衡阳方言孤立词识别研究①

李荣华, 赵征鹏

(云南大学 信息学院, 昆明 650500)

摘 要: 目前, 汉语识别已经取得了一定的研究成果. 但由于中国的地域性差异, 十里不同音, 使得汉语识别系统在进行方言识别时识别率低、性能差. 针对语音识别系统在对方言进行识别时的缺陷, 构建了基于 HTK 的衡阳方言孤立词识别系统. 该系统使用 HTK3.4.1 工具箱, 以音素为基本识别单元, 提取 39 维梅尔频率倒谱系数(MFCC)语音特征参数, 构建隐马尔可夫模型(HMM), 采用 Viterbi 算法进行模型训练和匹配, 实现了衡阳方言孤立词语音识别. 通过对比实验, 比较了在不同因素模型下和不同高斯混合数下系统的性能. 实验结果表明, 将 39 维 MFCC 和 5 个高斯混合数与 HMM 模型结合实验时, 系统的性能得到很大的改善.

关键词: HTK; 隐马尔可夫模型; 衡阳方言; 梅尔频率倒谱系数; Viterbi 算法

Isolated Word Recognition of Hengyang Dialect

LI Rong-Hua, ZHAO Zheng-Peng

(College of Information, Yunnan University, Kunming 650500, China)

Abstract: At present, Chinese speech recognition has made some achievements. However, due to regional differences in China, different place has different dialect, the Chinese recognition system has low recognition rate and poor performance in the dialect recognition. In order to solve the shortcomings of speech recognition system in dialect recognition, an isolated word recognition system of Hengyang dialect based on HTK is proposed. This method constructs the Hidden Markov Models (HMM), using phoneme as the basic recognition unit and using the HTK3.4.1 toolbox to extract the speech feature parameters of 39-dimensional Mel frequency cepstral coefficients (MFCC). Viterbi algorithm is used to train and match the model to achieve the isolated word speech recognition system of Hengyang dialect. The system's performances are compared under the different phoneme models and different Gaussian mixture numbers. The experimental results show that the system performance can be greatly improved by combining the 39-dimensional MFCC with 5 Gauss mixed numbers and HMM model.

Key words: Hidden Markov Model Toolkit(HTK); Hidden Markov Model(HMM); Hengyang dialect; Mel Frequency Cepstral Coefficients (MFCC); Viterbi algorithm

语音是人与人之间方便快捷、准确高效的交流方式, 也是人类最重要的沟通手段. 随着社会的进步和科学技术的不断发展, 语音识别技术逐渐出现在了人们生活当中, 如智能手机、机器人餐厅等. 目前, 较为常用的语音识别模型主要有隐马尔科夫模型(HMM)和人工神经网络(ANN), 而针对 HMM 模型的语音识别工具箱有两个, 即美国约翰·霍普金斯大学开发的 Kaldi 和英国剑桥大学开发的 HTK(Hidden Markov Model Toolkit)[1]. 这两个工具包都提供了一套完整的语音处理功能, 如预处理、训练及识别等. Kaldi 还支持深度神经网络(DNN); 而 HTK 主要设计用于构建基于隐马尔科夫模型(Hidden Markov Models, HMM)的语音处理工具[2]. 其广泛应用于语音识别领域, 它由一系列的功能模块库组成, 实现了对语音进行预处理、模型训练和识别, 并对系统的识别性能进行分析.

由于我国人口众多、地域广阔、各个地区都有自

① 收稿时间:2016-08-25;收到修改稿时间:2016-09-23 [doi: 10.15888/j.cnki.csa.005741]

己独特的方言, 这将给语音识别工作的研究增加了挑战, 虽然有些语种相同, 但在发音上却存在较大的差异, 使得语言识别类型呈直线增长趋势. 然而, 863 计划以来, 汉语识别系统性能得到了很大程度的提高, 但汉语语音识别仍存在一系列的关键技术难题尚未解决, 其中之一就是针对方言的语音识别难题. 因此, 本文在 Windows7 的环境配置下, 利用 HTK 工具箱, 针对衡阳方言孤立词语音识别进行了深入研究.

1　阳方言特征分析

衡阳地区的通用方言大体上可分为两类: 湘语和赣语. 其中, 耒阳话和常宁话属于赣语, 其余地方属于湘语[3]. 从广义上说, 衡阳方言是指衡阳境内所有的汉语地方方言. 而从狭义上讲, 衡阳方言是指流行于衡阳市辖四主城区、衡阳县和衡南县的衡阳话, 也就是之前所说的衡阳小片. 本文的研究对象正是我们狭义上所指的衡阳方言(若无特殊说明, 在后文中的“衡阳方言”均指衡阳话).

