语音理解中的容错技术的研究

张建平,王作英,赵庆卫,陆大[illegible]
(清华大学电子工程系,北京 100084)

摘 要: 本文研究了大词汇量非特定人汉语连续语音识别和理解系统中的容错技术. 首先,声学识别器产生 N 个最优(*N*-best)音节候选及其相应的声学层的概念,再由 *N* 个最优音节候选构成一个音节网格(syllable lattice). 一个容错语言分析器被用来搜索该音节网格并发现最优的汉字串. 由于考虑了额外的可能候选音节,该最优汉字串的某些字的音节可能不在原来的音节网格中. 这样,声学层的一些错误被纠正,语言分析器的稳健性(robustness)得以提高. 实验表明容错分析器能将字的理解正确率从 91.83 %提高到 94.15 %. 与传统的无容错技术的基于三元文法模型的分析器相比,错误率下降了 28.4 %.

关键词: 容错技术; 语言模型; 复杂度; levenshtein 距离
中图分类号: TN912 **文献标识码**: A **文章编号**: 0372-2112 (2000) 03-0084-03

A Study of Error-Tolerant Techniques for Speech Underst anding

ZHANG Jian-ping, WANG Zuo-ying, ZHAO Qing-wei, LU Da-jin
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: In this paper, error-tolerant techniques are studied for large-vocabulary speaker-independent Chinese continuous speech recognition and understanding systems. When N-best syllable candidates with their corresponding acoustic scores are generated and a syllable lattice is constructed by the acoustic recognizer, an error-tolerant linguistic parser is used to search through the syllable lattice and find the optimal Chinese character sequence in which syllables of some characters may not be included in the lattice because additional possible candidates are considered. In such a way, some acoustic errors are corrected and the robustness of the parser is improved. Our experiments show that the error-tolerant parser can increase the understanding rate from 91.83 %, a 28.4 % error reduction over the conventional trigram model based parser without error-tolerant techniques.

Key words: error-tolerant techniques; language mode; perplexity; levenshtein distance

1 引言

本文将提供语音理解中的容错技术的一些实验结果. 从一个声学处理器得到的 N 个最优候选可能包含替换、插入和删除错误. 尽管我们能够通过增大 *N* 来减少这种错误,但当 *N* 大于 10 时这种方法不是很有效[3],并且会增加语言分析器的负担. 对语音识别系统而言,非常需要一个容错分析器能够通过考虑 *N* 个最优候选外的可能音节以容许一些音节错误. 在例 1 中,分析器的正确输出可以利用语言学知识,考虑额外可能的候选得到. 虽然正确音节 xin 和 qing 并不在 *N* 个最优候内,但 xing 和 qin 在 *N* 个最优候选内. 一个容错分析器将在提出的尺度(metrics)下考虑这些正确音节.

另外,由于不同的说话人可能有不同的口音,在汉语中, in 常被误差读成 ing, sh 误读成 s 等. 在这种情况下,正确的音节可能不在 N 个最优候选内. 还有,对于象听写机这样的系统,要求用户读一段长的文章而不出现任何发音错误是很困难的. 基于以上事实,可以看到容错技术对实验语音识别系统有重要的作用. 尽管本文提出的方法被用于语音理解,它们对词法处理,拼音校正和信息提取中的字符串近似匹配都有用处[1].

正 确 句 子:每天有新鲜的事情发生
间 节 序 列:met tian you xin xian de shi qing fa sheng
第 *i* 个候选: 1 2 1 * 5 1 1 * 2 1
基线(baseline)系统输出:每天有兴现的是亲发生
容错分析器输出:每天有新鲜的事情发生
(* 表示正确音节不在 N-*best* 候选内)

例 1 两种分析器的分析结果

2 容错算法概述

容错算法的关键是如何衡量两字符串的差异. 在模式识别应用,例如手写体识别、语音识别、信息提取和机器翻译中,我们常常要比较文本类型或音素. 问题的核心是如何度量字符串间的差异. 下面介绍三种尺度以比较字符串的差异. 并且设计了一递归算法来实现这些尺度.

