DOI: 10.16511/j.cnki.qhdxxb.2006.01.025

ISSN 1000-0054 清华大学学报 (自然科学版) 2006年 第 46卷 第 1期 25/41
CN 11-2223/N J Tsinghua Univ (Sci & Tech), 2006, Vol. 46, No. 1 94-97

基于上下文相关置信度打分的语音确认方法

孙 辉, 郑 方, 吴文虎

(清华大学 计算机科学与技术系, 智能技术与系统国家重点实验室, 北京 100084)

摘 要: 口语对话系统中,集外词的存在会引起很多识别错误,为了有效地发现并拒绝集外词,提高系统性能,研究利用置信度打分进行语音确认的方法,发现并拒绝识别错误。提出上下文相关的置信度特征,充分考虑当前待确认词与其前序词和后序词之间的相关性。实验结果表明: 上下文相关的置信度特征能够很好地提高拒识性能,对符合识别文法的句子,错误拒绝率为 2.5% 或 5% 时,对比没有使用上下文相关的置信度特征时,错误接受率分别下降了 29% 和 36%; 基于置信度打分的语音确认策略在拒识性能上优于系统已有的在线垃圾模型。

关键词: 信息处理; 声音识别; 置信度; 语音确认; 口语对话系统

中图分类号: TP 391; TP 391.42 文献标识码: A

文章编号: 1000-0054(2006)01-0094-04

Confidence scoring using context dependent features for word verification

SUN Hui, ZHENG Fang, WU Wenhu

(State Key Laboratory of Intelligent Technology and System, Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China)

Abstract In a spoken dialogue system, out-of-vocabulary words will lead to recognition errors. A confidence scoring strategy for word verification was developed to detect and reject such recognition errors. Context-dependent confidence features were used based on the information in the previous and following words. The results show that the context dependent confidence features greatly improve the word verification performance. For in-grammar sentences, the false acceptance rate is reduced by 29% at a false rejection rate of 2.5% and by 36% at a rejection rate of 5%, compared with a system not using the context-dependent confidence features. The verification strategy based on confidence scoring also has better performance than the verification method using the online filler model.

Key words information processing; speech recognition; confidence measures; word verification; spoken dialogue system

对于这些特定领域的对话系统来说,识别词表总是有限的,在实际应用中不可避免地会碰到词表外的词,集外词的存在引起了很多识别错误,因此如何拒绝集外词就成为设计系统时必须考虑的问题。

很多对话系统的识别前端是基于关键词识别技术的。在关键词识别中,采用补白模型来吸收集外词是比较流行的方法[1-3]。一般说来,主要有 3种补白模型: 一是集外补白模型,用训练数据中除关键词之外的其他部分训练补白模型,当关键词发生变化时,这种补白模型需要重新训练; 另一种补白模型由子词模型直接构成或拼接组合而成,当关键词发生变化时,无须重新训练补白模型,具有较好的灵活性,但需要的训练数据较大; 还有一种方法用在线补白模型来吸收集外词,不建立特定的补白模型,而是在搜索过程中动态地形成一个补白,同关键词进行竞争,此方法计算简单,且有一定的抗噪性。口语对话系统得意自动总机 (d-Ear Attendant)[4]原先正是采用了这种拒识方法。

与补白模型在识别同时吸收集外词的方法不同,基于置信度打分的语音确认方法需要在识别结束后对结果进行二次确认: 评价所有候选结果的可靠性,拒绝那些置信度得分比较低的识别候选。基于置信度打分的语音确认中可以利用多种置信度特征,本文在研究常用置信度特征的基础上,提出上下文相关的置信度特征,考虑上下文对当前待确认词的影响情况,提高了语音确认的性能。

1 置信度特征

本文用到了两类置信度特征: 第一类特征仅仅描述了当前待确认词的置信程度,称为局部置信度特征; 第二类特征是在局部置信度特征基础上提出的上下文相关的置信度特征,主要考虑了当前待确认词与其相邻词之间相关性。

收稿日期: 2005-01-18

作者简介: 孙辉 (1977-), 女 (汉), 山东, 博士研究生。

通讯联系人: 郑方, 教授, E-mail: fzheng@tsinghua.edu.cn

1.1 局部置信度特征

选取那些具有区分度的置信度特征用于语音确认这一问题已有大量的研究[5,6]。根据使用的知识源的不同,置信度确认特征大致可分为 3类: 声学层面特征、语言模型特征、词图特征。

