计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200500097

基于 PIFA 的语音识别系统评测平台

崔　阳[1]　刘长红[2]
1 中国劳动关系学院应用技术学院　北京 100048
2 江西师范大学计算机信息工程学院　南昌 330022

摘　要　语音识别技术的应用领域众多,而语音识别系统的性能评测对语音识别技术的发展起着重要的推动作用。为了能够更好地对比各类语音系统的性能,在总结现有各种语音识别评测方法的基础上,提出了一种基于性能影响因素分析(PIFA)的语音识别平台体系结构,并据此开发了一个通用的语音识别系统评测平台。该平台以评测库和评测项目为核心概念,包含评测数据生成、数据分析、性能评价指标计算和性能影响因素分析等主要模块,能够面向多种任务和多种语音数据对语音识别系统性能进行快速、准确的自动化评价,尤其适用于大词汇量、连续性的语音识别情景。评测结果可以由平台加以统计分析,揭示各数据属性对识别系统性能的影响,指导语音识别系统的改进和提高。

关键词:语音识别;PIFA;评测库;评测项目;性能影响因素分析
中图法分类号　TP311.1

PIFA-based Evaluation Platform for Speech Recognition System

CUI Yang[1] and LIU Chang-hong[2]
1 College of Applied Technology, China University of Labor Relations, Beijing 100048, China
2 College of Computer Information Engineering, Jiangxi Normal University, Nanchang 330022, China

Abstract　There are many application fields of speech recognition technology, and the performance evaluation of the speech recognition system plays an important role in promoting the development of speech recognition technology. PIFA (Performance Influencing Factor Analysis) based architecture of evaluation platform for speech recognition system is proposed by summarizing various existing speech recognition evaluation methods to compare the performance of various speech systems better, and a platform with PIFA is implemented. The platform involves two key concepts, evaluation database and evaluation project, and includes modules of evaluation data generation, data analysis, performance evaluation index calculation and performance influencing factors analysis. It can deal with multiple recognition tasks and many kinds of data, especially for speech recognition with large vocabulary and continuity. The evaluation results can be statistically analyzed by the platform to reveal the influence of various data attributes on the performance of the recognition system, and help the improvement of the speech recognition system.

Keywords　Speech recognition, PIFA, Evaluation library, Evaluation items, Performance influencing factor analysis

1　引言

语音识别技术作为一种自然的人机交互方式,已经在身份识别、计算机文本输入、电话语音服务器、广播语音检索等领域获得了广泛的应用。互联网技术的进步带来了音频信息的迅速增长,因此对语音识别技术提出了新的需求[1]。近年来,国内外基于隐马尔可夫模型[2]、改进型卷积神经网络算法[3-4]、标签同步解码算法[5]、去噪技术[6]等多种新型算法和技术的语音识别系统层出不穷。这些系统在提高语音识别性能方面有了很大提升,但普遍还存在着评测能力较弱的问题。语音识别评测在语音识别技术中具有重要作用,因此它已成为当前语音识别领域的主要研究方向之一。

语音识别评测的主要目的是了解和比较各语音系统的性能,其不仅包括性能评价指标,还涉及系统的其他特性,如针对某些数据属性(例如噪音、方言口音)的鲁棒性等[7]。语音识别评测技术和评测平台的区别是:前者大多是关于性能评价指标的计算和各系统性能的比较,而后者的主要功能是针对一种识别任务,在统一的界面下方便地完成数据采集、系统运行、性能指标计算等一系列工作。目前比较好的一种语音识别评测方法是性能影响因素分析(Performance Influencing Factor Analysis,PIFA)[8]。这种方法采用统计学中的实验设计、方差分析和极差分析等方法,推断属性对性能的影响在统计意义上是否显著,并可对不同系统的鲁棒性进行比较,从而为研究者改进系统提供参考。PIFA 方法的不足之处在于它是一套独立于其他评测操作的流程,大部分工作需要人工完成,过程较为繁琐。