收稿日期:1998-10-12;修订日期:1999-07-30
基金项目:国家 863 高科技计划项目(No. 863-306-03-02-1)支持课题

2.1 广义 Levenshtein 距离(Generalized Levenshtein Distance, GLD)

广义 Levenshtein 距离是:为了将一字符串转变为另一字符串所需的与插入、删除和替换操作相对应的编辑距离之和的最小值. 首先,对它进行公式化. 设 A 是被考虑的字母表,A^* 为基于 A 的字符串集. $\lambda \notin A$ 是空符号. 字符串 $S = x_1 x_2 \Lambda x_N$ 的长度为 $|S| = N$,其中,$S \in A^*$ 并且 $s_i \in A$. S_i 代表字符串 S 的前 i 个字符 $(1 \leq i \leq N)$. 插入、删除和替换操作定义如下:

(1) $d_s(a, b)$:与 a 用 b 替换相对应的距离,$a, b \in A$. (一般,$d_s(a, a) = 0$)

(2) $d_i(a)$:与插入 $a \in A$ 相对应的距离.

(3) $d_e(a)$:与删除 $a \in A$ 相对应的距离.

然后,长度为 $|S| = N$ 的字符串 S 与长度为 $|T| = M$ 的字符串 $T = y_1 y_2 \Lambda y_M$ 的 GLD 能递归地公式化为:

$$D(s, i, j) = \min[D(s, i-1, j) + d_e(x_i), D(s, i, j-1) + d_i(y_j), D(s-1, i-1, j-1) + d_s(x_i, y_j)] \quad (1)$$

上式中,$D(s, i, j)$ 是 S_i 和 T_j 间的广义 Levenshtein 距离,其中最优转换包含 s 次替换. 一般满足:$0 \leq s, i \leq N$ 及 $0 \leq j \leq M$.

2.2 受限编辑距离(Constrained Edit Distance, CED)

由于 GLD 不能有效地用于识别有噪声的子序列[4]和骨骼图像[5],人们提出了受限编辑距离. 设为了将字符串 S 编辑为 T 对所需操作的数量和类型的任何编辑约束为 τ. 于是,S 和 T 间的关于约束 τ 的 CED 可写成:

$$\mathrm{CED}_\tau(S, T) = \min_{S \in \tau}[D(s, N, M)] \quad (2)$$

一般,s 满足 $0 \leq s \leq N$. 如果在编辑 S 为 T 过程中,发生了 s 次替换,那么删除和插入的次数分别为 $N - s$ 和 $M - s$.

2.3 标准编辑距离(Normalized Edit Distance, NED)

标准编辑距离可看作 CED 的一种特殊情形. 可表示为:

$$\mathrm{NED}(S, T) = \min_{s}[D(s, N, M)/(N + M - s)] \quad (3)$$

NED 的含义是:为了将 S 编辑为 T 所需的一系列操作的编辑距离之和与操作次数的比率的最小值.

2.4 计算 CED 和 NED 的算法

根据上面的讨论,一个有效的算法被用于计算字符串 S 和 T 的 CED 和 NED.

算法:Dist(S, T);

输入:S, T 和编辑距离集;

输出:$\mathrm{CED}_\tau(S, T)$ 和 $\mathrm{NED}(S, T)$

开始:

/ * 计算 $D(s, i, j)$, $s = 0$ */

$D(0, 0, 0) \leftarrow 0$

for $i \leftarrow 1$ to N do

$D(0, i, 0) \leftarrow D(0, i-1, 0) + d_e(x_i)$

for $j \leftarrow 1$ to M do

$D(0, 0, j) \leftarrow D(0, 0, j-1) + d_i(y_j)$

for $i \leftarrow 1$ to N do

for $j \leftarrow 1$ to M do

$D(0, i, j) \leftarrow \min[D(0, i-1, j) + d_e(x_i), D(0, i, j-1) + d_i(y_j)]$

/ * 计算 $D(s, i, j)$, $1 \leq s \leq N$ */

for $s \leftarrow 1$ to N do

$D(s, 0, 0) \leftarrow \infty$

for $i \leftarrow 1$ to N do

$D(s, i, 0) \leftarrow \infty$

for $j \leftarrow 1$ to M do

$D(s, 0, j) \leftarrow \infty$

for $i \leftarrow 1$ to N do

for $j \leftarrow 1$ to M do

$D(s, i, j) \leftarrow \min[D(s, i-1, j) + d_e(x_i), D(s, i, j-1) + d_i(y_j), D(s-1, i-1, j-1) + d_s(x_i, y_j)]$

/ * 计算 $\mathrm{CED}_\tau(S, T)$ 和 $\mathrm{NED}(S, T)$ */

$\mathrm{CED}_\tau(S, T) = \min_{S \in \tau}[D(s, N, M)]$

$\mathrm{NED}(S, T) = \min_{S}[D(s, N, M)/(N + M - s)]$

全部的计算可在 $0 \leq i \leq N$ 和 $0 \leq j \leq M$ 的二维平面上实现,并且 s 由 0 增加到 N. 该算法的时间复杂度为 $O(N^2 M)$,空间复杂度为 $O(NM)$. 在第四部分,该算法被用于计算两音节间的基于拼音的距离.

