文章编号:1671-4598(2018)11-0252-04 DOI:10.16526/j.cnki.11—4762/tp.2018.11.055 中图分类号:TP399 文献标识码:A

小语料库重庆话语音识别的研究

张 策,韦鹏程,石 熙
(重庆第二师范学院 数学与信息工程学院,重庆 400065)

摘要:随着计算机技术的发展,人工智能产品已经开始广泛地应用在各个领域;利用地区方言与人工智能产品进行交流成为了人机交互技术领域一个重要的研究方向;地处西南的重庆市为国家定位的国际大都市,世界各种文化伴随着人流汇聚于此;承载着重庆本土文化的重庆话作为重庆地区的主要交流语言,研究重庆话语音识别在推动人工智能产品本土化有着积极的作用;文章以重庆话为研究对象,建立了重庆话和重庆话口音的普通话小语料库,搭建了以 HMM 为声学模型的语音识别系统,分别以重庆话和重庆话口音的普通话作为声学模型去分别识别重庆话和带重庆话口音的普通话;实验表明,重庆话和重庆话口音的普通话声学模型去识别对应语音的正确识别率均为 100%;重庆话声学模型识别重庆口音的普通话的正确识别率达到 78.89%,重庆话口音的普通话声学模型去识别重庆话的正确识别率达到 91.67%。

关键词:重庆话;小语料库;HMM 模型;语音识别;识别率

Study on Speech Recognition in Chongqing Dialect of Small Corpus

Zhang Ce, Wei Pengcheng, Shi Xi
(College of Mathematics and Information Engineering, Chongqing University of Education, Chongqing 400065, China)

Abstract: Regional dialect is an important carrier of cultural heritage in the region, and is an important part of the Chinese national language library. Located in the southwest, Chongqing city as a national positioning of the international metropolis, various cultures of the world gather here together with the flow of people. Chongqing dialect with Chongqing local culture is as the main communication language in Chongqing. The study of speech recognition in Chongqing dialect plays a positive role in promoting the localization of artificial intelligence products. This paper takes Chongqing dialect as the research object, establishes the small corpus of Chongqing dialect, and constructs a speech recognition system using HMM as the acoustic model, and compares the recognition results of Chongqing dialect and Mandarin in Chongqing dialect accent. Experimental results show that the correct recognition rate of acoustic model of Chongqing dialect and Mandarin in Chongqing dialect accent to recognize the corresponding speech was 100%; the correct recognition rate of the Chongqing dialect acoustic model recognize Mandarin in Chongqing dialect accent was 78.89%, the correct recognition rate of Mandarin in Chongqing dialect accent acoustic model to recognize Chongqing dialect was 91.67%.

Keywords: Chongqing dialect; small corpus; HMM; speech recognition; recognition rate

0 引言

语音识别技术[1]是人机交互领域的重要研究内容,解决了人机交互过程中计算机不能够听懂人说话的问题。语音识别技术起步于上世纪五十年代,发展至今已经取得了长足的进步。国内外很多科技公司都在语音识别领域进行了深入的研究。如谷歌、微软以及科大讯飞等公司已经走在了语音识别领域最前沿。目前研究语音识别主要的研究对象是主流的语言,而关于方言的研究就相对少些。

重庆话是重庆地区方言文化,承载着重庆本地的传统文化,人口覆盖超过 3 000 万。近些年来,重庆的电子信息产业已经成为重庆经济的重要增长极,在 2012 年重庆市提出的“两江有云,西永有端,南岸有网”的电子信息产业发展总战略布局背景下,人工智能产品将在重庆各个领域广泛应用。语音识别是人工智能领域的重要组成部分,实现了计算机能“听懂”人的语音。重庆话语音识别的研究将有助于实现重庆地区的人们能够自然地利用重庆话与人工智能产品进行交流,实现“人机对话”,从而让人们享受到科技发展给生活带来的便利和高效。

1 重庆话的发音特点

重庆方言虽然也属于汉语,但是和汉语普通话存在一些差异。在声母方面的差异,汉语普通话有 21 个声母[2],而重庆话中没有翘舌声母/zh/、/ch/、/sh/以及鼻音声母/n/;重庆话只有 17 个声母[3-6],重庆话不能区分/n/和/l/,也就是说没有鼻音声母/n/,并且通常把声母/h/读成/f/。在韵母方面的差异,汉语普通话共有 39 个韵母[2],而重庆话只有 37 个韵母[3-6],没有/ing/和/eng/这两个后鼻音韵母,多一个/vu/,少一个/ui/。汉语普通话、重庆话的声母和韵母分别如表 1、表 2、表 3 和表 4。

