文章编号:1674 - 2869(2018)06 - 0691 - 05

基于HMM的声调语音模型研究

易雪蓉[1],黄 巍*[1,2],胡 迪[1],蒋 怡[1]
1. 武汉工程大学计算机科学与工程学院,湖北 武汉 430205;
2. 智能机器人湖北省重点实验室(武汉工程大学),湖北 武汉 430205

摘 要:针对声韵母相同但声调不同的近音字识别问题和声韵母及声调都相同的同音字识别问题,提出在语音模型和语言模型中分别引入声调和字转移概率,以提高近音字和同音字的识别率。首先将声调划分为5种表现形式添加到汉语音节的最后一个音素中构成新音素,使用高斯混合隐马尔科夫模型建模新音素。然后通过统计方法计算特定语境下的字间转移概率。最后使用HTK工具包实现了带声调的语音模型和有字转移概率的语言模型。实验结果证明添加声调可以提高近音字的识别率,使用特定语境下字间转移概率可以提高同音字的识别率。

关键词:语音识别;隐马尔科夫模型;声调模型;转移概率
中图分类号:TP391 **文献标识码**:A **doi**:10.3969/j.issn.1674-2869.2018.06.021

HMM-Based Tone Speech Model

YI Xuerong[1], *HUANG Wei*[*1,2], *HU Di*[1], *JIANG Yi*[1]
1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. Hubei Key Laboratory of Intelligent Robot (Wuhan Institute of Technology), Wuhan 430205, China

Abstract: To improve the recognition rate of approximant characters with the same initial but different tones and the recognition accuracy of the homophonous characters with the same initial and tone, we introduced the tone and word transition probabilities into the models of speech and language respectively. Firstly, the tone is divided into five forms and added to the last phoneme of Chinese syllable to form a new phoneme, which was afterwards modeled by Gaussian mixed hidden Markov model. Then, we calculated the word transition probabilities in a specific context. Finally, we adopted the Hidden Markov Model Toolkit to realize the models of tonal speech and language with word transition probabilities. The experiments show that the tones can improve the recognition rate of approximant characters, and the use of word transition probabilities in a specific context can promote the recognition rate of homophonous characters.

Keywords: speech recognition; Hidden Markov Model; tone model; transition probability

语言是人类沟通的重要工具,语音识别是人工智能研究的重要领域。20世纪50年代,贝尔实验室设计了第一个语音识别系统,实现了对孤立数字的语音识别[1]。20世纪60年代,提出了时间归一化打分机制、音素动态跟踪技术和动态规划算法,有效地解决了语音信号的特征提取和不等长语音匹配问题[2]。20世纪70年代,模式识别思想、线性预测编码等技术被应用于语音识别中,识别对象从孤立词转移到连续语音[3]。20世纪90年代及以后,隐马尔科夫模型(hidden markov modol, HMM)、高斯混合模型(gaussian mixed model, GMM)被提出[4],基于GMM-HMM的语音识别框架得到广泛使用和研究,文献[5]通过改进语音特征参数相邻帧的相关性,进一步提高GMM-HMM的准确度;文献[6-7]使用GMM-HMM识别了连续语音的声调。目前,深度学习技术也被应用于语音处理

收稿日期:2018-08-14
作者简介:易雪蓉,硕士研究生。E-mail:1143152674@qq.com
***通讯作者**:黄 巍,博士,副教授。E-mail:wei.huang@foxmail.com
引文格式:易雪蓉,黄巍. 基于HMM的声调语音模型研究[J]. 武汉工程大学学报,2018,40(6):691-695.

