

文章编号:1001-9081(2005)02-0291-03

语音识别错误的分类分析

付跃文,杜利民

(中国科学院 声学研究所,北京 100080)

(fuyw@iis.ac.cn)

摘 要:大词表连续语音识别系统由多个组件构成,识别错误受多种因素的影响。系统开发者需要分析错误发生的不同原因。根据语音识别的基本理论给出了对错误进行分类分析的原理,将识别错误按错误原因分为解码错误、声学模型错误、语言模型错误、声学 and 语言复合错误四大类,并对分类后的错误做了统计分析。实验证明,识别错误的分类分析为系统的改进提供了参考依据。

关键词:大词表连续语音识别;识别错误;分类

中图分类号:TP391.42 **文献标识码:**A

Classification analysis of speech recognition errors

FU Yue-wen, DU Li-min

(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100080, China)

Abstract: Large vocabulary continuous speech recognition system consists of several components, and recognition errors are caused by different factors. Developers need to know how the errors occur. In this paper we deduced the principle to classify the recognition errors from the recognition theory and put each error according to its cause into one of four classes: the decoding, the acoustic model, the language model, and the acoustic and language model. We then performed statistical analysis of classified errors. Experiments show that classification analysis of recognition errors provides guidance for improving the system.

Key words: large vocabulary continuous speech recognition; recognition error; classification

0 引言

在研究大词表连续语音识别的过程中,对于语音识别系统的识别错误,人们通常仅对总的识别错误率做出统计,以便对系统本身和系统的某些改进做出评价,但是对于系统的开发者来说,除了知道识别错误率外,还应知道错误发生的具体原因。由于语音识别系统由声学模型、语言模型和解码器多个模块构成,其性能受几个方面因素的影响,当发生错误时,引起错误的模块各不相同。因而对识别错误进行分类分析,评价组件及算法的性能,从而为改进系统提供依据,就成为必要的工作^[1]。我们从识别的基本理论出发,说明了识别错误可以按照引起错误的模块进行分类,并提出了进行分类的具体方法,给出了我们的分类分析及应用的试验结果。所用的分类分析方法及分类分析应用方法具有一般性。

1 基本原理和实现方法

1.1 错误分类的基本原理

大词表连续语音识别过程的标准表述为:给定声学信号

$$\begin{aligned} p(X|W) &= p(X|w_1 w_2 \cdots w_i) = p(X|[h_{11} h_{12} \cdots h_{1p}]_{w_1} [h_{21} h_{22} \cdots h_{2q}]_{w_2} \cdots [h_{i1} h_{i2} \cdots h_{iu}]_{w_i}) = \\ &= p(x_1 x_2 \cdots x_n | [s_1 s_2 \cdots s_{k_1}]_{w_1} [s_{k_1+1} s_{k_1+2} \cdots s_{k_2}]_{w_2} \cdots [s_{k_{i-1}+1} s_{k_{i-1}+2} \cdots s_{k_i}]_{w_i}) = \\ &= [p(x_1 | s_1) p(x_2 | s_{q_1}) \cdots p(x_{i_1} | s_{k_1})]_{w_1} [p(x_{i_1+1} | s_{k_1+1}) p(x_{i_1+2} | s_{k_1+2}) \cdots p(x_{i_2} | s_{k_2})]_{w_2} \cdots \\ &= [p(x_{i_{i-1}+1} | s_{k_{i-1}+1}) p(x_{i_{i-1}+2} | s_{k_{i-1}+2}) \cdots p(x_n | s_{k_i})]_{w_i} \end{aligned} \quad (3)$$

上式的下标中的 p, q, \cdots, r 分别表示词 w_1, w_2, \cdots, w_i 的 HMM 模型的总数, k_1, k_2, \cdots, k_i 分别表示词 w_1, w_2, \cdots, w_i 的状

X , 求最可能产生该声学信号的词序列 W 。利用贝叶斯准则:

$$p(W|X) = \frac{p(X|W)p(W)}{P(X)} \quad (1)$$

识别结果为上述后验概率公式取最大值时的词序列,又因为 $P(X)$ 与词序列是相互独立的,所以可以不予计算,因此识别结果可以表述为:

$$\hat{W} = \arg \max_W p(X|W)p(W) \quad (2)$$

公式(2)的第一部分 $p(X|W)$ 是词序列 W 的声学概率部分,即词序列 $W = w_1, w_2, \cdots, w_i$ 的声学模型产生声学信号 X 的概率。声学信号由若干按时间顺序排列的声学特征矢量构成($X = x_1 x_2 \cdots x_n$)。每个词被分解为由发音基本单位构成的读音序列。每个发音基本单位(音素、音子或半音节)是一个声学模型,现有的主流声学模型是 HMM 模型(用 h 表示),每个 HMM 模型由若干状态(用 s 表示)构成,每个状态按照一定概率产生声学特征矢量。该词序列所有的声学模型状态连接在一起,构成一个大的声学模型状态序列,然后用该状态序列即可计算产生声学信号 X 的概率。表述如下:

