文章编号：1005-8451（2015）06-0022-04

铁路语音识别引导购票应用的设计与实现

翁涅元，单杏花，蒋秋华，周永健
（中国铁道科学研究院，北京　100081）

摘　要：本文对铁路语音识别引导购票应用进行了需求分析，并分析了开源的Sphinx语音识别平台的特性，在此平台上完成了铁路语音识别引导购票应用的整体设计与实现。

关键词：语音识别；JAVA；Android开发；客票系统
中图分类号：U293.22：TP39　**文献标识码：**A

Speech recognition guided railway ticketing

WENG Shengyuan, SHAN Xinghua, JIANG Qiuhua, ZHOU Yongjian
(China Academy of Railway Sciences, Beijing　100081, China)

Abstract: The paper analyzed the demand for speech recognition guided railway ticketing and the characteristics of Sphinx speech recognition open source platform, designed and developed a demo based on the platform.

Key words: speech recognition; JAVA; Android development; Railway Ticketing and Reservation System

随着铁路客票系统的发展以及互联网联网售票功能的实现，互联网购票、电话购票、手机客户端购票等功能的出现为旅客购票提供了极大的便利，但是购票过程仍需要旅客进行多次的点击操作与查询。而语音引导购票让旅客通过语言表达购票需求，使购票过程变得更为高效和便捷。

1　背景与算法综述

1.1　背景

语音识别技术始于20世纪50年代的语音学与声学基础理论的研究，动态规划方法的提出以及线性预测编码的提出推动了语音识别技术的发展。矢量量化与隐马尔科夫链模型理论的提出，推动了语音识别技术的长足进步。目前，语音识别技术正逐步成为人机交互的关键技术。国内外的语音识别应用平台在开源与商业化两大阵营均有较为成熟的产品。商业化平台中以微软Speech API，IBM Via Voice，讯飞语音识别平台最具代表性。开源软件阵营中，Sphinx、HTK以及Julius等项目也正在高速发展[1]。

1.2　算法

语音识别系统的模型通常由声学模型和语言模型两部分组成，分别对应语音到音节概率的计算和音节到字概率的计算。连续语音识别的大致流程如图1所示。

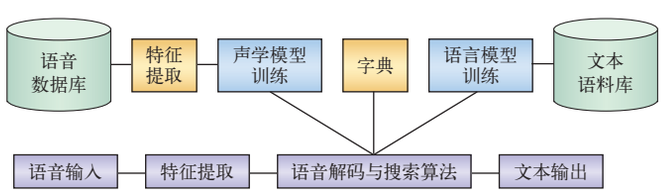

图1　连续语音识别流程

首先对输入的原始语音信号进行端点检测、语音分帧以及预加重等处理[2]。使用Mel倒谱系数去除语音信号中的冗余信息，采用隐马尔可夫模型来模拟人的语言过程，采用N-Gram模型通过词汇出现先后顺序的概率计算概率最大的单词序列，根据已经训练好的声学模型、语言模型或字典建立一个由语音音素组成的有向网络并寻找最佳的路径，即确定了识别的文字信息[3]。

2　需求分析

语音引导购票平台应能迅速高效地识别用户的语音命令，提取信息并快速查询并反馈查询结果。为适应离线与在线识别两种应用场景，语音识别功能应分别在PC与移动端实现。

PC端作为服务端应实现的功能有：接受移动端上传的语音，将识别的文字结果推送给移动端；支

收稿日期：2014-11-13
作者简介：翁涅元，在读硕士研究生；单杏花，研究员。

持更高识别精度的算法；模型文件可以随时更新；有良好的多线程识别的支持；具备语音样本采集与模型训练等功能。

移动端应该同时支持离线与在线两种语音识别模式；保证识别精度的同时，使用的模型体积应该较小，代码精简，启动快速；同时应以组件的形式存在，方便与当前 12306 手机购票软件整合。

在语音转换为文本之后，应有对应的语义分析程序将乘车日期、发站、到站需求抽取出来，为后期购票过程提供信息。语义分析程序应支持在语音识别错误的情况下，从文本中提取有效部分的信息，增加识别应用的容错能力。