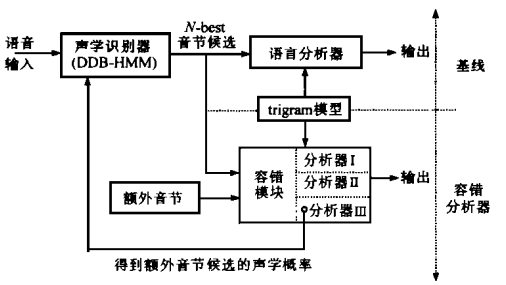

图 1　N 个最优音节候选网格的不同译码方案示例

3 基线系统

基线系统是一个大词汇量非特定人汉语连续语音识别系统. 声学识别器的模型是基于段长分布的隐含马尔可夫模型(Duration-Distribution-Based Hidden Markov Model, DDB-HMM)[2]. 语言模型是一个传统的基于词的三元文法模型,该模型用《人民日报》中的 2100 万词训练得到. 在声学层,408 个无调音节中的每个都可分成声母和韵母格式,类似于英语中的辅音和元音的关系. 在汉语中,有 100 个声母和 43 个韵母. 所有训练数据是在实验室环境以 16kHz 采样率录制. 声学识别器把 N 个 $(N \ll 408)$ 最优音节候选送给用 Viterbi 方法进行译码的语言分析器. 基线系统如图 1 的上半部分所示.

4 语音理解中的容错技术

本节讨论三种误差尺度和相应的三个分析器(参看图 1 的下半部分). 对于给定的一段语音,设 Ω 是所有音节候选的集合,Θ 是 N 个最优音节候选的集合. 定义 X 为 N 个最优候选中的任意一个音节,即 $X \in \Theta$;定义 Y 为 N 个最优候选外的任意一个音节,即 $Y \in \Omega - \Theta$. 那么三种距离尺度定义如下:

4.1 基于半音节的距离

设 $X = I_X F_X$ 表示声母为 I_X,韵母为 F_X 的一音节 X. 设 $Y = I_Y F_Y$ 表示声母为 I_Y,韵母为 F_Y 的一音节 Y. 那么,X 和 Y 间

的语言学的距离可表示为:

$$ld^{<1>}(X,Y) = ld^{<1>}(I_X,I_Y) + ld^{<1>}(F_X,F_Y) \quad (4)$$

其中,$ld^{<1>}(I_X,I_Y)$ 表示 I_X 和 I_Y 间的语言学距离,定义为:

$$ld^{<1>}(I_X,I_Y) = \begin{cases} 0 & I_X = I_Y \\ 1 & I_X \neq I_Y \end{cases} \quad (5)$$

同样,$ld^{<1>}(F_X,F_Y)$ 表示 F_X 和 F_Y 间的语言学距离,定义为:

$$ld^{<1>}(F_X,F_Y) = \begin{cases} 0 & F_X = F_Y \\ 1 & F_X \neq F_Y \end{cases} \quad (6)$$

这样,X 和 Y 间的声学距离可用下式来计算:

$$d^{<1>}(X,Y) = ld^{<1>}(X,Y) \cdot d_0/2 \quad (7)$$

式中 d_0 可从 N 个最优音节候选的声学概率估计得到. d_0 的含义是两音节间的平均声学距离. $d^{<1>}(X,Y)$ 可看成第二部分讨论的 NED 的特殊情形,此时将 X 转变为 Y 时没有插入和删除发生.

4.2 基于拼音的距离

两音节间的基于拼音的距离衡量了两音节音拼音上的差异. 语言学距离 $ld^{<2>}(X,Y)$ 定义为:为了将 X 的拼音转变为 Y 的拼音所需的插入、删除和替换操作数的最小值[1]. 同时,声学距离 $d^{<2>}(X,Y)$ 可计算为:

$$d^{<2>}(X,Y) = ld^{<2>}(X,Y) \cdot d_0/L \quad (8)$$

上式中 L 是拼音中的字母的平均数. 在这种情形下,$d^{<2>}(X,Y)$ 是第二部分讨论的 NED 的改进形式. 同样,第二部分给出的算法可用于计算 $ld^{<2>}(X,Y)$.