接下来从三个方面对衡阳方言发音特征进行分析:

1) 声母系统

衡阳方言的声母系统包括零声在内共有 19 个. 其中, [f]、[x]易混读, 如“峰”“哄”部分; [l]与齐齿呼韵母配合的时候多半发音为[t], 如“礼”“理”发音为[ti]; 而在北京话中发音为零声母的时候, 衡阳方言里大部分发音为舌根鼻音, 如“我”发音为[ŋo], “按”发音为[ŋan], 等等.

2) 韵母系统

衡阳方言中韵母共有 37 个. 其中, [o]、[i]、[y]发音时的唇形相对普通话而言更随意, [o]、[y]发音的圆唇度稍微偏弱; [iu]发音时中间有过渡音[o], 在口语中的实际发音为[iou], 同样[ui]的实际发音为[uei]等.

3) 声调系统

衡阳方言有 6 个单声调, 即阴平、阳平、上声、阴去、阳去、入声; 其中阴平调是一个高平升调, 在六个声调中音调最高, 入声失去了塞尾声而变成了舒声.

根据衡阳方言的声韵母发音特点建立的音素模型, 能够很好的代表方言语音信号中的特征, 避免了汉语语音音素模型在方言中的不足, 从而在很大程度上提高了系统识别性能.

2　HTK工具基本介绍

HTK 最早是由英国剑桥大学开发用于建立基于 HMM 的语音识别工具. 它包括一系列的运行库和工具, 使用基于 ASNIC 模块化设计, 可以实现语音录制、分析、标示、HMM 的训练、测试和结果分析. HTK 以源代码的方式发布, 开发者在其官方网站上下载到最新版本的代码后, 即可在自己的计算机上编译获得可执行工具. 如图 1 所示, 基于 HTK 的语音识别系统大致可分为四个阶段: 数据准备, 模型训练, 识别测试和性能评测[4].

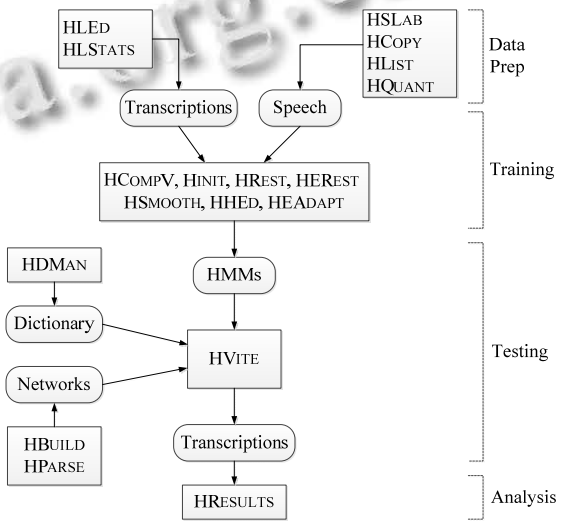

图 1　HTK 处理流程图

3　系统实现

本文所针对的方言是湖南省衡阳市标准的衡阳话, 利用 HTK 实现了衡阳方言的孤立词语音识别系统, 系统框图如图 2.

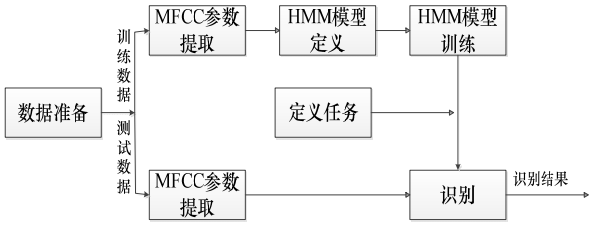

图 2　语音识别系统框架图

3.1 数据准备

为了对衡阳当地方言进行系统的研究, 找到当地讲衡阳方言相对流畅稳定的说话者进行录音. 使用 Cool Edit 软件进行录音, 保存为.WAV 格式, 采样频率为 16KHz, 采样精度为 16bit, 采用单声道, 录音环境

为室内相对安静的情况下, 针对常用的部分孤立词进行识别研究.

用包含变量的正则表达式来定义语法, 其内容为:

$name=guodi|nali|nilin|jilin|laiji|meiji|yiji|jiajia|jiafu|xingji|qiafan|qiaxu|kougan|qihuan|guanghua|luoyu|……;

(SENT-START ($name) SENT-END)

再用 HParse 函数工具根据任务语法生成 SLF 底层网络格式文件 wdent. 接下来采用声韵母的结构方式定义字典, 根据衡阳方言声韵母发音特性, 对上述的衡阳方言孤立词语进行构建字典, 部分内容如下:

guoli k o t i sp

nali n a t i sp

nilin n i n in sp

jilin tɕ i n in sp

……

最后标注数据, 用 HLEd 工具将字词级真值文本转换成音素级真值文本, 并保存为.mlf 格式.

