整来优化, 直到目前为止对参数自动优化的研究很少。

文献[1]对 N-best 的识别结果进行重打分, 通过实验得到词错误率(Wer)大小对参数进行优化。但是, 这种参数的优化是手动调整的, 所需时间太长, 且往往不能达到最优。

文献[2]通过对 N-best 进行预选后进行优化使优化的速度大大提高。

文献[3]提出了利用区分性训练得到的正确解与候选解的差值, 运用递归线性规划的方法对其差值方程进行优化, 估计 2 种参数。显然, 这种方法只需进行线性规划的优化, 比较简单, 但是其需要对语料进行区分性训练, 并且容易使参数的收敛陷入局部最优, 影响其优化的效果, 在一定程度上限制了它的应用。

文献[4]将 2 种参数表示成 Log 线性模型, 再通过递归的

生成词图进行重估计算其最大后验概率, 直到其收敛为止。这种方法能够得到较优的参数, 且对不同的初始值具有一定的稳定性, 但这种方法需要反复生成词图, 对于大词汇连续语音识别系统来说其速度显然是不能容忍的。

本文提出了一种基于词图重估的语音解码参数优化方法。新方法利用词图重估, 并结合线性搜索与模拟退火(Simulated Annealing, SA)的算法, 自动优化语言模型权值与插入代价。与文献[3]相比, 这种方法无需特殊的训练方法, 简洁且精度高, 同时也避免了文献[4]需要反复生成词图的缺点。在采用线性搜索与 SA 相结合的方法后能有效地避免 SA 容易陷入局部最优的缺点, 实现全局最优的优化。实验结果表明, 该方法不仅能提高识别率, 而且对不同的初始值具有一定的稳定性。

2 N-best 优化估计

N-best 优化估计是指对识别生成的 N 个结果进行重估, 通过识别率的反馈优化参数。

假设语音集中句子数为 M , 加入语言模型权值与插入代价参数的第 i 句话中识别的第 j 个候选的总分为 $f(i, j, \lambda_L, \lambda_I)$, 则:

基金项目: 国家“863”计划基金资助项目(2006AA01z146)

作者简介: 尹明明(1986—), 男, 硕士研究生, 主研方向: 语音识别; 屈丹, 副教授、博士; 李弼程, 教授、博士生导师; 黄山奇, 硕士研究生

收稿日期: 2010-12-30 **E-mail:** hiyingmingming@gmail.com

$$f(i, j, \lambda_L, \lambda_I) = f_A(i, j) + \lambda_L f_L(i, j) + \lambda_I n_w(i, j) \quad (1)$$

其中, $f_A(i, j)$ 为第 i 句话中识别的第 j 个候选句的总声学得分; $f_L(i, j)$ 为第 i 句话中识别的第 j 个候选句的总语言学得分; λ_L 、 λ_I 分别表示语言模型权值与插入代价。语音识别的 Wer 表示为:

$$Wer = \frac{\sum_{i=1}^M E(i, \theta_i(\lambda_L, \lambda_I))}{\sum_{i=1}^M N_i} \quad (2)$$

其中, N_i 表示第 i 句话的标准识别结果的词的个数; $\theta_i(\lambda_L, \lambda_I)$ 表示加入语言模型权值与插入代价参数后的第 i 句话的最佳候选句, 即:

$$\theta_i(\lambda_L, \lambda_I) = \arg \max_j f(i, j, \lambda_L, \lambda_I) \quad (3)$$

其中, $E(i, j)$ 为第 i 句话中识别的第 j 个候选句中插入/删除/替换的单词数。

获得的最优化参数为:

$$\lambda_{L, I}^o = \arg \min_{\lambda_L, \lambda_I} Wer = \arg \min_{\lambda_L, \lambda_I} \sum_{i=1}^M E(i, \theta_i(\lambda_L, \lambda_I)) \quad (4)$$