收稿日期:2004-07-21;修订日期:2004-12-22 基金项目:国家 973 计划项目(G1998030505)

作者简介:付跃文(1968-),男,山西孝义人,博士研究生,主要研究方向:信号处理与语音识别;杜利民(1957-),男,四川营山人,研究员,博士生导师,主要研究方向:信号处理与语音交互技术。

态总数。

公式的(2)的第二部分 $p(W)$ 是词序列 W 的语言模型概率部分。表示词序列 $W = w_1, w_2, \dots, w_i$ 在整个语言空间的发生概率。在大词表识别中,语言模型概率计算被表述为:

$$p(W) = p(w_1 w_2 \dots w_i) = p(w_1) p(w_2 | w_1) p(w_3 | w_1 w_2) \dots p(w_i | w_1 w_2 \dots w_{i-1}) \quad (4)$$

限于实际的统计和计算能力,实践中用 n -gram统计语言模型来近似上述计算,即认为某词发生的概率仅受其前 $n-1$

$$\begin{aligned} \hat{W} = \arg \max_W \{ \log p(X | W) p(W) \} &= \arg \max_W \{ \log p(X | W) + \log p(W) \} = \\ \arg \max_W \{ [\log p(x_1 | s_1) + \log p(x_2 | s_{q1}) + \dots + \log p(x_{i1} | s_{k1})]_{\text{声学得分}} + [\log p(w_1)]_{\text{语言得分}} \}_{w_1} + \\ [\log p(x_{i1+1} | s_{k1+1}) + \log p(x_{i1+2} | s_{q2}) + \dots + \log p(x_{i2} | s_{k2})]_{\text{声学得分}} + [\log p(w_2 | w_1)]_{\text{语言得分}} \}_{w_2} + \dots + \\ [\log p(x_{i-1+1} | s_{k_{i-1}+1}) + \log p(x_{i-1+2} | s_{q_i}) + \dots + \log p(x_n | s_{k_i})]_{\text{声学得分}} + [\log p(w_i | w_{i-2} w_{i-1})]_{\text{语言得分}} \}_{w_i} \end{aligned} \quad (6)$$

一个大词表连续语音识别系统的基本组成如图1所示,该系统用来实现(6)式的计算。由于语言空间是庞大的,无论采用何种搜索算法,例如 Viterbi 算法或是堆栈译码算法,进行全部词序列空间的搜索计算是不可能的,必须进行大量的剪枝,即依据某种准则在还未完成全部搜索计算时就提前放弃对于某些序列路径的搜索,只搜索最有希望的一些路径。这种剪枝会在有时造成高概率的词序列在解码中被丢弃。

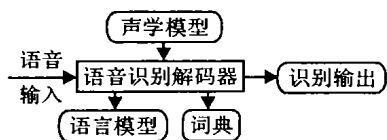


图1 语音识别系统基本构造

在解码器输出结果之后,将识别结果词序列和参考文本词序列比较,就可以得出其中的不一致的部分,即识别错误区域,一个区域可以是一个词或是数个词。当我们按照图2的流程比较错误区域和相应的参考文本区域的声学 and 语言概率得分情况时,可以将该错误区域分成几种不同的类型,同时查明了识别错误的原因是系统的何种模块。经过分类之后,错误被分为了解码类型错误、语言模型和声学模型复合错误、声学模型错误、语言模型错误四大类。在对错误类型进行整个识别结果集合的统计之后,为系统的进一步改进提供了参考。

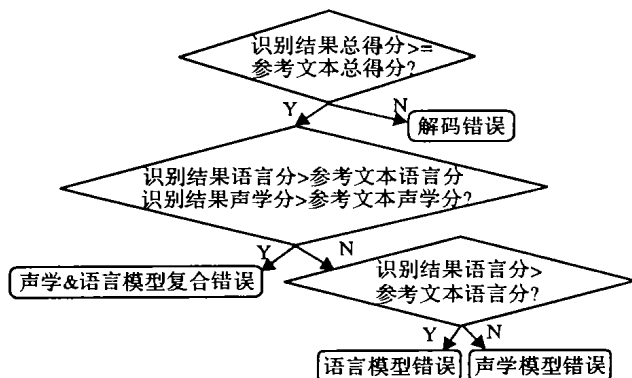


图2 错误区域分类流程

1.2 分类分析的具体实现

1.2.1 构造基本信息文件

对于同一语音输入,参考文本和对应的识别结果都存在一个用(6)式计算的结果。对于识别结果来说,由于无论何种解码器,都已经在解码过程中进行了实际计算,所以从解码过程当中就可得到识别结果每个词的声学概率得分和语言概