声学层面的置信度特征主要是基于最大后验概率(maximum a posterior probability, MAP)的。理论上,给定的语音特征序列 X, 词模型 W 对应的后验概率 $p(W|X)$本身就是 W 的置信特征 $C_{MAP}(W|X)$。根据 Bayes公式,有

$$C_{MAP}(W|X) = p(W|X) = \frac{p(W)p(X|W)}{p(X)}.$$

在实际的语音识别系统中,一般都忽略不计 $p(X)$, 因为对给定的 X 来说, $p(X)$是一个常量。但是,在计算代表置信程度的后验概率时,必须考虑 $p(X)$。可以通过一个全词网络来计算 $p(X)$, 即

$$p(X) = \sum_{W} p(W)p(X|W).$$

这种方法的计算量非常大。大多数系统中都采用不同的近似方法来算 $p(X)$, 如用前 N 个候选来近似。

本文中用到的局部置信度确认特征主要是基于最大后验概率的。此外,还从识别结束后得到的词图中提取了特征。d-Ear Attendant系统的识别算法中并没有使用语言模型,所以语言模型特征不在局部确认特征之中。具体说来,待确认词的局部置信度特征如下:

1) 词一级的对数后验概率得分,也称为正规化对数似然得分(normalized log-likelihood, NLL)。词的 NLL得分定义为该词内所有音素的 NLL分的平均值,而音素的 NLL得分是音素中所有帧的帧一级 NLL得分的算术平均值。给定一帧语音信号 X, 帧一级 NLL得分计算公式为

$$C_{NLL}(W|X) = \lg\frac{p(X|W)}{p(X)}.$$

它反映了模型与整个待确认候选匹配的平均情况。

2) 词中最小的音素级 NLL得分,该特征反映待确认序列中置信度水平最低的部分。

3) 词结束时搜索空间中活动的路径条数。

4) 识别结束后得到的词网格中,与当前待确认词处于并列位置的其他候选词的个数。竞争词的数目越多,候选词的可靠性就越低。

5) 识别结束后得到的词网格中,在并列位置包含当前待确认词的路径的条数。候选词在不同路径的并列位置中出现的次数越多,说明它越可靠。

1.2 上下文相关置信度特征

N-gram语言模型刻画的是某个词与其相邻词之间的关系,文[7]中的研究表明: 在口语对话系统中,来自于语言模型层面的置信度特征其性能要优于声学层面的特征。这个实验结论是可以理解的,因为语言模型特征不同于声学特征,它反映的是语言层面的知识,在置信度打分中使用语言模型特征相当于在语音确认时引入了一种新的知识源,其效果理应优于置信度打分中单纯使用声学模型特征。另外,在语音识别中,往往会出现识别结果互相干扰的现象: 如果某一个词的识别发生替代和插入错误,很可能会发生识别结果的前后时间边界与正确的边界不一致的情况; 某个词识别结果的边界定位不准确时,必然会导致识别结果中与其相邻的前后词的边界定位也不准确,从而影响前后相邻词的识别准确度。这也可以从另一个角度解释为什么来自于语言层面的置信度特征的性能要优于声学层面的特征。

考虑到来自语言模型的置信特征在语音确认中的积极作用,希望能够引入一些新的特征,起到与语言模型特征类似的作用,即能够描述当前待确认词的上下文对其置信程度的影响。于是,把与当前词相邻的前后两词的特点和可信程度作为一种置信度特征,用于评价当前待确认词的可靠程度,并把这种特征称之为上下文相关的置信度特征。

待确认词的上下文相关特征具体包括:

1) 前序词的局部置信度特征,

2) 后序词的局部置信度特征,

3) 前序词是否补白模型,

4) 前序词是否静音模型,

5) 后序词是否补白模型,

6) 后序词是否静音模型。

除此以外,还加入了一维特征,用于描述整个句子的可靠程度,即整个句子中补白的数目。一般说来,补白越多,整个句子的识别结果就越不可靠。

2 实验说明

2.1 实验背景

本文所有的试验都是以口语对话系统 d-Ear Attendant为背景的。该系统的任务是将呼入的电话正确地转接到相应人员或者部门。由于该任务的特殊性, d-Ear Attendant只关心句子中的某些关键词,如人名、部门名称,因此本文中所指的识别正确率就是关键词检出的正确率。dEar-Attendant系

统采用语境知识指导下的关键词识别,结合了文法网络识别率高和关键词识别鲁棒性强的优点,有很好的性能表现[4]。

2.2 实验数据

d-Ear Attendant系统的关键词表包括 110个被拨叫人名。实验中用到的所有数据均通过 8 kHz采样率的电话信道采集得到。全部数据分成 4个子集。

1) IG-IV (in-grammar and in-vocabulary)集:完全符合识别文法,被呼叫人名在词表内。

2) IG-OOV (in-grammar and out-of-vocabulary)集:完全符合识别文法,被呼叫人名不在词表内。

3) OOG-IV (out-of-grammar and in-vocabulary)集:不完全符合识别文法,被呼叫人名在词表内。

4) OOG-OOV (out-of-grammar and out-of vocabulary)集:不完全符合识别文法,被呼叫人名也不在词表内。

其中 IG-IV和 IG-OOV子集的句子都是符合识别文法的,统称为 IG(in-grammar)集,而 OG-IV和 OG-OOV子集的句子都是不完全符合识别文法的,统称为 OOG(out-of-grammar)集。