3.2 MFCC 参数提取

本文中使用 MFCC 进行语音的特征参数提取, 同时还会用到 MFCC 参数的一阶、二阶 delta 系数. HTK 提供 HCopy 命令来实现特征参数点的提取功能. 在提取特征参数时, 需要定义一个符合 HTK 格式的配置文件, HCopy 按照配置文件中设定的参数来提取相应的特征参数. 在本文中, 配置参数文件内容如图 3 所示.

```
# Coding parameters
SOURCEKIND     = WAVEFORM
SOURCEFORMAT   = WAV
TARGETKIND = MFCC_0_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = T
SAVEWITHCRC = T
WINDOWSIZE = 250000.0
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = F
```

图 3 配置参数内容

在命令行(Command Processor, CMD)中, 提取特征矢量的命令为:

HCopy -T 1 -C config -S code.txt

运行 HCopy 命令后, 会按照 config 文件内的参数和 code.txt 文件左侧的路径进行语音的特征参数提取, 并将得到的特征参数按照 code.txt 文件内右侧的路径保存到特征文件中.

3.3 HMM 模型定义

语音特征参数准备好后, 接下来进行声学模型的训练, 即 HMM 模型训练. 在模型训练前, 我们需要定义一个初始化 HMM 模型, 保存为文本文件, 这个初始化模型的参数并不重要, 它的目的只定义 HMM 的初始结构[5]. 如图 4 所示, 本文中采用的 HMM 初始结构为含 5 个状态, 转移状态从左至右, 并且没有跨状态间的转移. 在描述模型的文件中, 包含了 HMM 模型的名称、特征参数向量大小(39 维)、特征提取方法(MFCC)、状态总数、模型的观察函数(使用带有对角矩阵的单一高斯观察函数)、观察函数的均值方差以及状态间的概率转移矩阵等信息.

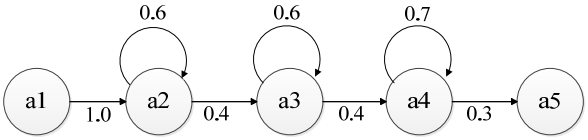

图 4 5 状态从左至右无跨越 HMM 模型结构

3.4 HMM 训练

定义好 HMM 模型后, 首先进行模型初始化, 利用 Viterbi 算法对 HMM 模型中的每一个参数进行训练, 生成完整的 HMM 模型库. HMM 模型训练的目的是通过对参数的重估迭代得到最佳值[6].

3.4.1 初始化

训练开始前, HMM 模型参数必须根据训练数据正确初始化, 这样可以加快训练速度, 使模型参数精准且收敛. HTK3.4.1 工具箱提供了两个不同的初始化工具: HInit/HRest 和 HCompV, 可完成所有模型的初始化工作. HInit/HRest 主要用于语料声学边界已完成标记的情况下进行单词分离式训练; HCompV 根据训练语音的特征参数计算其均值和方差, 再对 HMM 模型的高斯参数进行初始化[7].

本文中使用 HCOmpV 进行 HMM 模型初始化, 在 CMD 中输入命令:

HCompV -C config -f 0.01 -m -S train.txt -M hmms/hmm0 proto

执行命令后将会在 hmms/hmm0 目录下生成两个文件, 一个更新后的 proto, 另一个是截止宏 vFloors, 这个宏是全局平均方差的 0.01 倍, 后续的训练过程中所有的方差值将不能小于这个数. 其中-f 选项表示将

方差下限设置成全局方差的 0.01 倍.

利用更新后的 proto 模型, 得到各个单音素的模型, 再加上静音段 sp 的模型, 共同组成了初始的 HMM 模型.

3.4.2 HMM 训练

HMM 模型训练的关键问题是估计模型的参数 , 使得 在这个模型下最大化. 使用 HTK 工具箱中的工具进行模型训练, 通常有两种方法: 一种方法是直接使用 HERest 工具进行嵌入式训练; 另一种方法是根据基元的标注信息, 使用 HInit 和 HRest 工具对初始模型进行训练, 再使用 HERest 工具做进一步的 Baum-Welch 参数重估.