收稿日期:2018-05-09; **修回日期**:2018-05-25。
基金项目:重庆市教委科学技术研究项目(KJ1714355);重庆第二师范学院校级科研项目资助(KY201726C)。
作者简介:张 策(1988-),男,四川宣汉人,助教,硕士,主要从事语音信号处理、教育大数据处理方向的研究。
韦鹏程(1975-),男,广西河池人,教授,博士,主要从事保密通信、计算智能方向的研究。
石 熙(1980-),男,重庆奉节人,副教授,博士,主要从事信息安全、教育大数据处理方向的研究。

表 1　汉语普通话声母表

	双唇音	前齿音	齿音	齿槽音	卷舌音	上腭音	软颚音
塞音不送气	b			d			g
塞音送气	p			t			k
塞擦音不送气			z		zh	j	
塞擦音送气			c		ch	q	
擦音		f	s		sh	x	h
通音					r		
鼻音	m			n			
边音				l			

表 2　汉语普通话韵母表

类型	开口呼	齐齿呼	合口呼	噘口音
单元音	a,o,e,er	i,ia,ie,ii,iii	u,ua,uo,ui	v,ve
双元音	ai,ei,ao,ou	iao,iou	uai,uei	
鼻音	an,en,ang,	ian,in,	uan,uen,uang,	van,vn,
	eng	iang ing	ueng,ong	iong

表 3　重庆话声母表

	双唇音	前齿音	齿音	齿槽音	卷舌音	上腭音	软颚音
塞音不送气	b			d			g
塞音送气	p			t			k
塞擦音不送气			z			j	
塞擦音送气			c			q	
擦音		f	s			x	h
通音					r		
鼻音	m						
边音				l			

表 4　重庆话韵母表

类型	开口呼	齐齿呼	合口呼	噘口音
单元音	a,o,e,er	i ,ia,ie, ii,iii	u,ua,ue	v,vo,ve
双元音	ai,ei,ao,ou	iao,ian,		uai,uei
鼻音	an,en,ang	iei, iou,in,iang	uan,uen, uang	van,vn, vu, iong

2　重庆话识别方法

简单来讲，重庆话语音识别是利用声学模型匹配方法将输入是语音识别系统的待识别语音与经过训练的声学模型进行模式匹配，并按照一定的判别规则得到待识别语音对应的文本信息。

2.1　训练方法

语音识别过程中需要对语料库中的语音基元建立声学模型，并对语音基元的声学模型的参数进行训练[7]，得到含有语音特征信息的声学模型。对建立的声学模型的状态转移概率进行重估训练，重估训练的方法如公式（1）所示。

$$\hat{a}_{ij}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^r(t)a_{ij}b_j(O_{t+1}^r)\beta_j^r(t+1)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^r(t)\beta_i^r(t)} \tag{1}$$

其中：$b_j(O_{t+1}^r)$ 为输出概率分布，$\alpha_i^r(t)$ 为前向概率，$\beta_j^r(t+1)$ 为后向概率。考虑到状态从以前模型中转换出来而可能占用状态转移入口的情况，需利用公式（2）对状态转移概率进行嵌入式重估。

$$\hat{a}_{ij}^{(q)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(q)r}(t)a_{ij}^{(q)}b_j^{q}(O_{t+1}^r)\beta_j^{(q)r}(t+1)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)} \tag{2}$$

从 HMM 模型的非发射入口状态进入 HMM 模型的由公式（3）嵌入式重估完成。

$$\hat{a}_{1j}^{(q)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(q)r}(t)a_{1j}^{(q)}b_j^{q}(O_t^r)\beta_j^{(q)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_1^{(q)r}(t)\beta_1^{(q)r}(t)+\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)r}(t)} \tag{3}$$

然后，从 HMM 模型进入 HMM 模型的非发射入口状态由公式（4）嵌入式重估完成。

$$\hat{a}\ a_{iN_q}^{(q)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_i^{(q)r}(t)a_{iN_q}^{(q)}\beta_{N_q}^{(q)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)} \tag{4}$$

最后，从 HMM 模型的非发射入口状态进入 HMM 模型的非发射入口状态由公式（5）嵌入式重估完成。

$$\hat{a}_{1N_q}^{(q)}=\frac{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r-1}\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)r}(t)}{\sum_{r=1}^{R}\frac{1}{P_r}\sum_{t=1}^{T_r}\alpha_i^{(q)r}(t)\beta_i^{(q)r}(t)+\alpha_1^{(q)r}(t)a_{1N_q}^{(q)}\beta_1^{(q+1)}(t)} \tag{5}$$