语音确认训练集总共包括 2000句,来自 IG-IV子集和 IG-OOV子集的各 1000句;测试集是 1 800句,包括 IG-IV、IG-OV集各 500句,OG-IV、OG-OOV集各 400句;另有 OG-IV、OG-OOV各 100句作为开发集,训练集、测试集、开发集完全独立。

2.3 评价指标

评价语音确认性能的指标通常有以下 2种:

错误拒绝率,定义为

$$R_{\mathrm{FR}}=\frac{\text{识别正确但被标记错误的词数}}{\text{待确认词中正确识别的词数}},$$

错误接受率,定义为

$$R_{\mathrm{FA}}=\frac{\text{识别错误但被标记正确的词数}}{\text{待确认词中错误识别的词数}}.$$

错误拒绝率和错误接受率两者之间是有关联的,错误拒绝率越高错误接受率越低,反之亦然。ROC(receiver operating characteristic)曲线[5]可以很好地描述两者之间的关系。等错误率(equal error rate, R_{EE})是 ROC曲线与坐标系左下角到右上角的对角线的交点,可以认为是错误拒绝率和错误接受率的最佳折中方案,也是评价语音确认效果的指标。

2.4 确认模型

提取出多种置信度特征后,必须联合多种特征,给出最后的确认结果。语音确认要解决的实际上是一个分类问题:即将待确认的识别结果分成“对”或“错”两类。

实验中,用最简单的 Fisher线性分类器作为确认模型。Fisher线性判决方法的原理是把高维空间的样本投影到一条最优分类直线上,即

$$g(\boldsymbol{x})=\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}.$$

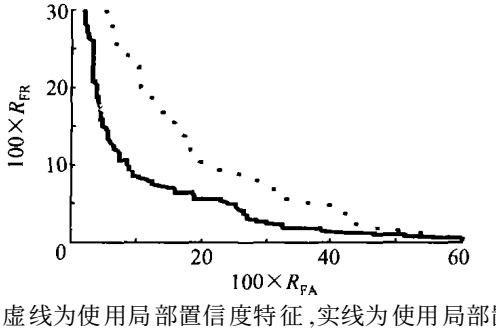
式中:$\boldsymbol{x}=[x_1,x_2,\cdots,x_d]^{\mathrm{T}}$是 d维特征向量,即样本向量,$\boldsymbol{w}=[w_1,w_2,\cdots,w_d]$为权向量。经过投影变换,即可在一维空间上对高维空间中的样本进行分类。

把待确认词的所有置信度特征结合并为一个大的向量,输入 Fisher分类器,根据分类器的输出可判断待确认词正确与否。

3 实验结果

实验 1 上下文相关置信度特征的性能

对符合识别文法的 IG集的测试数据进行实验,结果见图 1所示。从图 1看出,加入上下文相关的置信度特征后,ROC曲线更加接近坐标轴,表示语音确认的性能越好。在 R_{FR}为 2.5%时,R_{FA}从 45%下降到 32%(错误率下降 29%);R_{FR}为 5%时,R_{FA}从 39%下降到 25%(错误率下降 36%);等错误率也从原来的 15.6%下降到 9.2%。

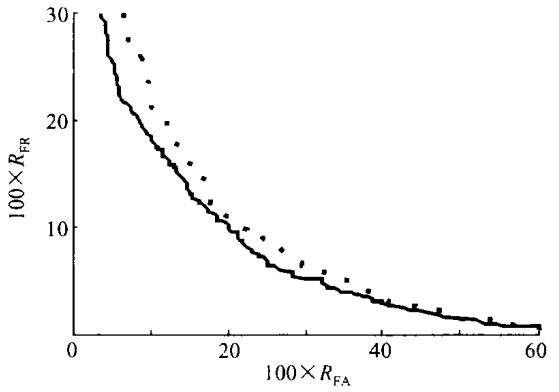


虚线为使用局部置信度特征,实线为使用局部置信度特征和上下文相关置信度特征。

图 1 不同置信度特征在 IG集上的确认性能

d-Ear Attendant在实际使用时必然会遇到很多不符合识别文法的句子,因此有必要进一步测试上下文相关的置信度特征在所有数据上(包括 IG集和 OOG集)的语音确认性能,实验结果见图 2。

