一种改进的语音识别词错误率评估算法

吴 边,兰时勇,刘重庆

(上海交通大学 图像处理与模式识别研究所,上海 200030)

摘要:在建立语音识别系统的过程中错误率评估起着非常重要的作用,传统的词错误率算法仅仅是基于最小错误率,具有显著的缺陷,因而不能准确评估系统的错误率。提出一种改进的基于最小错误率和时间信息的词错误率评估算法,能够准确评估系统的错误率,为声学模型的优化提供指导,同时列举了该评估算法在建立语音识别系统过程中的应用。

关键词:模式识别; 词错误率; 时间信息; 语音识别

中图分类号:TP391.42 **文献标识码**:A **文章编号**:1007-2276(2005)01-0106-04

Improved Word Error Rate evaluation algorithm for automatic speech recognition

WU Bian, LAN Shi-yong, LIU Chong-qing

(Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: The error rate evaluation is very important in building an Automatic Speech Recognition (ASR) system. The conventional algorithm for Word Error Rate (WER) evaluation is based on the minimum error rate. The improved WER algorithm is proposed on the basis of minimum error rate time information, which makes the WER evaluation on the overall system more accurate. Then the acoustic model then can be improved according to the evaluation.The applications of the new algorithm in building a practical ASR system is also presented.

Keywords: Pattern recognition; Word Error Rate; Time information; Automatic speech recognition

0 引 言

在建立语音识别系统的过程中,评价语音识别系统的性能是非常重要的。通常,通过错误分析,剔除对声学模型训练有害的数据,优化声学模型的训练以达到改进整个系统性能的目的。另一方面,可以通过测试系统的词错误率,分析错误的原因,有针对性地解决问题,改进系统的性能。错误率分析在语音识别系统的研究和开发中起着重要的作用。在建立一个与说话人无关命令识别系统的过程中,需要一个人工神经网络用于拒识所有不"清晰"的命令,而神经网络的训练要用到所有识别正确的词,要求能为识别结果中每个词的识别正确性进行准确的标注,产生这些标注是错误率评估的另一作用。

收稿日期:2004-03-10; **修订日期**:2004-05-10

作者简介:吴边(1977-),男,江西南昌人,博士生,从事模式识别、语音识别、多媒体技术方面的研究,在 SCI 检索期刊上发表论文2篇。

1 问题的提出

目前词错误率被广泛用于评价语音识别系统性能，词错误率能够直接反映识别系统声学模型的性能,也是其他评估指标如串错误率(句错误率)的基础。传统的词错误率评估算法[1]在语音识别中存在三种典型的词错误:

(1) 替换错误(Substitution)

在识别结果中,正确的词被错误的词代替;

(2) 删除错误(Deletion)

在识别结果中,丢失了正确的词;

(3) 插入错误(Insertion)

在识别结果中,增加了一个多余的词。

在传统的词错误率评价算法中,词错误率为:

$$WER=\frac{S+D+I}{T}\times 100\% \qquad (1)$$

式中 S 为替代错误词数;D 为删除错误词数;I 为插入错误词数;T 为参照句子中所有的词数。

传统的算法可以描述为:假设参照的句子表示为 $w_1,w_2,w_3,\cdots,w_m$,其中 w_i 表示第 i 个词,识别的句子是 $\hat{w}_1,\hat{w}_2,\hat{w}_3,\cdots,\hat{w}_m$,其中 $\hat{w}_j$ 表示第 j 个词。

第一步初始化:

$$R[0,0]=0 \quad B[0,0]=0$$

第二步生成评估矩阵:

for $i=1,\cdots,n$ {

for $j=1,\cdots,m$ {

$$R[i,j]=\min\begin{pmatrix} R[i-1,j]+1 \\ R[i-1,j-1] \\ R[i-1,j-1]+1 \\ R[i,j-1]+1 \end{pmatrix}\begin{matrix} ,B[i,j]=\text{Deletion} \\ ,B[i,j]=\text{Match} \\ ,B[i,j]=\text{Substitution} \\ ,B[i,j]=\text{Insertion} \end{matrix}$$

