文章编号:1001 - 9081(2005)12 - 2795 - 03

在线废料建模在特定领域语音识别中的应用

幸璐璐,谢莎莎,孙甲松,王作英
(清华大学 电子工程系,北京 100084)
(xll03@mails.tsinghua.edu.cn)

摘 要:严格按照语法规则模型指导声学层识别的特定领域语音识别系统,难以处理未经规则描述的插入语或语气词等语言现象。针对这一问题,将在线废料建模方法应用于该系统,详细讨论了此方法中模型参数 N 的选择策略,分析验证了语料的信噪比 SNR 值与参数 N 之间的相关性,提出了基于此相关性的模型参数优化方法,使得系统的句子识别率和槽识别率相对基线系统分别提高了 18.34% 和 11.47%。

关键词:语音识别;废料;在线废料建模;词图搜索
中图分类号: TP391.42 **文献标识码**:A

Application of on-line filler modeling in specific domain speech recognition system

XING Lu-lu, XIE Sha-sha, SUN Jia-song, WANG Zuo-ying
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: It is difficult for a speech recognition system in specific domain to deal with parentheses or sentence particles which is not depicted in rule since it instructs acoustic layer recognition in strict syntax rules model. To solve this problem, an on-line filler modeling method based on wordgraph decoder was applied to specific domain speech recognition system. The choice of the parameter N in this method was discussed in detail, and the correlation between materials' SNR and the parameter N was also analyzed and validated. Moreover, a model parameter optimization method, which benefits from the correlation, was proposed. Computer simulation validates that the proposed method increases recognition rate by 18.34% for sentence and 11.47% for slot relatively.

Key words: speech recognition; filler; on-line filler modeling; wordgraph decoder

0 引言

语音识别是人与机器进行口语交流的关键技术,在智能家电、智能办公以及机器人技术等方面有着广阔的应用前景。而语言模型是高性能语音识别系统不可或缺的重要组成部分,它包括统计模型和规则模型两种。目前常用的统计模型为 N-gram 模型;而规则模型则是指传统的规则文法。对于特定领域的语音识别任务,如果采用传统的 N-gram 统计语言模型,由于很难获得足够的文本语料,常出现严重的数据稀疏现象,往往不能达到令人满意的效果;与之相对,由于特定领域的语言规则比较简单,语法和句型相对统一,因此一般采用规则模型效果会更好。基于图搜索的特定领域的语音识别系统[1]即直接使用规则模型指导声学层识别。

但是在日常生活中,人们说话时常含有如“请问”,“是吗”这类的插入语或语气词,它们不表示实际意义,不影响整句话的语义理解,通常称之为废料。语法规则难于表述这一现象,所以直接按照语法规则进行识别搜索的系统,对于这种含有废料的情况难以得到正确的识别结果,这将导致整个系统性能的降低。

为了处理这种情况,多采用离线废料建模的方法[2],它预先对废料建立模型。在数据量较小时,离线废料模型能比较精细地刻画废料特性。但是,由于废料非常广泛,使得模型的设计和训练相当困难,不同应用环境下的模型需要重新训练。同时,Hever 等人[3]也提出了在线废料建模的方法,此算法简单,易于实现。扩展的模板匹配方法[4]把在线废料建模方法应用到模板匹配中,增强了系统的稳健性,取得了较好的识别效果。

本文将在线废料建模方法应用于特定领域的语音识别,

收稿日期:2005 - 06 - 24 **基金项目**:国家 863 计划资助项目(2001AA114071)
作者简介:幸璐璐(1981 -),女,四川乐山人,硕士研究生,主要研究方向:特定领域中的语音识别; 谢莎莎(1982 -),女,山西运城人,硕士研究生,主要研究方向:语音识别中的搜索和剪枝算法; 孙甲松(1962 -),男,山东潍坊人,副教授,主要研究方向:语言模型,中文信息处理; 王作英(1935 -),男,江西赣县人,教授,博士生导师,主要研究方向:非特定人连续语音识别.