本文中, 使用 HERest 工具对模型进行嵌入式训练. 在 CMD 中的命令为:

HERest -C config -I phones0.mlf -t 250.0 150.0 1000.0 -S train.txt -H hmms/hmm1/macros -H hmms/hmm1/hmmdefs -M hmms/hmm2 lists/monophones0

为了使所有模型达到收敛, 训练次数越多, 模型越趋于稳定, 但考虑到时间成本, 一般在训练时, 重复进行两到三次即可得到较为稳定的模型.

3.4.3 建立绑定状态的三音素模型

建立基于上下文相关的三音素模型能够使声学模型更加适配语境, 从而增强 HMM 模型参数估算的鲁棒性和准确性. 首先利用 HHEd 工具初始化三音素模型, 再利用 HLEd 工具将字词级单音素标注文件生成三音素标注文件, 部分内容如下:

guoli k+o k-o+t o-t+i t-i sp
nali n+a n-a+t a-t+i t-i sp
nilin n+i n-i+n i-n+in n-in sp
jilin tɕ+i tɕ-i+n i-n+in n-in sp
……

最后利用 HHEd 工具函数采用决策树的聚类方法对三音素进行绑定.

3.5 识别

语音识别过程就是将待识别语音信号的特征参数与训练好的 HMM 模型进行匹配[8]. 先使用 HCopy 工具对待识别语音信号进行预处理, 得到待识别语音的 MFCC 特征参数; 然后使用 HTK 工具箱中的 HVite 工具进行识别, 在 CMD 中的命令为:

HVite -H hmms/hmm7/macros -H hmms/hmm7/hmmdefs -S test.scp -l * -i results/recout_step7.mlf -w wdnet -p 0.0 -s 5.0 dict/dict1 lists/monophones1

其中, -H 选项会载入识别所需要的声学模型, -s 选项确定语言学模型的权重因子, -S 选项会载入所需要识别的语音, -i 选项确定输出结果的存放位置和文件名, -w 选项会载入语音学模型.

最后使用 HTK 工具箱中的的 HResults 工具求出系统识别率[9]. 在 CMD 下执行命令为:

HResults –I labels/testwords.mlf lists/monophones1 results/recout_step7.mlf

其中, testword.mlf 为标注文件, monophones1 为单音素列表, recout_step7.mlf 为 HVite 的识别结果.

4 实验结果分析

实验选取了 40 个常用的衡阳方言词语, 包括了称谓名词 10 个、行为动词 10 个、自然名词 10 个和穿着词语 10 个共 40 个, 具体词汇见表 1. 由 4 个人(3 男 1 女)录音. 测试集中, 每个词录 3 遍, 共有 40*4*3=480 个语音样本; 测试集中, 每个词录 2 遍, 共有 40*4*2=320 个语音样本.

表 1 40 个常用衡阳方言孤立词

称谓名词	行为动词	自然名词	穿着词
果地(这里)	恰饭(吃饭)	窗叶子(窗户)	罩衣(外套)
阿地(那里)	恰诉(喝水)	麻拐凳(板凳)	罩裤(外裤)
你郜(你们)	口干(口渴)	丁板(砧板)	汗衣子(衬衫)
几郜(他们)	起欢(喜欢)	麻拐(青蛙)	汗褂子(背心)
赖几(男孩)	广话(讲话)	荚子(茄子)	盖衣(棉衣)
妹几(女孩)	落雨(下雨)	班椒(辣椒)	伐子(袜子)
姨几(阿姨)	天长(天晴)	王瓜(黄瓜)	还子(鞋子)
假甲(姐姐)	出报(丢脸)	汗菜(苋菜)	里衣(内衣)
假夫(姐夫)	抢客(请客)	洋芋子(土豆)	兜子裤(内裤)
醒几(婶婶)	晓得(知道)	豆瓜子(豆角)	红锁衣(毛衣)

注: 括号内的词语是根据衡阳方言孤立词音译的汉语普通话.

在实验中, 我们采用音素作为语音识别单元, 并比较了单音素和三音素模型下的识别结果, 实验结果如图 5 所示.

```
D:\htk\fangyan2>HResults -I labels/testwords.mlf lists/monophones1 recout9.mlf
====================== HTK Results Analysis =======================
  Date: Sat Aug 20 15:05:21 2016
  Ref : labels/testwords.mlf
  Rec : recout9.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=84.38 [H=270, S=50, N=320]
WORD: %Corr=94.79, Acc=94.79 [H=910, D=0, S=50, I=0, N=960]
===================================================================
```