为了将 PIFA 融入通用的评测平台中,本文提出了一个以“评测库”和“评测项目”为核心的评测平台体系结构,并基

基金项目:中国劳动关系学院科研项目(20XYJS004);国家自然科学基金项目(61662030)
This work was supported by the Research Project of China University of Labor Relations (20XYJS004) and National Natural Science Foundation of China (61662030).
通信作者:崔阳(cuiyang14@163.com)

于这一体系结构设计和实现了一个通用的语音识别自动评测平台。该平台针对连续语音识别、连续语音关键词识别、孤立词识别等多种任务,既可以方便地进行评测库生成、性能评价指标计算等普通评测操作,又可以通过数据分析、自动实验设计、模拟实验数据等功能来进行性能影响因素分析,从而得到对各系统特性更完整、深入的评价。

2　语音识别系统评测中的 PIFA 方法

PIFA 的核心思想是根据选定的因素和水平,采用统计学中的实验设计方法,将评测中的语音数据分为若干个固定水平搭配的处理组,在每组内按照数据的性能指标计算出实验数据。根据这些数据,采用方差分析推断各因素对系统性能的影响在统计学意义上是否显著,并使用极差分析法比较各识别系统对因素的鲁棒性。PIFA 中要考查的数据属性,如语音的信噪比、语速,以及说话人口音等称为性能影响因素(简称因素)。因素的离散值称为水平。对于取连续值的因素,一般将其值域分为几个区间,将每个区间作为一个水平,如信噪比的一个水平为(11 dB,14 dB]。PIFA 方法的总体框架结构如图 1 所示。

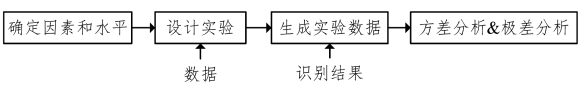

图 1　PIFA 方法的总体框架

Fig. 1　Framework of PIFA

对于一个给定的任务,PIFA 分析过程包括 4 个主要步骤:

(1)根据经验和现有的数据特点性能影响因素以及确定每个因素的水平。常见的口音、SNR(信噪比)等都属于性能影响因素,而 SNR 可以分为 3 个水平,分别是(−∞,11 dB],(11 dB,14 dB],(14 dB,+∞)。

(2)根据选定的因素和水平设计实验,即按照统计学中的实验设计(Design of Experiment,DOE)方法按水平把数据分为若干个处理组,每个处理组是各因素水平的一个组合。例如,{说话人性别为男,无口音,信噪比为 11 至 14 dB}为一个处理组。

(3)在每个处理组内得到实验数据。实验数据可以由各处理组内数据的性能评价指标计算得到,但这些指标不能直接采用。因为方差分析要求各处理组的数据服从正态分布(正态性),且不同处理组间的数据方差相等(方差齐性),而各条数据的性能指标不满足这两个条件。

(4)对得到的各处理组的实验数据,应用方差分析和极差分析获得各因素对系统性能的影响。

3　基于 PIFA 的语音识别系统评测平台

设计通用语音识别评测平台的目的是将 PIFA 方法应用于评测。由于该平台可以面向多种任务和多种语音数据,其功能比只能对一种任务进行性能指标计算的评测平台要强大,应用领域也更为广泛。

3.1　实验数据生成算法

实验生成数据算法主要针对的是 PIFA 分析过程的步骤 3)和步骤 4)。其主要思想是把各处理组中的数据分成若干个子集,在每个子集中计算性能评价指标作为多次重复实验的实验数据,但这种方法只适用于各处理组中数据量基本相等的情况。当各组数据量相差较大时,各处理组内产生的实验数据的方差有较大差别,导致其不适用于方差分析。为此,本文提出了一种类似 bootstrap[9] 的基于模拟的自动实验数据生成算法。通过统计学中的 Shapiro-Wilk 检验和 Levene's 检验进行验证,这种方法在各处理组中数据量相差较大的情况下产生的数据满足正态性和方差齐性。生成实验数据算法的伪代码如图 2 所示。

```
1. 选定N, N小于或等于每个处理组中的数据条数;
2. 选定样本数B;
3. For 每个处理组
        For i=1 to B
                采用有放回抽样方法, 从当前处理组中抽出N条数据;
                对这N条数据计算其总体性能评价指标, 记为Pi
        End for;
End for.
```