}}

第三步回溯:

最小错误率路径$(s_1,s_2,s_3,\cdots,s,0)$

$s_1=B[n,m]$

for $t=2,\cdots,$ until $s_t=0$

$$s_t=\begin{pmatrix} B[i-1,j] \\ B[i,j-1] \\ B[i-1,j-1] \\ B[i-1,j-1] \end{pmatrix}\begin{matrix} ,\text{case } s_{t-1}=\text{Deletion} \\ ,\text{case } s_{t-1}=\text{Insertion} \\ ,\text{case } s_{t-1}=\text{Match} \\ ,\text{case } s_{t-1}=\text{Substitution} \end{matrix}$$

$$WER=\frac{R(n,m)}{n}\times 100\%$$

可见传统的算法是基于错误率最小准则的,而没有深入细致的分析。由这种评价算法得到的词错误率,只能粗略地反映语音识别系统的错误率,对进一步改进整个系统性能的作用有限，有时甚至会给出错误的结果。语音信号不仅仅是字符串,更重要的是一个时间序列，传统的错误率评估算法忽略了语音识别结果的时间信息，因为同样的一个识别句子与参考句子相比时，考虑时间信息和忽略时间信息会得到完全不同的评估结果。

表 1 中 start 和 end 分别指起始帧数和结束帧数,一般情况下帧长为 10 ms,前三列显示的是人工标记的结果,后三列是通过语音识别系统输出的结果,识别出来的结果按照传统的方法评估(Res.1)发现所有

表 1 识别标记和参考标记的比较

Tab.1 Comparison of the recognition and reference sign

Start	End	Label	Start	End	Label	Res.1	Res.2
0	17	sil	0	15	sil	M	M
17	51	6	15	51	6	M	M
51	84	5	51	129	5	M	A
84	127	5	129	143	5	M	I
127	148	sp	143	150	sp	M	M
148	198	3	150	197	3	M	M
198	229	6	197	227	6	M	M
229	266	0	227	266	0	M	M
266	294	4	266	294	4	M	M
294	324	sil	294	323	sil	M	M

的标记完全匹配(M)。通过引入时间信息,发现中间的两个 5 的识别出现了错误,第一个 5 吸收了实际的两个 5, 而第二个 5 完全是插入的错误，将中间的短停顿(short pause,sp)的一部分识别成了 5。因此新的评估算法的结果(Res.2)发现了一个吸收错误(A)和一个插入错误(I),至少 5 的声学模型的训练出现了问题,可能 sp 的声学模型也存在问题。通过以上的实验分析发现,传统的算法忽略了时间信息,基于错误率最小准则会得出错误的词错误率评价结果。如果是为了评价词错误率,传统的算法是可以接受的,但为每个识别结果作出标注不仅是为了得到错误率,而也是希望通过标注发现在声学模型训练过程中的问题,以达到优化声学模型的目的,显然仅给出错误

率是很难达到这一目的的。

2 基于时间信息的词错误率评估算法

基于时间信息的词错误率评估算法引入了许多判别准则。在时间标记上识别结果不可能和人工标记的结果完全一样,允许在一定范围内的误差。另一方面如表 1 中第一个 5 的时间段,尽管它包含实际第一个 5 的全部时间段,也不能认为这个识别结果是正确的。设 S 和 E 分别表示参考标记中一个词的起始帧数和结束帧数, $\hat{S}$和$\hat{E}$分别表示识别结果中同一个词的起始帧数和结束帧数, 则这四个变量间存在以下六种关系:$S<E<\hat{S}<\hat{E}$;$S<\hat{S}<E<\hat{E}$;$S<\hat{S}<\hat{E}<E$;$\hat{S}<\hat{E}<S<E$;$\hat{S}<S<\hat{E}<E$;$\hat{S}<S<E<\hat{E}$。