3 整体设计与实现

为确定语音识别的词汇范围，通过采集 2012 年以来所有旅客的购票需求统计站名出现的频率。通过统计，旅客购票记录中出现的发站站名共 2 598 个，到站站名共 3 599 个。其中查询频率出现高的站名分布集中，若以累计频率达到 90% 为线，发站、到站合计仅需 522 个站名即可。在系统开发初期可以重点训练这些热门站点的语音样本，保证热门站点的查询准确率，其余站点可以陆续完善补充。

3.1 平台选取

Sphinx 语音识别平台的分支 Sphinx4 和 Pocket Sphinx 各具特色。Pocket Sphinx 具有代码精简、识别速度快、准确率高等优点，是移动端开发的良好选择。Sphinx4 可以提供更高的识别准确度，完全基于 JAVA 平台，有良好的跨平台优势，适合作为服务器端的开发工具。

3.2 识别模式的选取

Sphinx4 和 Pocket Sphinx 均支持语言模型和字典两种识别方式。语言模型在声学模型搜索的基础上综合词汇出现顺序计算所有可能的文本概率。字典模式仅利用声学模型搜索可能的词语，结合字典记录的文本结构在不同的词汇范围中搜寻最匹配的词语组成文本。语言模型考虑词语先后出现的概率，所以训练样本的采集需要考虑所有发到站的组合，样本数量巨大，采集工作繁杂。语言模型识别模式计算复杂，识别耗时长于字典模式。因此针对语音引导购票功能的实现，使用字典模式识别较为合理。

3.3 语音训练样本设计

声学模型的训练是最关键的环节。根据 Sphinx 官方训练规范文档的要求，对于有限样本的多人语音识别需求，需要准备约 200 人，50 h 的训练样本进行训练 [4]。为了便于保证各命令要素（发站、到站、时间）必要的训练强度，将训练文本设计为“今天 从 北京 到 福州”形式。

在原型的开发阶段选取了 9 个站名，3 个时间要素进行识别训练。站名包括：北京、福州、上海虹桥、苏州、广州、太原、厦门、杭州；时间要素包括：今天、明天、后天。共采集了 9 人共 750 个语音样本片段，并尝试进行不同的组合训练并对系统的识别效果进行分析以了解平台特性。

3.4 平台特性分析

根据样本数量、获取时间、设备的不同对样本进行分类和组合，生成 12 个不同的样本集合进行训练。并使用生成的训练结果进行识别测试，以此得到识别准确度的交叉统计表。样本集合信息如表 1 所示。识别准确率统计数据的交叉表如表 2 所示。

表1 样本详情列表

模型列表	样本量	描述
1	158	（男）男声1号提取部分样本
2	200	（男）男声1号
3	89	（女）女声1~3号
4	61	（男）男声2号
5	60	（男）男声3号
6	136	（女）女声4~6号
7	260	（男）男声1、3号
8	261	（男）男声1、2号
9	350	（男女）男声1、2号，女声1号
10	486	（男女）男声1、3号，女声1、4~6号
11	321	（男）男声1~3号
12	546	（男女）男声1~3号，女声1~6号

以上数据展示了不同的样本集合与模型识别表现的关系，为语音样本录制的语料范围与数量提供了参考。通过对 Sphinx 语音识别系统的小样本训练与测试，针对单一词汇的训练仅需 13 次左右的样本即可达到很高的识别率，男声女声不需要单独训练，多人语音样本混合训练在提高整体识别率的同时并不会显著降低个人的语音识别准确率。