图 2　实验数据生成算法的伪代码

Fig. 2　Pseudocode of testing data generation algorithm

3.2　评测库和评测项目

与只计算性能指标的评测平台相比,基于 PIFA 的平台除需要语音数据外,还需要每条数据对应的数据属性作为实验设计、方差分析和极差分析的依据。这就需要把语音数据和对应的数据属性作为一个整体来创建、导入和处理,这个整体称为评测库。由于平台针对多种任务,因此当进行一次评测时,除指定数据和属性外,还需指定任务、指标,并给出系统的识别结果。这里把一个选定的任务、一个或几个性能评价指标、一个指定的评测库、若干个作为输入的识别结果的集合称作一个评测项目。在基于 PIFA 的平台中,评测库和评测项目的结构和内容如表 1 和表 2 所列。

表 1　评测库的结构和内容

Table 1　Structure and content of evaluation library

内容	取值或说明
语音数据	PCM 编码的 WAV 文件,支持多种采样率和量化比特数
采集通道	PC、电话、嵌入式设备、广播
数据特点	朗读、单人自然语音、双人自然对话、孤立词
ID	说话人姓名或编号
性别	男、女
口音	有口音、无口音
字节数	文件字节数
时长	语音的时长
信噪比	基于能量的算法自动估计语音信噪比
语速	自动根据文本和时长计算语速
语音对应文本	标注文本
噪音种类	背景噪音的种类

其中,ID、性别、口音属于说话人信息,而字节数、时长、信噪比、语速是可自动计算的信息。

表 2　评测项目的结构和内容

Table 2　Structure and content of evaluation items

内容	说明
任务	连续语音识别、连续语音关键词检测、孤立词识别
评测库	由用户指定已创建的库
识别结果	各参评系统在评测库上运行后的识别结果,由用户选择文件导入

围绕评测库和评测项目这两个核心概念,就可以进行各种操作,这些操作构成了评测平台的主要内容。因此,评测平台的主要体系结构就是由评测库、评测项目及这两者之上的操作构成,如图 3 所示。

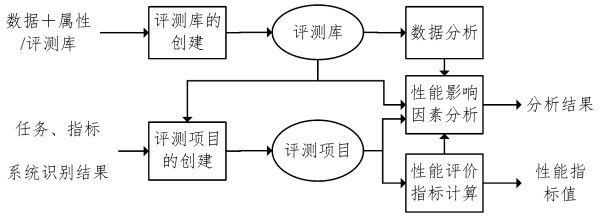

图 3 基于 PIFA 的语音识别评测平台体系结构

Fig. 3 Architecture of speech recognition evaluation platform based on PIFA

3.3 系统功能和架构设计

根据图 3,可以把平台划分为 4 个主要功能模块:评测数据生成模块、数据分析模块、性能评价指标计算模块和性能影响因素分析模块。

(1)评测数据生成模块。其主要功能是采集或导入评测数据以及对应的数据属性。数据属性的导入分为两种:1)语音文本、说话人信息、发音方式、噪音种类等标注信息从用户指定的标注文件中导入;2)文件字节数、语音时长、信噪比、语速等只依赖于数据本身的信息在导入时自动计算。

(2)数据分析模块。其主要功能是对数据属性的分布进行统计分析,作为 PIFA 选择因素和水平的依据。其采用直方图和假设检验估计各属性在数据中的概率分布,采用相关分析和回归分析判断各属性之间是否存在相互影响的关系。

(3)性能评价指标计算模块。其主要功能是将识别结果与参考答案相比较,为评测库中的全体数据计算性能评价指标值。与同类评测系统不同的是,本平台不只计算所有数据的总体指标值,还对每一条数据单独计算指标值,从而有利于 PIFA 计算各处理组内的性能指标。

(4)性能影响因素分析模块。这是基于 PIFA 的评测平台的核心模块,主要功能是按照用户给指定的因素和水平,自动进行实验设计,根据每条数据的性能评价指标值,由图 2 所示的算法自动产生实验数据,由文献[8]中的统计方法进行方差分析和极差分析,以确定各个属性(因素)对系统识别性能的影响。