在公式（2）、公式（3）、公式（4）以及公式（5）中的下标 q 表示嵌入式重估的次数，如果 q 没有明显的标注出来，嵌入式重估的输出概率分布公式和单个模型的输出分布是一样。然而，概率计算公式必须将公式（6）变成公式（7）才能实现从入口状态的转移。

$$U_j^r(t)=\begin{Bmatrix}a_{1j} & if\ t=1\\ \sum_{i=2}^{N-1}\alpha_i^r(t-1)a_{ij} & \text{otherwise}\end{Bmatrix} \tag{6}$$

$$U_j^{(q)r}(t)=\begin{Bmatrix}\alpha_1^{(q)r}(t)a_{1j}^{(q)} & if\ t=1\\ \alpha_1^{(q)r}(t)a_{1j}^{q}+\sum_{i=2}^{N_q-1}\alpha_i^{(q)r}(t-1)a_{ij}^{(q)} & \text{otherwise}\end{Bmatrix} \tag{7}$$

语音识别中训练声学模型的方法较多，以上 7 个公式仅仅是语音识别中对声学模型进行重估训练所涉及的基本公式。

2.2 识别方法

语音识别过程就是待识别语音的声学模型和声学模型库中的模型进行匹配，得到匹配度最高的声学模型即是识别结果。待识别语音的声学模型和声学模型库中的语音的匹配过程采用维特比算法实现。本文基于 HMM 模型的维特比算法基本思想是从观测序列 $O=(o_1,o_2,o_3,\ldots,o_t)$ 中求取给定模型 $\lambda=(A,B,\pi)$ 下的最大似然概率。维特比算法用于语音识别解码的公式[7]如下所示。

给定一个模型 M，设 $\Phi_j(t)$ 表示在 t 时刻观测到语音序列从 O_1 到 O_t 处于 j 状态的最大似然，那么 $\Phi_j(t)$ 如公式（8）所示。

$$\Phi_j(t)=\max_i\{\Phi_i(t-1)a_{ij}\}b_j(o_t) \tag{8}$$

其中：i 和 j 为不同的状态，a_{ij} 为状态转移概率，$b_j(o_t)$ 为输出概率密度如公式（9）所示。

$$b_j(o_t)=\prod_{s=1}^{S}\left[\sum_{m=1}^{M_s}c_{jsm}N(o_{st};\mu_{jsm},\sum\nolimits_{jsm})\right]^{\gamma_s} \tag{9}$$

其中：c_{jsm} 是第 m 个分量的权重，$N(o;\mu,\Sigma)$ 是具有均值向量 μ 和协方差矩阵 Σ 的多元高斯模型，o_{st} 是在时间 t 观测向量被分成 s 个独立的数据流。

3 重庆话识别过程

重庆话语音识别是将重庆话语音识别成文本的过程。重庆话语音识别分为两个过程，即训练过程和识别过程。其中训练过程是利用语料对声学模型进行训练，最终得到声学模型库；识别过程是将待识别的语音进行预处理，然后提取语音的特征参数，最后利用相应的识别方法实现语音识别，并对识别结果进行分析得到识别结果。重庆话语音识别过程如图 1 所示。

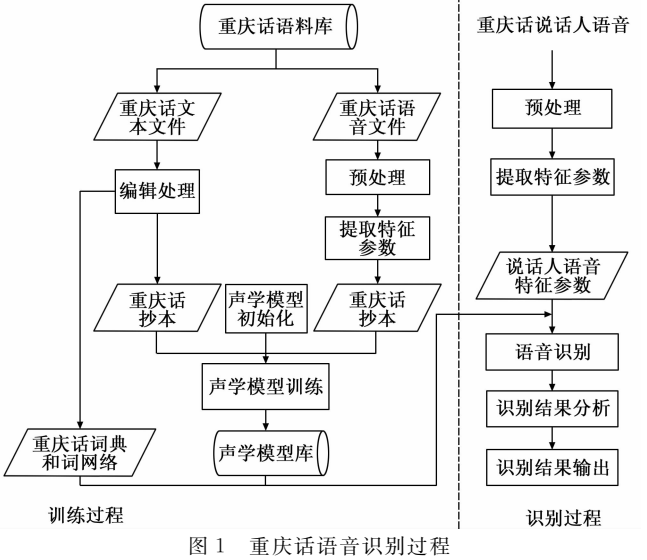

图 1 重庆话语音识别过程

3.1 建立重庆话语料库

首先采集本实验需要的语音文件对应的文本，选择重庆话和普通话发音标准的录音人；然后按照实验方案分别录制重庆话语音 30 句，重庆话口音的普通话语音 30 句，每句语音发音 10 遍，共得到（30＋30）＊10 句语音；最后由重庆话语音以及重庆话口音的普通话语音文本和与之对应的语音文件形成语料库。