特征参数。本文提出的基于 MBIC 的决策树聚类方法,对训练数据集规模和声学特征参数维数的变化则具有较好的适应能力,该方法可以应用于声学模型的自适应训练。另外,该方法对声学特征参数的类型变化也有较好的适应能力。

参考文献:

[1] YOUNG SJ, ODELL JJ, WOODLAND PC. Tree-Based State Tying for High Accuracy Acoustic Modelling[A]. Procedings ARPA Workshop Human Language Technology[C]. Berlin, 1994. 286 - 291.
[2] 龚光鲁, 钱敏平. 应用随机过程教程及在算法和智能计算中的随机模型[M]. 北京: 清华大学出版社, 2004.
[3] YOUNG S, EVERMANN G, HAIN T, *et al*. The HTK Book(for HTK Version 3.2)[M]. Cambridge University, 2002.
[4] 吴宗济, 林茂灿, 等. 实验语音学概要[M]. 北京: 高等教育出版社, 1989.
[5] 黄伯荣, 廖序东. 现代汉语(增订三版). 上册[M]. 北京: 高等教育出版社, 2002.

详细讨论了模型参数 N 的选择策略，旨在大致保持对不含废料语句的识别率的基础上，提高含有废料的语句的识别率。同时还分析并验证了语料的信噪比（SNR）和参数 N 之间的相关性，提出了基于此相关性的在线废料模型参数优化方法。实验证明，该方法使系统性能得到了进一步提高，具有良好的实用前景。

1 在线废料建模

1.1 基本原理

在线废料建模方法没有明确地建立废料模型，而是在线计算每一帧语音的声学得分，取其中 N 个最好得分的平均分作为废料的似然得分，如图 1 所示。

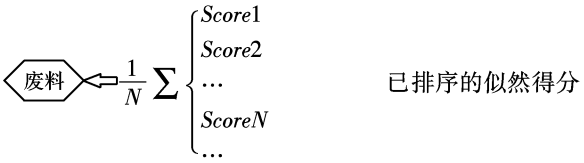

图 1 在线废料建模原理

似然得分 S 定义如下：

$$S = -\log P(o^s)$$

其中，s 是当前匹配的状态，o 是观测矢量。

当语料中存在废料时，虽然该得分不是最优的，但是选择合适的模型参数 N，该得分往往能位于前列，使得此模型具有较好的废料吸收能力。

1.2 模型参数的选择策略

该方法中的模型参数 N 将直接影响识别系统的性能。如果待识别语音含有废料，较小的 N 值可以使废料模型获得较好的吸收能力，从而提高识别率。对于不含废料的情况，在线废料模型的存在一定程度上增大了搜索时的混淆度，为了尽量减小这种负面影响，N 应该越大越好。

然而在实际应用中，我们不能先验地判断待识别语音是否含有废料。为此，需要对上述两种情况综合考虑，选择适当的 N，以提高整个系统的识别性能。

1.3 模型参数的优化

当训练语料和测试语料的 SNR 值相差较大时，通过训练语料得到的 N 在测试语料中的表现不太令人满意。为此，可以假定待识别语音的 SNR 值和参数 N 之间存在一定相关性。

理论上，如果语料的 SNR 值越小，那么它对于声学模型的距离就会越大，相应的似然得分也会增大。因此，低信噪比时，由于废料和正常语音的得分之间的差距被缩小了，语音更倾向于被废料吸收，因此需要适当增大 N 值。

由此，考虑利用带有噪声的训练语料得到的 SNR 和 N 之间的对 N 进行优化，从而降低由于训练语料和测试语料的 SNR 差异带来的影响，进一步提高系统性能。

2 实现方法

2.1 基线系统简介

基线系统采用基于图搜索的语音识别系统，基本框图如图 2 所示，它直接使用语法规则指导声学层识别，将声学层识别搜索范围约束在该语法规则内。

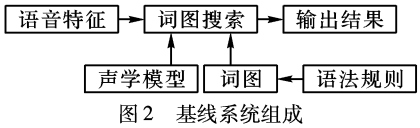

图 2 基线系统组成

语法规则是指通过对某一类语言的句法和语义等进行分析后总结得出的语法规律；词图则是严格按照语法规则所生成的由弧和节点组成的有向图；词图搜索依靠词图来约束路径的扩展，每条搜索路径在到达某个节点后只能沿着与该节点有连接的路径扩展，因此识别结果一定符合词图所表示的语法规则。