4.3 基于音节的距离

两个音节 X 和 Y 音的基于音节的距离定义如下:

$$ld^{<3>}(X,Y) = \begin{cases} 0 & X = Y \\ 1 & X \neq Y \end{cases} \quad (9)$$

声学层的距离用下式计算:

$$d^{<3>}(X,Y) = ld^{<3>}(X,Y) \cdot |d_X - d_Y| \quad (10)$$

式中,d_X 和 d_Y 分别为 X 和 Y 的声学概率. 声学概率可从声学识别器获得. 这种情形下,因为 d_Y 不在 N 个最优候选音选内,声学识别器和语言分析器应该以交互的方式工作.

4.4 容错分析器

对应上面讨论的三种尺度,三种语言分析器:Parser-Ⅰ、Parser-Ⅱ和 Parser-Ⅲ被实现和测试. 图 1 的下半部分画出了一个容错识别和理解系统的关键部件. 为了减少计算量,对语言距离 $ld^{<i>}(X,Y)$,$1 \le i \le 3$ 设置了一阈值 ld_{TH}. 因此,只有当 Y 满足 $ld^{<i>}(X,Y) \le ld_{TH}(1 \le i \le 3)$ 时,Y 才是一个有效的候选. 如果 $Z_{\min}$ 是 N 个最优候选中与 Y 有最小语言学距离的音节,且 $ld^{<i>}(Z_{\min},Y) > ld_{TH}$,那么 Y 就被丢弃. 同样,出于时间上考虑,语言学距离的计算可通过矩阵 B 得到,B 得到,B 为 408 ×408 矩阵,元素为 $B(X,Y) = ld^{<i>}(X,Y)(1 \le i \le 3)$. 那样,两音节间的语言学距离可查找矩阵 B 得到.

5 实验和结果

实验中的录音数据包括三部分. 一部分是男音,包含 519 句. 其它两部分分别是 5 男 5 女录音,每部分包含 240 句. 基线声学识别器产生 5 个最优候选,通过一音节网格传给语言分析器.

5.1 声学识别器结果

表 1 音节识别率(N:1 - 5)

	前 N 候选正确率(%)				
数据	1	2	3	4	5
*m*1. dat	88. 20	93. 30	94. 97	95. 65	96. 10
*m*5. dat	87. 95	93. 48	94. 72	95. 23	95. 61
*f*5. dat	89. 67	94. 39	95. 56	96. 08	96. 42

从上表看出,由 5 个女生发音的测试数据(*f*5. dat)有最高的识别率,m1. dat 次之.

5.2 测试集复杂度(Test-Set Perplexity,PP)

使用测试集复杂度去衡量一个语言模型的性能. 在此情况下,具有较小复杂度的模型一般认为优于具有较高复杂度的模型. 三部分测试数据的复杂度显示如下:

表 2 测试数据的复杂度

数据	*m*1. dat	*m*5. dat	*f*5. dat
PP	115. 01	74. 40	86. 75

注:$PP = \exp\left\{-\frac{1}{Q}\log P(W_1 W_2 \Lambda W_Q)\right\}$,$W_i$ 是测试文本的第 i 个词,Q 为总词数

5.3 不同分析器的理解正确率

理解结果总结于下表中:

表 3 不同分析器的性能比较

	理解正确率(%)			
数据	基线系统	Parser-Ⅰ	Parser-Ⅱ	Parser-Ⅲ
*m*1. dat	91. 84	92. 29	92. 70	94. 33
*m*5. dat	90. 21	92. 50	92. 02	94. 85
*f*5. dat	93. 43	95. 32	94. 97	95. 28
平均	91. 83	93. 37	93. 23	94. 15

从表 3 中看出,容错分析器达到了比基线系统更好的性能. 虽然 Parser-Ⅲ获得了最好的平均理解正确率(94. 15 %),但这种分析器需要 N 个最优候外的音节的声学概率值,比其它分析器需要更多的时间. 总之,容错分析器能有效地纠正声学层的错误并且提高非特定人识别系统的稳健性.

6 结论

本文实现了容错分析器,这种分析器能有效地减少声学层的识别错误. 错误率平均减少 28. 4 %. 本文还提出了三种尺度以衡量两音节间的语言学距离. 相应地,它们的声学层的距离可以被有效地估计和计算. 这种方法也适用于处理声学识别器的插入和删除错误. 在将来的工作中,将研究这些问题并考虑在信息提取和命令理解系统中的潜在应用.

张建平 1992 年获清华大学电子工程系学士学位,1995 年获北京邮电大学电信工程系硕士学位,1999 年毕业于清华大学电子工程系信号与信息处理专业. 研究方向为语音识别与理解、语言建模、容错算法、词义模型、电话语音识别、关键词检测、语音识别中的算法.