语料库由训练集和测试集组成，训练集中包含（30＋30）＊7 句语料，测试集中包括（30＋30）＊3 句语料。其中测试集细分为（30＋30）＊1 句、（30＋30）＊2 句以及（30＋30）＊3 句。语料库中语音对应的文本如表 5 所示。

表 5 重庆话语料库文本与发音对照表

文本	重庆话拼音	重庆话口音的普通话拼音
中国银行	zong gue yin hang	zhong guo yin hang
南山街道	lan san gai dao	nan shan jie dao
运动鞋	yun dong hai	yu dong xie
等一下	den yi ha	deng yi xia
三间房子	san gan fang zi	san jian fang zi
牛角沱	liu ge tuo	niu jiao tuo
叫花鸡	gao hua ji	jiao hua ji
巷子	hang zi	xiang zi
敲门	kao men	qiao men
红岩广场	hong anr guang cang	hong yan guang chang
儿童研究院	er tong lian jiu yuan	er tong yan jiu yuan
眉毛	mi mao	mei mao
的确	di quo	di que
江湖菜	jiang fu cai	jiang hu cai
光脚板儿	guang juo banr	guang jiao banr
网约车	wang yuo ce	wang yue che
西南医院	xi lan yi wan	xi nan yi yuan
去吃饭	qi ci fan	qu chi fan
解放碑	gai fang bei	jie fang bei
咸味	han wei	xian wei
螃蟹	pang kai	pang xie
业绩	lie ji	ye ji
雷雨天气	luei yu tian qi	lei yu tian qi
长得很像	zang de hen qiang	zhang de hen xiang
眼泪	yan luei	yan lei
能量	len liang	neng liang
提不动	dia bu dong	ti bu ding
乘客	sen kie	cheng ke
孕育生命	run yu sen min	yun yu sheng ming
磨合	mo huo	mo he

3.2 语音预处理

语料库中的语音是连续且非平稳的信号，而非平稳的信号不便于处理，因此需要对语音信号进行预处理。预处理包括采样量化、分帧加窗以及预加重等过程。

1）采样量化是将连续语音信号转换成离散数字信号。本实验的语料库采用的采样量化标准是 16 kHz 采样、16 bit 量化。

2）分帧加窗是为了将语音信号进行短时化处理，我们可以认为语音信号长度在 10～30ms 时为准平稳信号，因此需要对语音信号加窗函数以实现短时化处理。常用的窗函数有矩形窗、哈明窗以及哈宁窗等，根据语音信号的特点，本文选取哈明窗函数，如公式（10）所示。

$$w(n)=\begin{cases}0.54-0.46\cos\left(\dfrac{2\pi n}{N-1}\right) & 0\leqslant n\leqslant N-1,\\ 0 & n<0 or n<N.\end{cases} \tag{10}$$

3）预加重是为了解决高频低功率谱的问题，即语音信号在高频部分呈现低能量，而低频部分呈现高能量的现象。在对语音信号进行处理分析过程中需要提高语音高频部分的功率谱，因此需要预加重处理。

3.3　提取特征参数

语音信号含有大量的信息，包括基频、时长以及频谱等基本声学参数，也包括语音韵律等信息。为了便于对语音信号的处理，去掉一些不太重要的冗余信息，因此需要对语音信号提取能够表征语音信号的相关参数，即语音信号特征参数。语音特征参数常见的语音特征参数有线性预测系数（linear predictive coefficients，LPC）、线性预测倒谱系数（linear predictive cepstral coefficients，LPCC）、基于 Mel 频率倒谱系数（mel frequency cepstral coefficients，MFCC）[8-9]。本论文根据声学建模的需要，选择接近人耳对语音信号频率的感知特性的特征参数。以上 3 种参数中的基于 Mel 频率倒谱系数（MFCC）作为特征参数。Mel 频率与 Hz 频率之间的映射关系如公式（11）所示。

$$f_{Mel}=(1000/\lg 2)\times\lg(1+0.001f_{Hz}) \tag{11}$$

MFCC 特征参数的产生过程如图 2 所示。

图 2　MFCC 参数提取过程

本文对语音信号提取的信号是 12 维的 MFCC 特征参数，为了反应语音信号的停顿及重音等参数需要加上 1 维短时平均能量构成 13 维特征参数，并且为了表示语音的动态特征，需对 13 维的特征参数求取一阶差分和二阶差分得到 39 维的特征参数。