3.5 应用设计与实现

表2　识别准确率交叉分析表

模型 \ 样本	1	2	3	4	5	6	7	8	9	10	11	12
1	0.886	0.835	0	0	0.016 6	0	0.646 1	0.577 8	0.477 1	0.343 6	0.481 3	0.307 6
2	0.860 7	0.885	0	0	0.033 3	0	0.688 4	0.612 4	0.505 7	0.364 1	0.512 8	0.327 8
3	0	0	0.508 1	0	0	0	0	0	0.0885	0.063 7	0	0.056 7
4	0	0	0	0.685 3	0.1	0	0.023	0.211	0.174 2	0.125 5	0.191 9	0.122 7
5	0.101 2	0.105	0.032 7	0.044 9	0.55	0	0.207 6	0.086 5	0.077 1	0.055 5	0.166 1	0.109 8
6	0	0	0.393 4	0.022 4	0.066 6	0.911 7	0.015 3	0.006 9	0.074 2	0.308 6	0.017 1	0.282
7	0.962	0.955	0	0.101 1	0.95	0	0.953 8	0.692	0.571 4	0.411 5	0.736 3	0.470 6
8	0.835 4	0.875	0	0.977 5	0.066 6	0.022	0.688 4	0.906 5	0.748 5	0.545 2	0.762 1	0.492 6
9	0.955 6	0.94	0.950 8	1	0.166 6	0.080 8	0.761 5	0.958 4	0.957 1	0.711 9	0.822 3	0.652
10	0.955 6	0.94	0.983 6	1	0.45	0.970 5	0.826 9	0.958 4	0.962 8	0.965	0.871	0.908 4
11	0.968 3	0.955	0.016 3	0.988 7	0.966 6	0.051 4	0.957 6	0.965 3	0.8	0.590 5	0.965 6	0.631 8
12	0.962	0.955	0.967 2	1	0.933 3	0.977 9	0.95	0.968 8	0.968 5	0.971 1	0.962 7	0.967

系统由 PC 端、移动端语音识别软件以及语音训练样本采集、训练工具组成，做到了：样本采集，模型训练，应用实现的整套方案原型。PC 端包含语音训练样本采集工具、训练工具以及基于 Sphinx4 的语音识别工具。移动端使用 Pocket Sphinx 作为语音识别核心，开发了原生 Android 应用以及 IBM WorkLight 框架下（12306 手机购票软件亦基于此框架）的语音识别两套应用。

3.5.1　样本采集与训练工具的实现

专门的语音样本采集工具使语音样本的采集更加方便，同时规范了语音音频参数以及文本的编码规范。语音样本采集工具界面如图 2 所示。

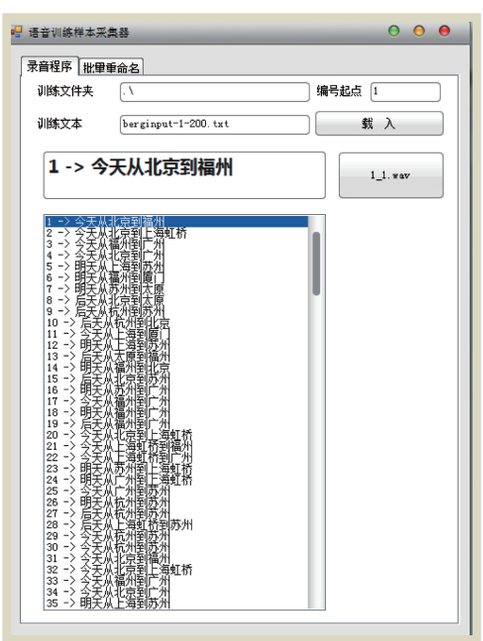

图2　语音样本采集工具界面

训练过程主要使用 Sphinx Train 套件完成。由于训练套件有各自独立的训练模块，使用批处理的方式完成模块的调用和过程文件的整理工作。开发人员只需执行编写好的脚本即可。模型训练工具界面如图 3 所示。

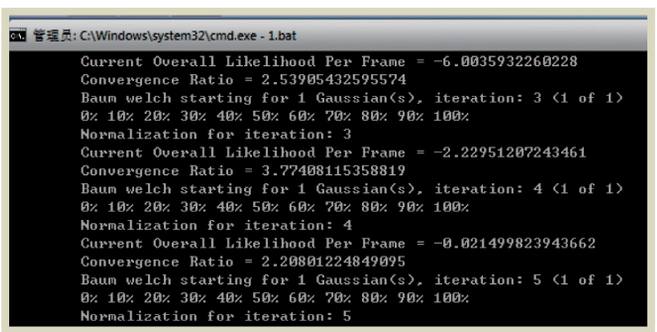

图3　模型训练工具界面

训练产生的模型文件体积仅 700 KB 左右，所占空间体积很小，适合作为移动端的语音识别模型存储使用。

3.5.2　PC端语音识别组件的实现

PC 端的语音识别模块利用 Sphinx4 作为识别核心，编写了多个不同功能的识别模块以满足不同的功能需求，包括：实时连续语音识别、WAV 文件识别、批量音频文件识别、识别结果详细分析等。将配置信息以 XML 形式保存与功能代码独立，并统一了输入输出接口以便后续开发与功能实现，程序结构如图 4 所示。