3 基于词图重估的语音解码参数优化

词图作为大词汇连续语音识别系统的中间识别结果的一种表示形式, 词图结构能紧凑地表示更多的候选识别结果。

图 1 是语料中第 i 句话的词图的一种简化表示形式, 图中圆点表示词图的节点, 边表示词图的弧, h 表示词图的开始节点, f' 表示词图的结束节点。

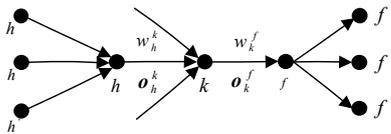


图 1 词图的简化表示

从图中可以看出词图中弧上保存了声学得分, 则第 i 句话的得分可表示为:

$$f_A(i) = \sum_{h, k=h}^{f'} p(o_h^k | w_h^k) \quad (5)$$

$$f_L(i) = \sum_{h, k, f=h}^{f'} p(w_k^f | w_h^k) \quad (6)$$

其中, o_h^k 、 w_h^k 是 h 节点到 k 节点之间弧的观测向量与词 ID。则根据式(1)、式(4)可以得到优化的参数。

虽然词图也是通过保存多个候选识别结果生成的, 但是词图结构能紧凑地表示更多的候选识别结果, 相比 N-best 能更快捷地重估出最优结果, 本文将选择词图作为参数优化的中介。目前词图已广泛用于二次解码, 广泛的研究也保证了重估的速度与精度, 最终按图 2 所示的过程进行优化。

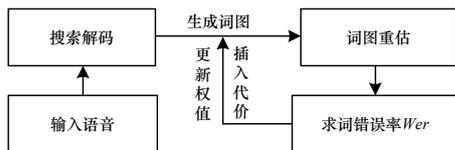


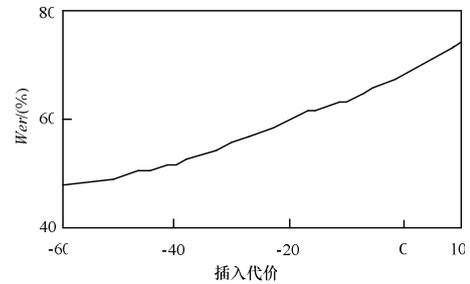
图 2 基于词图重估优化过程

3.1 优化估计思想

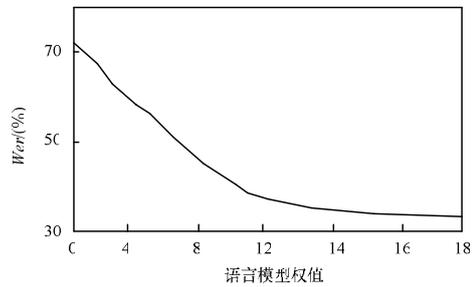
假设设定搜索区间, 如果直接采用线性搜索对所有的候选值搜索, 精度较高, 但这种线性搜索策略所需时间非常长, 尤其是搜索步长较小时耗时更多。增大步长虽然可以减少重估次数, 但显然会使参数的估计精度不高, 因此, 如何结合

两者的优点进行参数评估是需要重点解决的问题。

图 3 给出了参数 λ_L 、 λ_I 对词错误率的影响曲线, 从图中可以看出, 2 种参数对识别率的影响呈现一种近似于单调的变化趋势, 根据这一特性, 可以利用步长选出一些较优点, 再对每个较优点的邻域范围内进行搜索, 得到最终的优化参数, 即将全局最优与局部最优相结合进行优化, 提升速度。



(a) 参数 λ_L



(b) 参数 λ_I

图 3 参数 λ_L 、 λ_I 单独作用对词错误率的影响

3.2 优化估计流程

假设权值变量集为向量 $A = (\lambda_L, \lambda_I)$, 词错误率为 $Wer(t)$, t 为循环的次数。

实际的参数范围设为 $\lambda_L \in (0, 20)$, $\lambda_I \in (-60, 10)$, 初始线性搜索设定步长为 $stepL = 2$, $stepI = 4$ 。初始 $A_0 = (-60, 0)$, 经过线性搜索统计出前 n 个 Wer 较优的向量集: $A_{opt} = (A_{opt}^1, A_{opt}^2, \dots, A_{opt}^n)$, 其中, $A_{opt}^n = (\lambda_{L, opt}^n, \lambda_{I, opt}^n)$ 。