(a)单音素模型下的识别结果

```
D:\htk\fangyan2>HResults -I labels/testwords.mlf lists/monophones1 recout15.mlf
====================== HTK Results Analysis =======================
  Date: Sat Aug 20 15:10:45 2016
  Ref : labels/testwords.mlf
  Rec : recout15.mlf
------------------------ Overall Results --------------------------
SENT: %Correct=89.69 [H=287, S=33, N=320]
WORD: %Corr=96.56, Acc=96.56 [H=927, D=0, S=33, I=0, N=960]
===================================================================
```

(b)绑定三音素 HMM 后的识别结果

图 5 音素模型下的识别结果

由图 5(a)中 SENT 这行可以看出, 此次用于测试的共有 320 个句子, 句子识别正确率为 84.38%, 由 WORD 这行内容可以看出, 字的正确率和准确率均为 94.79%; 图 5(b)可以看出, 绑定三音素 HMM 后句子的识别率和字的正确率分别为 89.69%、96.56%, 较单音素模型分别提高了 5.31%、1.77%; 比较图 5 中(a)、(b)可以得出, 与单音素模型相比, 三音素模型具有很好的抗噪性能, 同时提高了系统的识别率.

然而, 在 HMM 模型中, 每个状态的高斯混合数也将影响系统性能, 特别是进行大词汇量连续语音识别时, 随着高斯混合数的增加, 也使得 HMM 模型更加精确, 同时加快收敛速度. 但是, 在小词汇量孤立词的语音识别中, 增加高斯混合数, 也许会得到相反的结果, 同时增加了运算量. 因此, 在进行实验时, 选择一个合适的高斯混合数与 HMM 模型结合起来变得尤为重要. 本试验中, 以单音素为识别单元, 特征参数为 MFCC_0_D_A, 在 HMM 模型上添加高斯混合数, 实验结果见表 2.

表 2 高斯混合数目识别率的比较

Mix	SC/%	WC/%	WA/%
1	84.38	94.79	94.79
2	92.19	97.40	97.40
3	95.00	98.33	98.33
4	95.31	98.44	98.44
5	97.19	99.06	99.06
6	96.25	98.75	98.75
7	96.25	98.75	98.75
8	91.00	97.00	97.00

其中, SC 表示句子识别正确率, WC 表示字识别正确率, WA 表示字识别准确率.

从表 4 中可以看出: 随着 GMM 分量个数的增多, 识别率也逐渐提升, 但是随着 GMM 分量个数的增加, 识别率的提升也有所放缓, 当 GMM 分量个数增加到 5 个时, 识别率不再增加, 当 GMM 分量个数超过 5 个时, 识别率开始下降. 因此, 在实验时使用 5 个 GMM 分量来建立 HMM 模型.

5 结论

本文通过对常用的衡阳方言词汇的语音进行分析、预处理、特征参数提取, 以音素为识别单元, 建立了每个方言词汇的 HMM 模型, 采用 Viterbi 算法进行 HMM 模型训练和测试, 利用 HTK 工具箱实现了对衡阳方言的语音识别系统的设计. 实验表明, 利用该方法能够得到较高的识别率, 在接下来对衡阳方言的识别研究中奠定了基础; 但是, 由于在词汇的选择和样本数量的大小上存在不足, 在今后的工作中, 应建立一个完善的衡阳方言语料库, 针对每个方言词汇发音的特点, 实现大词汇连续方言的识别.

参考文献

1 李余芳.基于 HTK 的普米语孤立词的语音识别.云南民族

大学学报,2015,24(5):426–430.
2 孙一鸣,刘葳.基于 HTK 的日语连续语音识别系统的建立与研究.计算机光盘软件与应用,2013,(16):86–87.
3 黄玉莲.衡阳方言假声研究[硕士学位论文].长沙:湖南大学, 2013.
4 孙爱中,刘冰,张琬珍,等.基于 DSP 的语音识别系统研究与实现.现代电子技术,2013,(9):76–78.
5 Kuamr A, Dua M, Choudhary T. Continuous Hindi speech recognition using Gaussian mixture HMM. 2014 IEEE Students' Conference on Electrical, Electronics and Computer Science (SCEECS). Bhopal. 2014. 1–5.
6 曾妮,费洪晓,姜振飞.基于 HTK 的特定词语音识别系统.计算机系统应用,2011,20(3):157–160.
7 樊帅.基于Xilinx Zynq的说话人识别的研究与设计[硕士学位论文].成都:电子科技大学,2015.
8 王爱芸.语音识别技术在智能家居中的应用.软件,2015,(7):104–107.
9 魏巍,张海涛.一种基于 HTK 的数字语音识别系统.计算机系统应用,2011,20(9):17–21.