个词的影响。以 trigram 近似为例,语言模型表述为:

$$p(W) = p(w_1 w_2 \dots w_i) = p(w_1) p(w_2 | w_1) p(w_3 | w_1 w_2) \dots p(w_i | w_{i-2} w_{i-1}) \quad (5)$$

在实际计算的时候,概率取对数值进行计算,将各对数概率值称为得分。由(2)式,(3)式和(5)式,识别结果如(6)式,可以看出,每词的概率总得分由声学得分和语言得分两部分构成:

率得分。对于参考文本,由于解码过程事先未知其结果,所以无法从解码过程直接获得,需要单独进行 Viterbi 强制校准搜索(forced alignment)^[2]后得出(6)式要求的每词的声学得分,然后再查找语言模型得到每词的语言得分。

1.2.2 错误区域的提取

比较识别结果和参考文本的相同的时间段内是否具有相同的词。相同时间段内具有不同词语的区域,即为错误区域,相邻的错误区域合并为一个。

1.2.3 错误类型的划分和统计

错误区域类型的划分依照图2的流程图进行。在得出错误区域类型后,最基本的统计是统计识别结果整个集合内错误区域的总数,各类错误的数目,以及错误类型的分布比例。

2 分类分析及应用实验

2.1 实验条件

我们在实验中所用的大词表连续语音识别系统,其声学模型为音子 HMM 模型,语言模型为 n -gram 模型,解码器由两遍搜索过程构成。第一遍搜索为 viterbi-beam 搜索,第二遍搜索采用堆栈译码算法。语音测试数据取自 863 语音识别语料库,四男四女共八名说话人(f93-f96, m93-m96),5146 句语音。

2.2 单句的错误提取和分类

按照上述方法对识别结果做了错误提取和错误类型的划分。图3是单个句子分析后的一个实例。图中提取了错误区域(时间区间 325 ~ 385)。在错误区域内,参考文本(REF)的语言概率得分(LM)大于识别结果(HYP)55分,但识别结果的声学概率得分(AC)大于参考文本56分,总的得分(TOT)识别结果大于参考文本1分,因此解码器选择了识别结果中的错误的词。该类错误属于声学模型错误。单句分析后,进行统计分析。

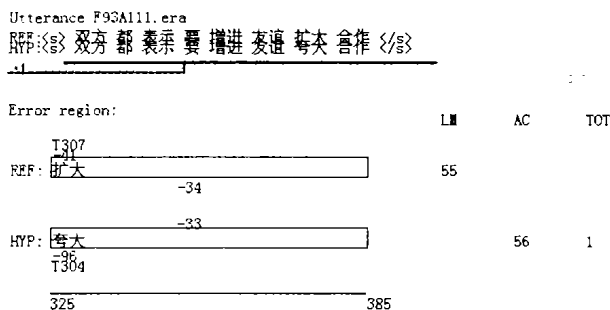


图3 错误区域及错误类型分析结果屏幕截图

2.3 分类分析后的统计分析用于系统诊断

表 1 中试验 1 的数据是我们在解码方法试验中对 5 146 句语音的识别结果进行错误分类和统计后的结果,统计了各类错误的数目以及分布比例。

表 1 863 语音库测试集中八名话者在两次试验中的错误分布比例比较

试验编号	识别率 (%)	错误区域比例 (总数)	解码错误比例 (数目)	声学错误比例 (数目)	语言错误比例 (数目)	声学 & 语言错误比例 (数目)
试验 1	76.7	100 (4 970)	52.8 (2 623)	10.1 (504)	22.7 (1 128)	14.4 (715)
试验 2	78.9	100 (4 628)	43.0 (1 989)	12.0 (555)	27.5 (1 272)	17.5 (812)
错误比例变化(%)	2.2	-6.9	-24.2	9.2	12.8	13.6

2.4 错误区域提取和分类用于语言模型分析

表 2 识别试验语言模型 Trigram 各类型分布百分比统计

统计项目	T1	T2	T3
整个参考文本中语言模型各类型的百分比	16.9	46.1	37.1
语言模型类错误区域对应的参考文本区域中语言模型各类型的百分比	48.2	40.6	11.3
语言模型类错误区域中语言模型各类型的百分比	31.6	51.0	17.4
错误区域对应的参考文本区域中语言模型各类型的百分比	39.0	41.0	20.0