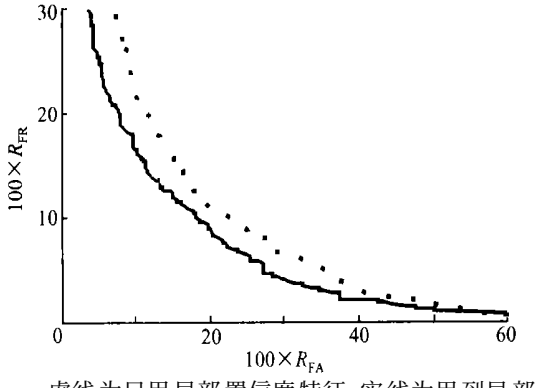
从图 2中看出,在全体测试集上,上下文相关的置信度特征对确认性能的提高不像在 IG集上那么明显,R_{EE}略有下降,从 15.6%下降到 14.2%。分析其中的原因:第一,对于 OOG的数据来说,每句话



虚线为只用局部置信度特征,实线为用到局部置信度特征和上下文相关置信度特征.

图 2 不同置信度特征的确认性能

的待确认关键词的前后相邻词通常都是补白或者静音. 相对于关键词,补白和静音来的局部置信度特征的物理意义并不是那么明确,亦即用前文所说的局部置信度特征对补白和静音的打分与这些特征对关键词的打分不具可比性. 将补白或者静音的局部置信程度作为待确认关键词上下文相关的置信度特征不能起到应有的作用. 第二,测试数据与训练数据的差异也是导致结果不理想的原因. 训练确认模型的数据全部来自于 IG集,这部分数据与 OOG的测试数据非常不一致,影响了测试结果. 针对第二个原因,将开发集的 200句话也加入确认模型的训练中,模型重新训练后的实验结果如图 3所示.



虚线为只用局部置信度特征,实线为用到局部置信度特征和上下文相关置信度特征.

图 3 增加训练数据后不同置信度特征的性能

从图 3中看出,对训练数据稍做调整后,上下文相关的置信度特征对确认性能有明显的改进,以错误拒绝率 2.5% 和 5% 为例,具体结果见表 1.

实验 2 加入语音确认后系统的性能

本实验考察基于置信度打分和基于在线补白模型的两种语音确认方法对系统识别性能的影响. 为了避免过多的口语现象对实验结果的影响,仅对符合识别文法的 IG集测试数据进行测试. 不进行语音确认时,系统的识别正确率为 97%, 加入语音确认后正确率略有下降. 在保证错误接受率相同的情况下,使用在线补白模型进行语音确认最后的识别正确率为 92.6%, 而用置信度打分的方法进行语音确认,识别正确率为 94.3%. 这一结果说明在错误接受率相同的前提下,相比在线补白模型方法,基于置信度打分的语音确认方法保证系统有更高的识别正确率.

表 1 不同置信度特征确认性能比较

R_{FR}	R_{FA}× 100		R_{FA}下降率 × 100
	局部置信度特征	局部特征+ 上下文相关的置信度特征	
2.5	43	37	14
5	35	27	23

4 结 论

本文研究基于置信度打分的语音确认策略,提出上下文相关的置信度特征法,这一方法充分考虑当前待确认词与其前序词和后序词之间的相关性. 实验结果表明: 新方法大大提高了语音确认的性能; 基于置信度打分的语音确认策略在整体性能上优于基于在线补白模型的语音确认方法.

参考文献 (References)

[1] Renals S, Morgan N, Bourlard H M, et al. Connectionist probability estimators in HMM speech recognition [J]. *IEEE Trans on Speech and Audio Processing*, 1994, **2**(1): 161-174.

[2] Rose R, Paul D. A hidden Markov model based keyword recognition system [A]. Proc ICASSP [C]. Albuquerque, USA: IEEE Press, 1990. 129-132.

[3] 刘俊, 朱小燕. 基于动态垃圾评价的语音确认方法 [J]. 计算机学报, 2001, **24**(5): 480-486.
LIU Jun, ZHU Xiaoyan. Utterance verification based on dynamic garbage evaluation approach [J]. *Chinese J Computers*, 2001, **24**(5): 480-486. (in Chinese)

[4] ZHANG Guoliang, SUN Hui, ZHENG Fang, et al. Robust speech recognition directed by extended template matching in dialogue system [A]. Proc WCICA 2004 [C]. Hangzhou: IEEE Press, 2004. 4207-4210.

[5] Hazen T J, Seneff S, Polifroni J. Recognition confidence scoring and its use in speech understanding systems [J]. *Computer Speech and Language*, 2002, **16**: 49-67.

[6] Wessel F, Schluter R, Macherey K, et al. Confidence measures for large vocabulary continuous speech recognition [J]. *IEEE Trans on Speech and Audio Processing*, 2001, **9**(3): 288-298.

[7] San-Segundo R, Pellom B, Hacioglu K, et al. Confidence measures for spoken dialogue systems [A]. Proc ICASSP2001 [C]. Salt Lake City, USA: IEEE Press, 2001. 393-396.