系统[8-11]，由于它对训练数据和硬件资源有着极高的要求，限制了其使用范围。

现代汉语是一种有声调的特殊语音，音素和声调组合可以构成无数个多音字和同音字的发音。一方面，同一个汉字在不同的声调下代表不同的意义，另一方面，相同的读音可能代表完全不同的汉字，因此，与印欧语系的语言相比，声调和上下文信息对汉语语音的识别具有更重要的作用[12]。

本文在语音模型中添加声调，并使用字转移概率捕获上下文信息，修改HTK[13]工具包以适应汉语语音识别问题，实验结果证明了声调对近音字识别的重要性，同时字转移概率的引入能有效提高同音字识别的准确率。

1 基于HMM的声调语音模型

GMM-HMM语音识别系统的框架图如图1所示，其结构主要由3部分组成：语言模型、字典和语音模型[14]。

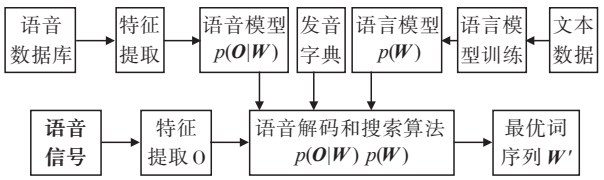

图1 语音识别系统框架

Fig. 1 Framework of speech recognition system

1.1 声调语音模型

GMM-HMM通常由 $\lambda=\{\boldsymbol{O},\boldsymbol{S},\pi,\boldsymbol{A},\boldsymbol{B}\}$ 来描述[15]，其中 $\boldsymbol{O}$ 代表 L 个观测向量集合 $\{o_1, o_2, \cdots, o_L\}$，$S$ 是 K 个HMM状态的集合 $\{s_1, s_2, \cdots, s_K\}$，$\pi=\{\pi_1, \pi_2, \cdots, \pi_K\}$ 是初始状态分布，$\boldsymbol{A}$ 是所有状态转移概率所构成的矩阵 (a_{ij})（a_{ij} 表示从状态 i 到状态 j 的转移概率），$\boldsymbol{B}$ 是状态观测符号的概率分布 $\{b_i(\boldsymbol{o}_j)\}_{K*L}$（$b_i(\boldsymbol{o}_j)$ 表示在状态 s_i 下观察到观测向量 $\boldsymbol{o}_j$ 的概率）。一个HMM的生成模型 M 如图2所示[13]，图2中1,2,3,4,5,6代表状态 s_1，s_2，s_3，s_4，s_5，s_6。

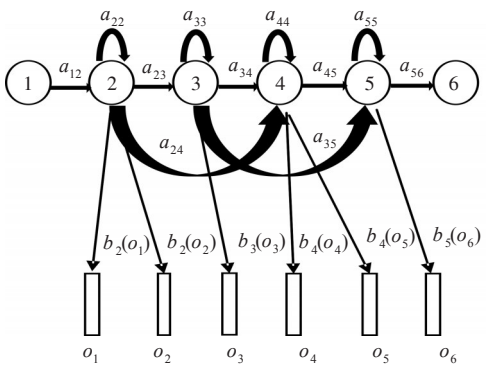

图2 HMM的生成模型

Fig. 2 HMM Generation Model

作为一个例子，在 M 的一个实例中，出现状态 $\boldsymbol{X}=(s_2, s_2, s_3, s_4, s_4, s_5)$ 并观察到观测向量序列 $\boldsymbol{O}=(o_1, o_2, o_3, o_4, o_5, o_6)$ 的概率为：

$$p(\boldsymbol{O},\boldsymbol{X}|M)=a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)a_{34}b_4(o_4)a_{44}b_4(o_5)a_{45}b_5(o_6)a_{56} \quad (1)$$

在基于GMM-HMM的语音识别应用中，$\boldsymbol{X}$ 是未知隐藏的，则：

$$p(\boldsymbol{O}|M)=\sum_X(a_{x(0)x(1)}\prod_{t=1}^{T}b_{x(t)}(o(t))a_{x(t)x(t+1)}) \quad (2)$$