3.4　训练声学模型

语音基元是发声的基本单元，本文是以重庆话的声韵母为语音基元，因此要为参与模型训练的声韵母建立声学模型。常用的声学模型较多，其中隐马尔可夫模型（hidden markov model，HMM）[10-12]是应用很广泛的声学模型。HMM 模型是由“单链”的马尔可夫演变为“双链”而来，其中一条隐藏的链描述了状态的转移，产生了不可观测的状态序列；另外一条可见的链描述了状态和观测值之间的统计对应关系。观察者只能通过可见的观测值来感知状态的转移关系。五状态的 HMM 模型如图 3[13]所示。

从图 3 中可以看出，由于 5 状态的 HMM 模型左右两端的状态只起到前后连接作用，这两个状态并没有高斯分布，因此 5 状态的 HMM 模型只有中间 3 个状态有状态转移。

声学模型 $p(y\mid x,\lambda)$ 在 HMM 模型中方可以变换如公式（12）所示。

$$p(y\mid x,\lambda)=\sum_{\forall q}p(y,q\mid x,\lambda)=\sum_{\forall q}P(q\mid x,\lambda)p(y\mid q,\lambda)=$$

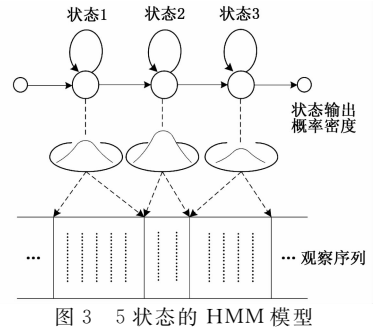

图 3　5 状态的 HMM 模型

$$\sum_{\forall q}P(q\mid x,\lambda)\prod_{t=1}^{T}p(y_t\mid q_t,\lambda) \tag{12}$$

其中：$P(\cdot)$ 表示一个概率密度函数，$p(y_t\mid q_t,\lambda)$ 是第 q_t 个状态的状态输出概率密度，它是一个典型对角协方差矩阵的单高斯分布，并且 $q=\{q_1,\ldots,q_T\}$ 是 HMM 状态序列。

为每一个语音基元建立了 HMM 模型之后，需要对 HMM 模型进行重估训练，训练方法如 2.1 节。对训练后的 HMM 模型建立 HMM 模型库，模型库中包含了（30＋30）＊7 句语料的所有基元对应的声学模型。

3.5　语音识别

语音识别是将待识别的语音识别成对应文本的过程，即在声学模型和语言模型下，对待识别语音的特征参数进行解码，从而将语音识别成对应的文本。

语音识别过程分为 4 个大组，每 1 个大组再以测试语句细分为 30 句、60 句以及 90 句 3 个小组，共计 12 组语音识别实验。具体的实验方案设计如下：

1）利用重庆话语音库中训练集的语料训练语音模型，重庆话语音库中测试集的语料为测试语句。

2）利用重庆话口音的普通话语音库中训练集的语料训练语音模型，重庆话口音的普通话语音库中测试集的语料为测试语句。

3）利用重庆话语音库中训练集的语料训练语音模型，重庆话口音的普通话语音库中测试集的语料作为测试语句。

4）利用重庆话口音的普通话语音库中训练集的语料训练语音模型，重庆话语音库中测试集的语料作为测试语句。

4　识别结果

根据以上 4 个大组，共 12 个小组的实验方案分别进行识别实验，并将实验结果整理如表 6 所示。

从表 6 中可以看出，重庆话和重庆口音的普通话对应识别自己本身的正确识别率为 100%，而两种语音交叉进行语音识别则呈现出不同的正确识别率。其中重庆话声学模型去识别重庆话口音的普通话在不同的测试集下呈现出不同的识别结果，当测试集为 30 句和 60 句时均为 76.67%，而在 90 句时达到 78.89%；重庆话口音的普通话声学模型去识别重庆话在不同的测试集下也呈现出不同的识别结果，随着测试集语句数的增加，正确识别率总体趋势上也随之增加，并在 30 句时达到 90.00%，60 句和 90 句时分别达到 91.67%和 91.11%。

（下转第 260 页）