显然上述优化是保持一种参数不变而对另一种参数的变化, 这也等同于对每个参数单独优化, 在设定步长进行全局线性搜索后, 得到一系列较优点。由图 3 可知, 当保持一种参数不变时, 另一种参数是一种近似单调的变化, 那么对这些较优点邻域进行搜索即能得到最优参数点。如果减小步长继续对每个较优点邻域进行线性搜索显然其所需时间是不能接受的, 这时需要对 2 种参数同时进行优化, 即需要采用组合优化算法对参数进行优化。由于 SA 描述简单且运行效率较高, 运用 SA 算法能使参数很快收敛到邻域范围内的最优, 本文选择 SA 算法优化得到最优解, 同时, 只需寻找到较优点的邻域范围内的最优, 这样在小范围邻域的搜索也避免了 SA 容易陷入局部最优的缺点。

结合上面的公式可以得到如下估计算法:

(1) 设定主循环初值 $t = 1$, t 为较优点序号, 设定 A 的初始值为 A_{opt}^1 , 根据初始 A_{opt}^1 进行词图得分重估, 得到 $Wer(1)$ 。

(2) 设 $T = 100$, $k = 0$, T 为 SA 搜索的初始温度, k 为设定的 SA 循环次数。

(3) 在 A_{opt}^t 的邻域内选择一个新 A_t' , 计算 $Wer(t)$, 则

$\Delta Wer = Wer(t) - Wer(t-1)$ 。

(4) 如果 $\Delta Wer(t) < 0$ ，则 $A_{opt}^t = A_i^t$ ， $Wer'(t) = Wer(t)$ ， $A_{best}^t(t) = A_{opt}^t$ ；如果 $\Delta Wer(t) > 0$ ，并且 $\exp(-\Delta Wer/T) > rand(0,1)$ ，则 $A_{opt}^t = A_i^t$ ， $Wer'(t) = Wer(t)$ ， $A_{best}^t(t) = A_{opt}^t$ ，若小于，则不接受；如果 $\Delta Wer(t) = 0$ ，所有值不变。

(5) $T(k+1) = \sigma T(k)$ ， σ 是小于 1 的正常数， $k = k+1$ ，继续步骤(3)。

(6) $k \geq 10$ 时， $t = t+1$ ，继续步骤 2，当 $t > 20$ 时，循环终止。

(7) 统计出 $Wer'(t)$ 中最小的值，将对应的 $A_{best}^t(t)$ 赋到 A_{best} 中，即 $A_{best} = A_{best}^t(t)$ 。

4 实验

4.1 实验条件

为了测试算法的有效性，随机从微软语料库 Speech Corpora (Version 1.0) 中选择 200 段不间断语句作为参数重估的训练集，500 句不间断的语句作为测试集。实验的特征是在窗长为 25 ms、帧移为 10 ms 下提取的 39 维 MFCC，其中包含 12 维 MFCC、一维能量谱以及它们的一阶与二阶差分。声学模型采用音节模型，每个状态混元数取 3，语言模型采用二元文法统计模型。为了使生成词图的初始参数对最终优化的结果影响最小，对训练集中的语料保存 N-best (本实验选取 $N=1000$) 候选来生成词图，进行重打分后得到最优参数，最后通过测试集语料进行评估。同时为了验证初始值选取对结果的影响，在生成词图时选取了 3 组不同的参数初始值进行测试，即 $\{A_0\} = \{(\lambda_1^0, \lambda_2^0)\} = \{(0, 1), (-30, 8), (-10, 5)\}$ 。

实验采用的平台是 HTK 的 Hvite 及以 HLRescore、HResults 为主搭建而成的评估系统，通过对 Hvite 生成的词图^[5]，HLRescore 重估参数，HResults 统计插入删除及替换的错误词数，计算词错误率，集合三者与同一平台实现对参数的自动估计。

4.2 实验结果及分析

4.2.1 优化结果分析

表 1 表示线性搜索法搜索参数集合得到的收敛参数组合。其中，第 1 列表示不同的初始值；其余列表示在不同的初始值下按词错误率从小到大的较优参数集，限于篇幅，每种初始值下选取 3 个较优参数集。

表 1 不同初始值得到的较优参数顺序

| 初始值 | 1 | 2 | 3 |
|----------|----------|----------|----------|
| (-30, 8) | (-52, 8) | (-48, 8) | (-32, 8) |
| (-10, 5) | (-52, 8) | (-48, 8) | (-36, 8) |
| (0, 1) | (-48, 8) | (-48, 6) | (-40, 6) |