$x(0)$ 是模型的初态，$x(T+1)$ 是模型的终态。

当观察到观测序列 $\boldsymbol{O}=[o(1), o(2), \cdots, o(t)]$ 时，最可能出现的未知状态序列 X 应该是使得观测向量序列 $\boldsymbol{O}$ 出现的可能性最大的状态序列，即：

$$\boldsymbol{X}=\max_x(a_{x(0)x(1)}\prod_{t=1}^{T}b_{x(t)}(\boldsymbol{o}(t))a_{x(t)x(t+1)}) \quad (3)$$

本文的实验中一个模型 M 对应一个音素 W，即 $p(\boldsymbol{O}|W)=p(\boldsymbol{O}|M)$。

在汉语中，一个汉字读音就是一个音节，每个基本音节由3个部分组成：声母、韵母和声调[16]，声母和韵母又是由音素组成的复合音。声母有23个，韵母有39个，音素包含辅音22个和元音10个，辅音对应声母，元音对应韵母。汉语拼音声母、韵母和音素对照见图3[17]，其中-i(前)为zi,ci,si发音的尾部部分，-i(后)为zhi,chi,shi发音的尾部部分。声调有4种，其中仅由声母和韵母构成并实际存在的声韵结合体据统计一共有400多个，将这些声韵结合体与音调组合成音节共记1 200多个[18]。在实际生活中，汉语口语中的音调不仅仅是一声、二声、三声和四声，还存在轻声。为了识别的准确性和全面性，在本文实验的声调模型中，除了标准规定的四种声调外，另加了一种轻声，构成5种声调，见表1。最后添加了声调的音素模型有81个，声调仅跟在每个音节的最后一个音素后(见图4)。

表1 声调模型对应表

Tab. 1 Mapping table of tone model

声调的表示	对应的声调
1	第一声(阴平)
2	第二声(阳平)
3	第三声[上(shǎng)声]
4	第四声(去声)
5	轻声

从图3和图4中可以看出：新模型与声韵母-音调组合相比较降低了复杂度，与传统音素模型相比较提高了精确度。部分传统音素从1个细分

声母：	b	p	m	f	d	t	n	l
音素：	b	p	m	f	d	t	n_e	l
声母：	g	k	h	j	q	x	zh	ch
音素：	g	k	h	j	q	x	zh	ch
声母：	sh	r	z	c	s	y	w	-
音素：	sh	r	z	c	s	i	u	ng
韵母：	a	o	e	i	u	ü	ai	ei
音素：	a	o	e	i	u	ü	a_i	ê_i
韵母：	ao	ou	ia	ie	iao	iou(iu)	ua	uo
音素：	a_u	e_u	i_a	i_ê	i_a_u	i_e_u	u_a	u_o
韵母：	uai	uei(ui)	üe	an	en	ang	eng	ong
音素：	u_a_i	uê_i	ü_ê	a_n	e_n	a_ng	e_ng	u_ng
韵母：	ian	in	iang	ing	iong	uan	uen(un)	uang
音素：	I_a_n	i_n	i_a_ng	i_ng	i_u_ng	u_a_n	u_e_n	u_a_ng
韵母：	ueng	üan	ün	er	-	-i（前）	-i（后）	
音素：	u_e_ng	ü_a_n	ü_n	e_r	ê	ix	iy	

图3　汉语拼音声母、韵母和音素对照图

Fig. 3　Comparison of Chinese Pinyin initials, finals and phonemes

b	p	d	t	g	k	f	h	x	z	m	sh
l	n	q	zh	ch	c	j	s	r			
a1	o1	e1	i1	o1	u1	ê1	ix1	iy1	r1	n1	ng1
a2	o2	e2	i2	o2	u2	ê2	ix2	iy2	r2	n2	ng2
a3	o3	e3	i3	o3	u3	ê3	ix3	iy3	r3	n3	ng3
a4	o4	e4	i4	o4	u4	ê4	ix4	iy4	r4	n4	ng4
a5	o5	e5	i5	o5	u5	ê5	ix5	iy5	r5	n5	ng5