2.2 在线废料模型的应用

为了使分析和处理简单，只考虑句首和句尾出现废料的情况。去除废料部分，剩下的语句符合该领域所定义的语法规则。

为此，在词图的起始节点和终止节点各增加一个在线废料弧 OFA（Online Filler Arc），用于吸收废料语音。该弧对应一个虚拟声学状态，如图 3 所示。

图 3 增加废料模型的词图示例

对每个弧的声学状态串进行维特比搜索，如果当前处理弧的语义信息为 OFA，则根据在线废料建模方法，计算其似然得分。

2.3 模型参数 N 的选择及优化

作如下定义，待识别的句子数目：N_i；识别正确的句子数目：CN_i；句子识别正确率：$CR_i = CN_i/N_i$。其中，$i = 1$ 表示含有废料的情况，$i = 2$ 则表示不含废料的情况。

采用不同 N 值对训练语料数据进行实验，得到以上所定义的指标；考虑系统整体性能，定义了下面三种参数 N 的选择策略：

N-1：$N^* = \arg\max_N (CR_1 + CR_2)$

N-2：$N^* = \arg\max_N (CN_1 + CN_2)$

N-3：$N^* = \arg\max_N (CR_2(N) \geqslant \lim_{p\to\infty} CR_2(P) - \delta$，其中 δ 为收敛阈值。

在确定模型参数 N 的选择策略基础上，通过考虑它与语料 SNR 之间的相关性对之进行优化。训练中对于不同 SNR 下的训练语料，根据同一选择策略找到一个最优的 N，对其进行数据拟合，以确定 SNR 同 N 之间的函数关系；测试时先估计待识别语句的 SNR 值，再利用 SNR 同 N 之间的函数关系调整模型参数 N，以去除噪音的影响。

3 实验结果与分析

分别对餐饮、旅游和天气三个领域进行了实验，其中语法规则由 2004 年 863 电话语音测试项目组提供；使用 45 维 MFCC 特征，声学模型是 Mixture 数为 16 的高斯混合模型。基线系统为不含在线废料模型的图搜索语音识别系统。实验所用的训练集为实验室环境下录制的电话语音，每个领域均包含 15 个说话人，600 句话，餐饮、旅游和天气三个领域中含有废料的句数分别为 135，150 和 150。测试集为 2004 年 863 评测中提供的测试语音。语料各领域均为 40 句话，餐饮、旅游和天气三个领域中含有废料的句数分别为 10，14 和 13。

实验结果均是三个领域的平均结果。实验中，通过已有的语法规则把具有一定语义概念意义的表达定义为槽，它可以是具有该概念意义的词或词的组合，也可以是整句话，当一句话中不含有此概念意义时，它的值设为空。只有槽中所有的词识别正确时该槽识别正确；只有一句话所有槽都识别正确时该句话识别正确。

3.1　实验一

本实验中,没有考虑语音 *SNR* 的影响,直接由训练语料得到模型参数 *N* 进行测试。

在不同模型参数 *N* 下,三个领域中含有废料及不含废料情况下的句子识别率如图 4 和图 5 所示。

从图4可以看到,CR_1 随着 *N* 的增大而增大,当 *N* 在35左右时,达到峰值,在越过这一峰值后,废料模型的吸收能力逐渐减弱,CR_1 随之减小并收敛。而图5中,CR_2 随着 *N* 的增大而增加,在 *N* 等于65左右时基本收敛。这与我们前面的分析是一致的。

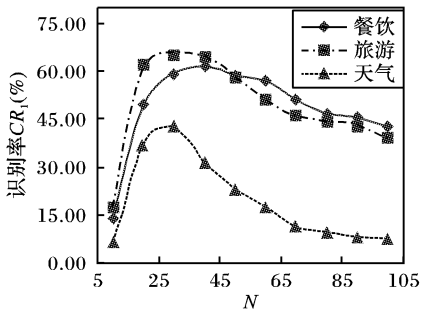

图4　含有废料情况下的句子识别率

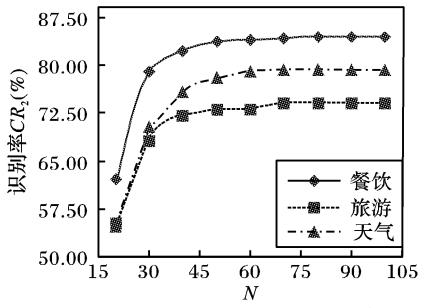

图5　不含废料情况下的句子识别率

表1列出了基线及三种不同选择策略下的测试结果,其中 $CR = (CN_1 + CN_2)/(N_1 + N_2)$ 为总的句子识别正确率,*CS* 则表示非空槽的识别正确率。