表 2 是对 8 名说话人 5 146 句语音的识别结果进行错误分类后统计错误区域内语言模型信息的结果。语言模型信息经查找语言模型获得。我们在识别时使用的模型为 trigram 模型。T1 表示 trigram 退化成了 unigram,T2 表示 trigram 退化为 bigram,只有 T3 表示完整的 trigram 信息。从统计的结果来看,在发生语言模型类型错误的区域中以及对于整个错误区域,对应的参考文本中 unigram 的比例都很高(48.2%和

从表 1 可以看出,解码错误所占比例较大,提示需要设法改善解码措施。在对解码方法做了修改后的试验 2 中,解码错误的比例下降了 24.2%。对错误类型进行比例统计可以观察错误主要模块来源。

39.0%),即语言模型统计信息严重不足。该项统计指示了识别错误与语言模型的相关性,同时提示开发者,对于错误区域内尤其语言模型错误区域内的词,需要考虑重新增加相应语料进行语言模型训练,以便获得相应足够的语言模型信息,用于提高识别率。

3 结语

本文根据大词表连续语音识别的基本理论公式,导出了错误分类的基本依据,将每个错误区域分类为解码错误、声学错误、语言模型错误、混合错误四大类,建立了错误分类分析的一般方法。实验证明,错误的分类分析有助于诊断识别系统并改进识别系统的缺陷。所建立的方法具有一般性。

参考文献:

[1] CHASE L. Error-Responsive Feedback Mechanisms for Speech Recognizers [D]. Carnegie Mellon University: 1997.65.
[2] YOUNG S, KERSHAW D, ODELL J, et al. The HTK Book (v3.0) [M]. Cambridge University Engineering Department: September 2000.

(上接第 290 页)

2) 过多的冗余。假设集合点中的 RPV 一致性特别差(每一个集合点中的 RPV 都和其他的不同),对于不同集合点发出的相同查询有可能有不同的映射点。根据我们的改进方案,索引将会被复制到许多集合点。这就增加了网络中的索引冗余。最差的情况是在极小的一段集合点中有大量的冗余。我们可以通过引入一个最小复制距离 L 来改善这种情况,即映射点和命中点的遍历距离等于或超过 L 时,我们才采用这种复制。也就是说,在一定距离内的遍历,我们是可以忍受的。这个 L 就是在索引冗余度和查询效率之间的折衷。当 L 设定得太小时,就趋向原始 JXTA2.0 规范中的查询机制;而当 L 设定的过小时,可能产生较大的索引冗余。因此,这个 L 应该能够根据不同的网络特性或用户要求进行调整。

2.2 改进后查询效率的分析

查询效率是 JXTA 网络性能的一个重要组成部分,其中一次远程查询的耗时比本地查询高一个数量级,并且随着网络规模的增大,两者耗时的差距不断增加^[5]。由于 P2P 网络本身特性所限,以及 RPV 的不严格一致,索引查询请求在网络中的有限范围巡查不可避免。一个远程查询请求在 RPV 中巡查时,需要消耗大量的网络带宽和机器处理时间,而且这些消耗随着巡查距离的增加而增加。更糟糕的是,在以后同样的查询请求要重复这个过程。

改进后的动态索引分布,目的是减少索引的巡查次数。在最理想的情况下,索引在经过一次巡查后就被复制到映射点之上,以后关于这个索引的查询就由映射点直接返回查询结果。即使在一般的情况下,索引也是向它的映射点方向移动,一定程度上减少了巡查的距离。因此,改进后的查询方案比原来单一的有限范围巡查机制有明显的优势。

参考文献:

[1] TRAVERSAT B. Project JXTA 2.0 Super - Peer Virtual Network [EB/OL]. <http://www.jxta.org/project/www/docs/JXTA2.0protocols1.pdf>, 2003-05.
[2] TRAVERSAT B, ABDELAZIZ M, POUYOULE E. A Loosely - Consistent DHT Rendezvous Walker[EB/OL]. <http://www.jxta.org/project/www/docs/jxta-dht.pdf>, 2003-03.
[3] RATNASAMY S, FRANCIS P, HANDLEY M, et al. A Scalable Content Addressable Network[EB/OL]. <http://netweb.usc.edu/cs51f02/papers/ratnasamy01scalable.pdf>, 2001-05.
[4] LV Q. Search and Replication in Unstructured Peer-to-Peer Network [EB/OL]. www.cs.princeton.edu/~qlv/download/searchp2p_full.pdf, 2002-06.
[5] HALEPOVIC E, DETERS R. The Costs of Using JXTA[EB/OL]. http://bosna.usask.ca/pub/P2P03_Halepovic_CostsOfUsingJXTA.pdf, 2003-10.