图4　音素-声调模型内容

Fig. 4　Content of phoneme-tone model

成5个，让识别过程中的分类更加精细，如图5和图6所示，图5是传统音素建立HMM模型的示意图，音素相同发音不同的汉字所生成的HMM模型是一样的；图6是声调-音素建立HMM模型的示意图，添加声调模型后，音素相同发音不同的汉字所生成的HMM模型是不一样的。传统模型中音素相同发音不同的汉字因为共用相同的HMM模型，最后计算的 $p(\boldsymbol{O}|M)$ 相等，无法选取最优字；声调–音素模型中音素相同发音不同的汉字因为HMM模型的不同，最后计算出的 $p(\boldsymbol{O}|M)$ 不一样，根据实际情况选择可能性最大的概率，可以让识别结果更准确。

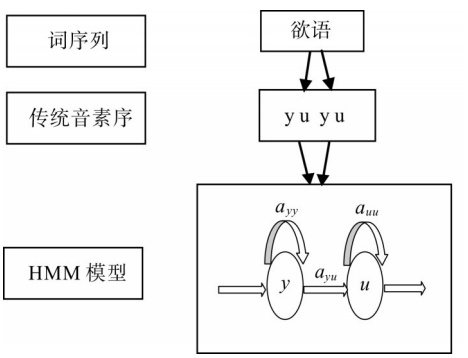

图5　基于传统音素的HMM模型示意图

Fig. 5　Schematic diagram of HMM model based on traditional phoneme

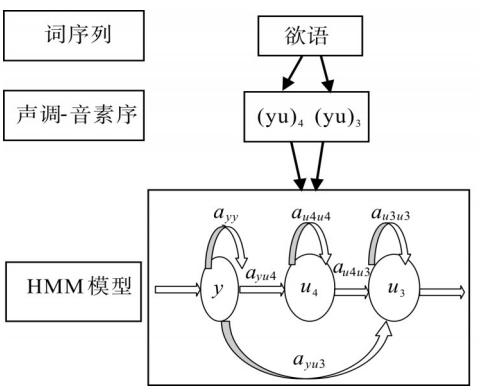

图6　基于声调-音素的HMM模型示意图

Fig. 6　Schematic diagram of HMM model based on tone-phoneme

1.2　字转移概率语言模型

语音识别应用中常用的语言模型是基于N-gram的统计语言模型。N-gram模型采用的是Markov假设[14]，即当前字出现的概率仅与前1个字有关系。

用 $A=(\text{start}, a_1, a_2, a_3, \cdots, a_m, \text{end})$ 表示一段待识别的字序列，a_i 表示其中的一个字，根据语音模型的处理结果，可以从词网中选取出 a_i 的所有同音字，然后计算每一个字出现的概率，选取概率最大的字组成最后识别出的字序列，若概率相同则选取同音字里出现的第一个字。

假设用 $w_1, w_2, w_3, \cdots, w_{m-1}, w_m$ 表示完整的句子中出现的每一个字，根据Markov假设，字 w_i 出现的概率为：

$$p(w_i|w_1, w_2, w_3, \cdots, w_{i-1}) = p(w_i|w_{i-1}) \quad (4)$$

整个句子出现的概率：

$$p(W)=p(w_1,w_2,w_3,\cdots,w_{m-1},w_m)=p(w_1)\times p(w_2|w_1)\times\cdots p(w_m|w_{m-1})=p(w_1)\prod_{i=2}^{m}p(w_i|w_{i-1}) \quad (5)$$