通过对不同模块的调用，可以满足不同的语音识别需求。PC 端语音识别界面如图 5 所示。

3.5.3　移动端的设计与实现

使用 Pocket Sphinx 作为语音识别引擎的核心，编写并实现了语音识别组件并设计了程序调用与数据传输接口，程序结构如图 6 所示。

上层应用与识别核心相互独立，通过规范调用

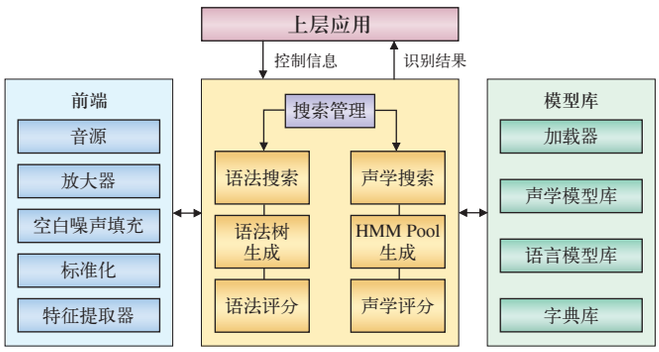

图4　PC端语音识别程序结构

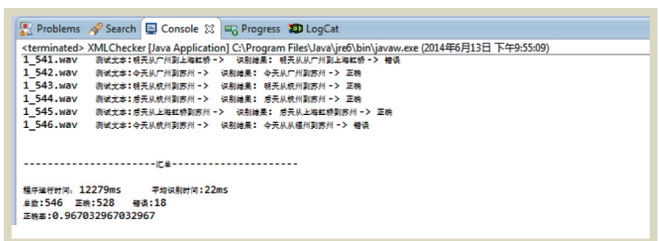

图5　PC端语音识别界面

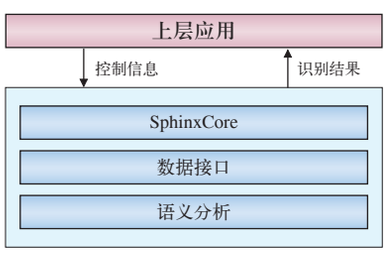

图6　移动端语音识别程序结构

函数与数据传输接口，即可实现方便的代码移植。目前在原生 APP 与 IBM WorkLight 框架下均调用成功，应用运行的效果如图 7 所示。

图7　移动端语音识别程序界面

4　结束语

本文列举了目前购票渠道流程的不足，提出了语音引导购票的思路，设计并实现了语音引导购票的系统原型，并对少量语音样本的不同组成成分与对应的识别结果进行了比较分析，为后续语音样本的采集提供了参考，但目前的成果还不能满足实际使用的要求。

目前，由于采集的样本数量较少，Sphinx 语音识别平台的表现不一定能反应真实应用情况。对 Sphinx 在大词汇量以及实际应用环境下的表现有待进一步检验，Sphinx 中识别器的所有可调参数的作用与调优策略也有待进一步研究。同时对于背景噪音的处理以及针对方言的识别功能也是亟待研究的问题。

参考文献：

[1] 王　韵．基于 Sphinx 的汉语连续语音识别 [D]. 太原：太原理工大学，2010.

[2] Stern R M, Acero A. Acoustical pre-processing for robust speech recognition[C]. Association for Computational Linguistics, 1989: 311-318.

[3] Ravishankar M K. Efficient Algorithms for Speech Recognition[R]. Carnegie-Mellon Univ Pittsburgh pa Dept of Computer Science, 1996.

[4] Acero A, Stern R M. Robust speech recognition by normalization of the acoustic space[C]. 1991. ICASSP-91, 1991 International Conference on. IEEE, 1991: 893-896.

责任编辑　方　圆