两个词的重叠部分可以表示为:

$$\text{Overlap}(w_i,\hat{w}_j)=\min(E_i<\hat{E}_j)-\max(S_i,\hat{S}_j) \tag{2}$$

定义词的分段正确率为:

$$SAR=\frac{\text{Overlap}(w_i,\hat{w}_j)}{E_i-S_i}\times 100\% \tag{3}$$

在引入时间信息以后, 新的评估算法能够检测到四种错误,即替代错误、删除错误、插入错误和吸收错误。此时词错误率为:

$$WER=\frac{S+D+I+A}{T}\times 100\% \tag{4}$$

新算法的第一步和第二步与传统的算法相同,

第三步回溯:

最小错误率路径$(s_1,s_2,s_3,\cdots,s,0)$

映射关系$(P_1,P_2,P_3,\cdots,P_m)$

$s_1=B[n,m]$

$e=R[n,m]$

$\alpha=\arg\limits_i s_1,\beta=\arg\limits_j s_1$

$P_\beta=\alpha$

for $t=2,\cdots$until $s_t=0$ {

if s_{t-1} not Deletion {

$\alpha=\arg\limits_i s_{t-1},\beta=\arg\limits_j s_{t-1}$

if Overlap $(w_\alpha,\hat{w}_\beta)<0$ and $E_\alpha<\hat{S}_\beta$

s_{t-1}=Insertion

if Overlap $(w_\alpha,\hat{w}_\beta)<0$ and $S_\alpha>\hat{E}_\beta$

s_{t-1}=Deletion

}

$$s_t=\begin{pmatrix} B[i-1,j] \\ B[i,j-1] \\ B[i-1,j-1] \\ B[i-1,j-1] \end{pmatrix}\begin{matrix} ,\text{case } s_{t-1}=\text{Deletion} \\ ,\text{case } s_{t-1}=\text{Insertion} \\ ,\text{case } s_{t-1}=\text{Match} \\ ,\text{case } s_{t-1}=\text{Substitution} \end{matrix}$$

$\alpha=\arg\limits_i s_t,\beta=\arg\limits_j s_t$

$P_\beta=\alpha$

if $R[\alpha,\beta]>e$ $e=R[\alpha,\beta]$

}

$$WER=\frac{e}{n}\times 100\%$$

从第三步可以得出实际的路径和实际的错误率,但这并不是最终的标注, 下一步要根据时间信息修正识别结果中的标注。 在算法的第三步同时得到了识别结果中的每个词与参考标记中的每个词的映射关系 $(P_1,P_2,P_3,\cdots,P_m)$,由于两个标记间的差异,映射并不是一一对应的关系。

对于在参考标记中的任意一个词 w_i, 在识别标记中可能存在多个词与其对应$(\hat{w}_{j1},\hat{w}_{j2},\cdots,\hat{w}_{jt})$,要选择最佳匹配的一个词与其对应, 其他的词设为 Insertion,最佳匹配遵循以下规则:

(1) 如果仅存在一个词满足$\hat{w}_{jt}=w_i$,则将这个词设为 Match,其他的词设为 Insertion;

(2) 如果存在多个词满足$\hat{w}_{jt}=w_i$,比较这些词的 $\text{Overlap}(w_i,\hat{w}_{jt})$,选择具有最大 $\text{Overlap}(w_i,\hat{w}_{jt})$的词作为与 w_i 对应的词,则将这个词设为 Match,其他设为 Insertion;

(3) 如果不存在任何一个词满足$\hat{w}_{jt}=w_i$,比较这些词的 $\text{Overlap}(w_i,\hat{w}_{jt})$、选择具有最大 $\text{Overlap}(w_i,\hat{w}_{jt})$的词作为与 w_i 对应的词,将其设为 Substitution,其他设为 Insertion;