其中 $p(w_1)$ 表示 start 后出现字 w_i 的概率。这些概率在原始模型中全等于1,以至在同音字识别中正确率是不高的。

本文首先对训练数据进行统计,构建一个为全0的矩阵 $\boldsymbol{C}=(c_{ij})_{(N+2)*(N+2)}$,$c_{ij}$ 表示字 i 后面出现字 j 的概率,N+2 表示有 N 个无重复的汉字和表示开始与结束的 start 与 end;然后依次读取训练集,读取到字符 $\boldsymbol{X}$,就在矩阵的行中找到 $\boldsymbol{X}$ 的位置 x,接着读取下一个字符 $\boldsymbol{Y}$,在矩阵的列中找到该字符的位置 y,则 $c_{xy}=c_{xy}+1$,表示字 $\boldsymbol{X}$ 后出现 $\boldsymbol{Y}$ 的次数;最后对矩阵中的数值进行计算 $c_{xy}=\ln\frac{1+c_{xy}}{2c_{xy}}$。则字转移概率:$p(w_i|w_{i-1})=c_{w_{i-1}w_i}$。

2 实验部分

2.1 实验工具和数据

本文研究修改了 HTK 工具包,以得到支持声调和字转移概率的 GMM-HMM 语音识别模型。为验证声调信息和字转移概率对汉语语音识别的影响,分别进行了两组实验。实验一是对声韵母相同声调不同的近音字的识别;实验二是对声韵母和声调都相同的同音字的识别。

实验一所使用的语音数据集一为本研究收集的6个人对5组声韵母相同但声调不同的单个汉字的发音,共1 000条语音数据,其中每5个相同声韵结合体不同声调的孤立汉字为一组,每组有180个训练发音,20个测试发音。5组数据分别为:

1、ma1 妈,ma2 麻,ma3 马,ma4 骂,ma5 吗;

2、ya1 压,ya2 牙,ya3 雅,ya4 讶,ya5 呀;

3、mo1 摸,mo2 磨,mo3 抹,mo4 末,mo5 魡;

4、zuo1 作,zuo2 昨,zuo3 左,zuo4 坐,zuo5 呫;

5、qi1 七,qi2 奇,qi3 起,qi4 气,qi5 啐。

这5组数据中,第1组和第2组有着相同的韵母,不同的声母,目的是验证声母对声调发音的影响;第1、3组数据有相同的声母和不同的韵母,目的是验证韵母对声调发音的影响。

实验二所使用的语音数据集二为本研究收集的10个人对10句连续字的发音,共110条,其中100条训练发音,10条测试发音。10组训练数据为:

1、慢 man4 慢 man4 喜 xi3 欢 huan1 你 ni3;

2、我 wo3 在 zai4 雨 yü3 中 zhong1 漫 man4 步 bu4;

3、我 wo3 在 zai4 洗 xi3 衣 yi1 服 fu5;

4、再 zai4 见 jian4;

5、我 wo3 在 zai4 做 zuo4 作 zuo4 业 ye4;

6、我 wo3 在 zai4 做 zuo4 手 shou3 工 gung1;

7、作 zuo4 息 xi1 时 shi2 间 jian1;

8、小 xiao3 荷 he2 才 cai2 露 lu4 尖 jian1 尖 jian1 角 jiao3;

9、保 bao3 持 chi2 沉 chen2 默 mo4;

10、蓝 lan2 色 se4 墨 mo4 水 shui3。

这10组数据中,第1组和第2组有相同发音的“慢”和“漫”,第2、3、4、5、6组有相同发音的“在”和“再”,第1、3组有相同发音的“喜”和“洗”,第5、6、7组有相同发音的“做”和“作”,第9、10组有相同发音的“默”和“墨”,这几组数据可以用来验证字转移概率对同音字识别的作用。

2.2 实验过程

第一步:统计实验数据中的汉字,编辑语法文件,实验一中的语法规则是多选一,然后通过 HTK 命令将语法文件转换成可供计算机识别的“词网文件”;实验二中的语法规则是多选多,然后建立两个“词网文件”,分别是 HTK 命令自动生成的无字转移概率的词网文件 wnet1 和添加了字转移概率的词网文件 wnet2。