3.4 系统实现

如前所述,系统中全部操作都围绕“库”和“评测”这两个概念进行。“库”是一批语音数据及其全部属性信息的集合,扩展名为 lrc。库记录文件存储了库中各语音文件及属性信息文件的路径及文件名。“评测”是一个选定的任务、一个指定的库、若干作为输入的识别结果及系统给出的评判结果和评判分析结果的集合。在系统中,一个评测由评测扩展名 erc 唯一确定。其中评测记录文件中存储了评测的任务,以及所用库的库记录文件、识别结果文件、评判分析结果文件的路径及文件名。图 4 为平台的主界面。

图 4 评测平台系统的主界面

Fig. 4 Interface of evaluation platform system

库的属性包括以下几类:1)采集通道,取值为 PC,PHONE,PDA,TV,BC 之一;2)数据特点,取值为连续朗读、双人对话、广播播音、孤立词之一;3)采样率;4)量化比特数;5)是否有头,取值为是、否之一。

评测包括 LVCSR(大词汇量连续语音识别)和 KWS(关键词检测评测)两种,如图 5 所示。其中 LVCSR 以当前主流的拼音音素为粒度构建音素的建模单元集,可以使评测结果更为充分和准确[10]。图 6 为某次评测结果的示意图。

图 5 平台的评测功能分类

Fig. 5 Classification of platform evaluation functions

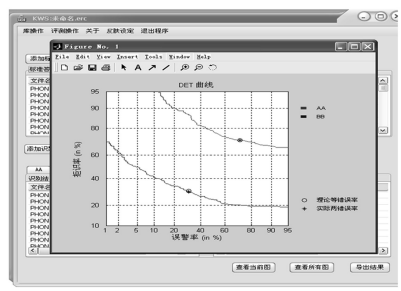

图 6 评测结果显示

Fig. 6 Evaluation results

4 评测平台应用举例

以大词汇量连续语音识别任务为例介绍平台的应用过程。该任务要求识别系统识别 200 个带噪音的汉语普通话句子,其中部分句子发音略带方言口音。该项评测采用的性能评价指标为字错误率。首先采集数据并导入表 1 所示的数据属性,然后采用数据分析模块画出部分属性的值分布的直方图,图 7 即为信噪比和语速对应的直方图。

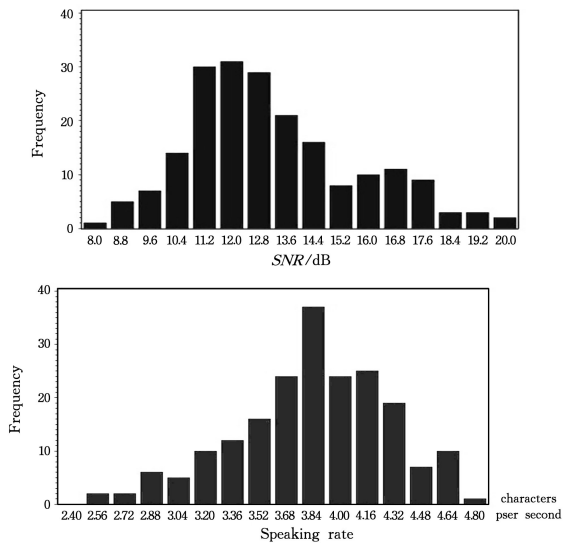

图 7 信噪比和语速的取值分布直方图

Fig. 7 Value distribution histogram of *SNR* and speech speed

由数据分析反映的特点,选定要考查的因素和水平,如表

3 所列。然后启动性能影响分析模块,完成自动设计、实验设计和平台应用正交设计,得到表 4 所列的各处理组。按照该设计,平台可进行实验数据生成、方差分析和极差分析,得到最终的 PIFA 结果。3 个系统的结果如表 5 所列。同时平台也输出各系统对全体数据的性能评价指标。

表 3　选定的因素和水平

Table 3　Selected factors and levels

水平	口音	信噪比/dB	语速/(字/s)
1	无	(−∞,11]	(0,3.40]
2	有	(11,14]	(3.40,4.35]
3	—	(14,+∞)	(4.35,+∞)