(下转第 56 页)

的截止波长,如表 3 所示. 由于高阶模的传输速度与基模不同,高阶模的出现将使传输信号产生色散. 图 7 示出了第一高阶模截止波长与双线导体直径与导体间距之比的关系. 当 $d/D = 0.195$ 时,第一高阶模的截止波长最短,可得到宽带基模传输. 这一结果对屏蔽平行双线的工程设计有参考价值.

表 3 屏蔽平行双线 TE 高阶模的截止波长

	TE_{sa1}	TE_{ss1}	TE_{as1}	TE_{aa1}	TE_{sa2}	TE_{ss2}	TE_{as2}	TE_{aa2}
λ_c/a	1.4427	0.9652	1.8009	1.0036	0.6845	0.5314	0.7118	0.6725

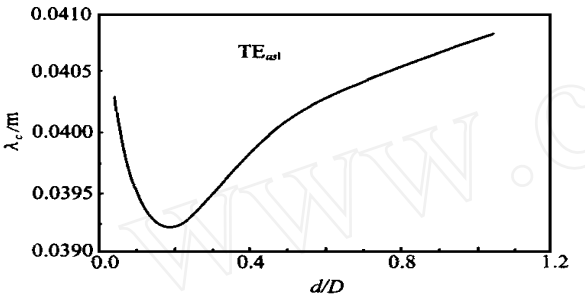

图 7 屏蔽平行双线第一高阶模截止波长

本文提出了保角变换 FDTD 算法. 以 Cassinian 变换为例,计算了椭圆波导、蒺形波导的截止波长与色散曲线,给出了屏蔽平行双线高阶模的截止波长. 其中椭圆波导截止波长计算结果与文献[7]完全一致,证明本文提出的基本算法、焦点的处理方法是正确的. 用其他保角变换,本方法还可求解其他复杂边界波导的传输特性.

参考文献

[1] Cheng Liao, Yusheng Zhao, Weigan Lin. New method for numerical solution Maxwell's equation. Electronics Letters, 1995, (31): 261～262

[2] 廖成,任朗. Thompson-FDTD 与经典 FDTD 方法之比较. 电波科学学报, 1997, (4): 356～360

[3] S. Chakravarthy, D. Anderson. Numerical conformal Mapping. Math. Comp., 1979, (33): 953～969

[4] A. Asi, L. Shafai. Dispersion analysis of anisotropic inhomogenous waveguides using compact 2D-FDTD. Electronics Letters, 1992, (15): 1451～1452

[5] S. Xiao, R. Vahldiech. An efficient 2-D FDTD algorithm using real variables. IEEE. Microwave & Guided Wave Letters, 1993, (3): 127～129

[6] P. Moon, D. E. Spencer. Field Theory Handbook. New York: Springer-Verlag, 1971

[7] 张善杰,沈耀春. 任意偏心率椭圆波导的本征模序. 电子学报, 1994, (3): 86～89

[8] 黄志洵,王晓金. 微波与传输线理论与实用技术. 北京:科学出版社, 1996, 370～376

周晓军 分别于 1983 年、1999 年在电子科技大学获硕士和博士学位,现为电子科技大学光电子技术系副教授. 主要从事电磁场数值计算和光纤传感器研究.

喻志远 教授,1987 年在电子科技大学获博士学位. 主要从事电磁场数值计算和微波电路的研究.

林为干 教授、博士生导师、中科院院士. 1950 年在美国 Berkeley 大学获博士学位,1949 年至 1951 年为 Berkeley 大学讲师. 1984 年和 1993 年分别赴加拿大 Manitoba 大学和日本九州大学作访问教授. 曾任中国电子学会微波分会主席、美国《电磁波和应用》杂志主编.

(上接第 86 页)

参考文献

[1] Oflazer Kemal. Error-tolerant finite-state recognition with applications to morphological analysis and spelling correction. Computational Linguistics, 1996, 22(1): 83～89

[2] Wang Zuoying, Gao Hongge. An inhomogeneous HMM recognition algorithm. Chinese Journal of Electronics, Jan. 1998, 7(1): 73～77

[3] Wang Hsin-Min, Ho Tai-Hsuan, et al. Complete recognition of continuous mandarin speech for Chinese language with very large vocabulary using limited training data. IEEE Trans. Speech & Audio Processing, March 1997, 5(2): 195～200

[4] Commen B J.. Recgnition of noisy subsequences using constrained edit distances. IEEE Trans. on Pattern Analysis and Machine Intelligence. 1987, 9: 676～685

[5] Marzal A, Vidal E. Computation of normalized edit distance and applications. IEEE Trans. on Pattern Analysis and Machine Intelligence, 1993, 15(9): 926～932