(4) 如果存在一个词满足$\hat{S}_{jt}<S_i$ 且$\hat{E}_{jt}<E_{i+1}$,则将其设为 Absorbent。

根据以上描述,提出的词错误率评估算法是基于时间信息和最小错误率的一种算法。以上所述的规

则可以根据应用目的和任务的不同进行修改，现在的规则是为了给一个为 OOV (Out-Of-Vocabulary)[2,3] 拒识进行确性度评估[4]的神经网络的训练提供参考标记而制定的。

3 应用及结论

本文提出的新算法可直接用于评估语音识别系统的错误率，并可根据评估结果为声学模型的优化提供指导。新算法给出了所有识别出的标记和参考标记比较后得到的标注，并且得到了平均的分段准确率和每个词的分段准确率。准确率越高表示这个词的模型训练越好,反之则越差。

将新算法用于评估一个中文数字串语音识别系统的声学模型，发现在所有的情况下本文提出算法的评估结果比传统算法的评估结果差，表明原来有些错误的识别结果被误判为正确识别。中文的 0~9 共十个数字,为每个数字都建立一个声学模型,一共有 12 个声学模型(其中数字 1 和 2 有两个发音)。评估的结果发现 5 的分段准确率(SAR)较低,大多数是由于多个连读的 5 造成的,连读的 5 产生大量的吸收错误。有两种解决办法,将 5 的声学模型单独进行进一步的训练，并增加包含连读的 5 的数据的样本数量；也可以为连读的两个 5 训练对应的声学模型,因为我们发现两个连读的 5 对声学模型的影响最大。以上两种方法都建立在分析新算法产生标注的基础上，而传统算法给出的标注由于没有时间信息而不准确。

词错误率评估算法可对识别结果进行标注,通过引入时间信息使得新算法产生的标注完全准确,可以得到所有识别正确的词。我们可以用这些词的语音信号产生的特征训练一个人工神经网用于确信度评估。这个确信度可以用于 OOV 的判别。OOV 词的检测在语音识别系统中是非常重要的。对于命令的识别,如果在命令识别系统中未引入确信度评估,传统识别系统会用命令集中的命令去“猜测”这个命令,可能会产生不可预知的结果。通过这个人工神经网络评估,可以拒识 OOV 词,即拒识所有未通过确信度评估的词。

综上所述,错误率评估算法在建立语音识别系统的过程中起着关键的作用，评估语音识别系统的错误率对改进整个系统的性能，使系统实用化都是非常重要的。文中列举了在建立语音识别系统中利用错误率评估算法的两个应用。传统的错误率评估算法是基于最小错误率的，只能对整个识别系统的性能粗略评估,对于改进整个系统性能的作用有限,同时它给出标注也是基于错误率最小的，其中会包含许多的错误。新的错误率评估算法同时基于错误率最小和时间信息,能够准确地评估系统的性能,同时根据对每个声学模型的 SAR 的计算（文中每个词有一个声学模型），为改进声学模型的训练提供指导，另外可以提供比传统算法更为准确的标注。

参考文献:

[1] Huang X D, Acero A Hon H W. Spoken Language Processing, A Guide to Theory, Algorithm and System Development [M]. Upper Saddle River,New Jersey:Prentice Hall,2001.

[2] Hazen T J, Bazzi I. A comparison and combination of methods for OOV word detection and word confidence scoring[A].Proceedings of the International Conference on Acoustics, Speech and Signal Processing[C].2001.

[3] 徐明星,郑方,吴文虎,等.连续语音关键词识别系统的拒识方法研究[J].清华大学学报(自然科学版),1998,38(S1):89-91.

[4] Wang H,Lin Y. Error-tolerant spoken language understanding with confidence measuring [A].Proceedings of the International Conference on Spoken Language Processing[C].2002.

《红外与光电系统手册》简介

《红外与光电系统手册》(内部资料)共八卷。主要内容:第一卷《红外辐射》、第二卷《辐射的大气传输》、第三卷《光电元器件》、第四卷《光电系统设计、分析和测试》、第五卷《被动光电系统》、第六卷《主动光电系统》、第七卷《光电对抗系统》、第八卷《新系统和技术》。有需求者请与本编辑部联系。