表 4　采用正交表 L9($2^1\times3^3$)的实验设计

Table 4　Experimental design with orthogonal table L9($2^1\times3^3$)

处理组	口音水平	信噪比水平	语速水平
1	1	1	1
2	1	2	2
3	1	3	3
4	1	1	2
5	1	2	3
6	1	3	1
7	2	1	3
8	2	2	1
9	2	3	2

表 5　863 语音识别评测的 PIFA 结果

Table 5　PIFA results of 863 speech recognition evaluation

系统	因素	显著性	极差比例
系统 1	口音	否	0.085
	信噪比	是	0.458
	语速	是	0.254
系统 2	口音	是	0.218
	信噪比	是	0.552
	语速	是	0.288
系统 3	口音	否	0.113
	信噪比	是	0.478
	语速	是	0.385

由表 5 中的方差分析结果来看,口音对系统 1 和系统 3 影响不显著,而对系统 2 影响显著;信噪比和语速对 3 个系统影响均显著。由于极差比例越大表明因素对性能的影响越强,可以看出,对于 3 个系统来说信噪比仍然是对性能影响最大的因素,且影响程度相近。对于系统 1 和系统 2,语速的影响几乎是信噪比的一半。而对于系统 3,语速的影响要大于其他两个系统。研究者可以据此了解各个系统的特性,互相交流、学习,进而改进系统性能。

结束语　基于 PIFA 的通用语音识别评测平台面向多种识别任务和语音数据,能够方便地进行评测数据生成、性能评价指标计算和影响因素分析,极大地减少了人工评测的工作量,实现了快速、准确的评测。更重要的是,融合了 PIFA 的平台能够对评测结果(性能指标)进行统计分析,得到不同数据属性对系统性能的影响,使得参评系统能够有效认识自身算法的优缺点,可帮助研究人员有针对性地对语音识别技术加以提高和改进。今后工作主要集中在进一步优化平台性能方面。

参 考 文 献

[1] LIU J,ZHANG W Q. Research Progress on Key Technologies of Low Resource Speech Recognition[J]. Journal of Data Acquisition and Processing,2017,32(2):205-220.

[2] ZHAO J H,GAO H B,LIU Y C,et al. Speech Recognition Algorithm Based on Neural Network and Hidden Markov Model [J]. The Journal of China Universities of Posts and Telecommunications,August,2018,25(4):28-37.

[3] YANG Y,WANG Y D. Speech Recognition Based on Improved Convolutional Neural Netword Algorithm[J]. Journal of Applied Acoustics,2018,37(6):940-946.

[4] CHEN Z H,ZHENG W L,YOU Y B,et al. Label Synchronous Decoding for Speech Recognition[J]. Chinese Journal of Computers,2019,42:1-15.

[5] WEI G W,FENG Z Y. Design of DSP Speech Recognition System Based on Denoising Technology[J]. Transducer and Microsystem Technologies,2017,36(1):108-118.

[6] JIN C,GONG C,LI H. Speaker Adaptation Research of Neural Network Acoustic Model in Speech Recognition[J]. Computer Applications and Software,2018,35(2):200-205.

[7] HU D,ZENG Q N,LONG C,et al. Front-end Robust Study for Continuous Speech Recognition[J]. Video Engineering,2015,39(24):43-46.

[8] WANG X D,XIE F,LIN S X,et al. DOE and ANOVA based Performance Influencing Factor Analysis for Evaluation of Speech Recognition Systems[C]// International Conference on Industrial. Singapore:ISCSLP,2013:431-442.

[9] WANG B C,WEI Y Y,DAI N. Pivoting and Approximate Pivoting of Bootstrap Statistics[J]. Statistics and Decision,2016,16:17-20.

[10] BAO Y B,HU Y,LIU C,et al. Phomeme Modeling Units Desing for Mandarin LVCSR Systems[J]. Journal of Tsinghua University(Sci & Tech),2011,51(9):1288-1292.

CUI Yang, born in 1979, Ph.D, lecturer. His main research interests include knowledge engineering and knowledge discovery.