对经过线性搜索得到的一组较优的参数，本实验取前 6 个参数进行 SA 搜索。对其邻域搜索，各初始值得到的最终收敛值分别为 (-49.7, 10.7)、(-49.7, 10.6)、(-48.05, 6.24)，可以看出，初始参数值不同，生成的词图对最终的结果会有影响，但是这种影响几乎可以忽略不计，也验证了算法是可行的。进一步观察表 1 还可以看出，在前 6 个较优的点集中，语言模型权值大多为 8，而插入代价与词错误率一致单调，而且对于同一个语言模型权值参数，不同的插入代价的值对应的词错误率都较小，这些特点说明只需对 (-52, 8) 邻域进行搜索即可得到最优解。实验表明对 (-52, 8) 邻域进行搜索得到收敛参数为 (-49.7, 10.7)，可见这种方法完全可以得到最优解。

4.2.2 优化时间分析

从表 2 可以看出，本文的方法相对于文献[6]中直接对 N-best 进行重估的时间 500 s 提升较大，但是文献[2]中对 N-best 结果预选后的 N-best 重估只需 10 s，所以，要对词图做进一步的优化，并对步长做改变等来提升速度。虽然比优化的 N-best 的速度慢很多，但是由于未舍弃任何解，其精度比文献[2]的方法高，且由于属于训练阶段确定的参数，170 s 的时间也能够接受。

表 2 3 种不同方法的所需时间

| 方法 | 优化所需时间/s |
|--------------------|----------|
| 本文方法 | 170 |
| 文献[1]的经典 N-best 估计 | 500 |
| 文献[2]的优化 N-best 估计 | 10 |

4.2.3 参数测试

随机选取测试集中的 500 句话对收敛的参数进行识别，对于阈值的设置仍采用与初始的阈值相同的值，并与经典的 N-best 优化得到的优化参数 (-53, 9) 比较，结果如表 3 所示。

表 3 不同的参数对识别结果的影响

| 参数值 | Wer (%) |
|---------------|---------|
| (-49.7, 10.7) | 53.26 |
| (-53, 9) | 54.21 |

从表 3 可以看出，在初始值 $A_0 = (-30, 8)$ 下得到的参数 (-49.7, 10.7) 识别的错误率最小，相比 N-best，优化精度略有提高，在实验中也发现，如果直接用系统默认的参数即 (0, 1)，识别的词错误率是完全不能够接受的，因为其插入错误太多，以至于几乎没有正确的词，这也说明参数设置的重要性。

5 结束语

尽管语言模型权值与插入代价知识只是语音解码中众多参数的一组参数，但是从表 3 可以看出，选择较好的参数对识别率的提升具有至关重要的作用。本文从词图重估方法入手，结合线性搜索与模拟退火 2 种搜索方式，对参数的搜索空间进行搜索找出最优的参数值。实验结果证明这种方法的稳定性与精确性，对测试集的实验也表明，这种方法能使识别率大幅提升。

语音解码的参数众多，本文探讨的只是在静态搜索空间中优化这 2 种参数。文献[5]提出了对动态搜索空间所有的阈值参数进行优化的方法，这种优化方法通过搜索过程的逐段优化来不断更新阈值参数，直到收敛为止，如何结合其方法从整体上对这些参数进行全面优化，将是下一步需要研究的重点与难点。

参考文献

[1] Horbe Y, Minematsu N, Nakagawa S. Theoretical/Experimental Investigation of Balance Between Acoustic Model Likelihood and Language Model Likelihood[J]. IPSJ SIG Notes, 2000, 54(3): 67-72.

[2] Ito A, Kohda M, Makino S. Fast Optimization of Language Model Weight and Insertion Penalty from N-best Candidates[J]. Acoustical Science and Technology, 2005, 26(4): 384-387.

[3] Mak B, Ko T. Min-max Discriminative Training of Decoding Parameters Using Iterative Linear Programming[C]//Proc. of the 9th Annual Conference of the International Speech Communication Association. Brisbane, Australia: [s. n], 2008: 915-918.