表1　测试结果1

测试结果	CR_1(%)	CR_2(%)	*CR*(%)	*CS*(%)
基线	13.51	66.27	50.00	65.83
N-1	48.65	42.17	44.17	62.95
N-2	51.35	49.40	50.00	67.63
N-3	43.24	62.65	56.67	72.66

由测试结果看出,由于训练集和测试集中两种情况所占的比例有所差异,所以前两种方法都没有取得令人满意的性能提高。而对于第三种方法,虽然 CR_2 比基线系统有所降低,但仍然达到了60%以上,同时 CR_1 提高了近30个百分点。系统的整体性能得到了较大提高。

3.2　实验二

本实验中,我们考虑了 *SNR* 与 *N* 之间的相关性。在训练 *N* 时,对训练集的数据叠加白噪声,使得每一句话的 *SNR* 值分别为10、15、20、25、30和35dB。得到每一个 *SNR* 值下最优的 *N* 值,其中 *N* 的选择采用实验一中的第三种选择方法。图6中,实线部分即为餐饮领域中二者的关系。

可以看出,该结果同我们前面的分析基本吻合,随着 *SNR* 的增大,*N* 值逐渐减小,并趋于收敛。为此,我们利用幂函数对以上数据进行拟合,如图6中的虚线所示。

测试中,根据待测语音的 *SNR* 值,利用拟合得到的函数表达式求得相应的 *N* 值,再进行搜索识别。测试结果如表2所示。

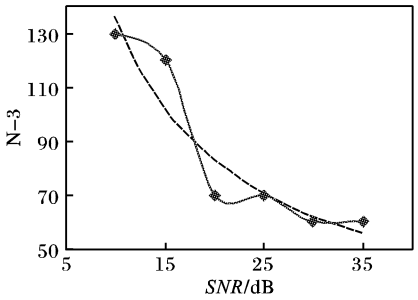

图6　餐饮领域 *SNR* 和 *N* 的关系

表2　测试结果2

测试结果	CR_1(%)	CR_2(%)	*CR*(%)	*CS*(%)
基线	13.51	66.27	50.00	65.83
V-1	43.24	62.65	56.67	72.66
V-2	43.24	66.27	59.17	73.38

V-1 和 V-2 分别代表 *N* 不随信噪比变化和幂函数拟合 *N* 的测试结果。从中可以看出,V-2 与 V-1 的 CR_1 持平;而对于 CR_2,V-2 比 V-1 的相对提高接近6个百分点。

所以,采用在线废料模型的同时,考虑语料的 *SNR* 与参数 *N* 之间的相关性,系统性能得到进一步提高。

同时由于废料模型的引入,在一定程度上增大了搜索空间和混淆度,使得识别速度有所降低,因此,我们对系统的实时率进行了测试,在 P4-3GHz,512M 内存的 PC 机上的测试结果见表3。

表3　实时率测试结果

测试结果	基线	V-2
实时率	0.70	1.01

从以上结果可以看到,改进后的系统也基本达到实时,可以满足实际应用的需要。

4　结语

本文将在线废料建模方法应用于基于图搜索的识别系统中,文章详细讨论了模型参数 N 的选择策略,并分析验证了语料的 *SNR* 值与参数 *N* 之间的相关性,提出了基于此相关性的模型参数优化方法,使得系统的句子识别率和槽识别率相对基线系统分别提高了18.34%和11.47%,同时还保持了系统的实时性,从而证明了该方法的实用性和可行性。

参考文献:

[1]　孟建庭,吴及,王作英.基于图搜索的特定领域语音识别[J].电声技术,2004,9:37-39.

[2]　ROSE RC, PAUL DB. A hidden Markov model based keyword recognition system[A]. Proceedings of ICASSP'90[C]. Albuquerque, NM: 1990.129-132.

[3]　BOURLARD H, D'HOORE B, BOITE JM. Optimizing recognition and rejection performance in wordspotting systems[A]. Proceedings of ICASSP'94[C]. Adelaide, Australia: 1994. 373-376.

[4]　ZHANG GL, SUN H, ZHENG F, *et al*. Robust speech recognition directed by extended template matching in dialogue system[A]. Proceeding of the 5th World Congress on Intelligent Control and Automation[C]. Hangzhou, P. R. China: 2004. 4207-4210.

[5]　CHIU CC, DELLER JR JR. Two-pass Decoding Algorithm for Partitioned Search in Continuous Speech Recognition[A]. Proceedings of APCCAS'94[C]. Taibei, Taiwan: 1994. 524-529.