第二步:提取供训练的汉字语音文件的梅尔倒谱系数,转化成为特征矢量文件。

第三步:结合实验数据构建两个字典。字典一直接使用 HTK 命令生成,由汉字和音素组成,不含音调信息;字典二是在字典一的基础上添加声调信息,将声调与每个字的最后一个音素相结合,生成含有音调的字典。

第四步:构建音素和音素-声调两个列表。音素表只包含音素,而音素-声调表在音素表的基础上加入声调信息,在每个元音后加上声调,声母不变。

第五步:构建基于音素的隐马尔科夫模型 HMM1 和基于音素-声调的隐马尔科夫模型 HMM2,HMM1 和 HMM2 都被迭代训练了7次。

第六步:实验一和实验二分别使用了语音数据集一和语音数据集二,对比了无声调模型 HMM1 和有声调模型 HMM2 对近音字和同音字的识别效果。

2.3 实验结果

实验中正确率(Correct,α)定义如式(6),准确率(Accuracy,β)定义如式(7),其中 N 表示语音转译文件中的标签总数,D 表示删除错误的数量,S 表示替换错误的数量,I 插入错误的数量[13]。

$$\alpha=\frac{N-D-S}{N}\times 100\% \quad (6)$$

$$\beta=\frac{N-D-S-I}{N}\times 100\% \quad (7)$$

从实验一的结果（见表2）中可以看出，在识别孤立汉字时，声调模型对近音字识别结果的影响很大。无声调模型的识别结果均是词网中的第一个汉字，所以只有20%的正确性；而有声调模型基本可以有效的识别声韵母相同但声调不同的汉字，但是仍然有些错误。从图7中可以看出，一声比较容易被识别成二声，轻声容易被识别为四声，其原因是一声和二声均以平声结尾，轻声和四声均有些短促，所以容易被混淆。

表2　孤立字识别的正确率和准确率比较

Tab. 2　Comparison of correct rate and accuracy of isolated word recognition　%

测试组	无音调模型		有音调模型	
	正确率	准确率	正确率	准确率
1	20.00	20.00	100.00	100.00
2	20.00	20.00	95.00	95.00
3	20.00	20.00	95.00	95.00
4	20.00	20.00	100.00	100.00
5	20.00	20.00	100.00	100.00

测试2　错误的识别结果：　讶
　　原语音文本：　呀
测试3　错误的识别结果：　磨
　　原语音文本：　摸

图7　有音调模型识别结果错误对比

Fig. 7　Errors comparison of tonal model recognition

从实验二2次测试结果的正确率和准确率的比较结果（见表3）中可以看出，在连续汉语语音识别中，声调信息与字转移概率结合使用对同音字识别结果影响很大。在相同数据下，有字转移概率的识别正确率比没有字转移概率的正确率提升了20%左右，准确率也提升了30%左右。在没有字转移概率的识别中，系统会默认选择词网中第一个出现相同发音的字，在添加字转移概率后，系统会通过计算概率选择概率最大的字，因此正确率会提升。

表3　连续语音识别正确率和准确率比较

Tab. 3　Comparison of correct rate and accuracy of continuous speech recognition　%

	正确率	准确率
无字转移概率	74.47	44.69
有字转移概率	93.62	76.00

3　结　语

将汉语中的声调信息和字间转移概率引入基于GMM-HMM的语音识别系统，通过改造语音模型和语言模型，提高近音字和同音字的识别率。但仍然存在，比如轻声和四声的误判；连续语音中的三声容易出现插入错误等问题，预期解决这些问题能够进一步提高系统的识别率。

参考文献：

[1]　何湘智. 语音识别的研究与发展[J]. 计算机与现代化，2002(3)：3-6.

[2]　聂敏. 语音识别及其关键技术[J]. 无线通信技术，1999(4)：53-56.

[3]　禹琳琳. 语音识别技术及应用综述[J]. 现代电子技术，2013(13)：43-45.

[4]　侯一民，周慧琼，王政一. 深度学习在语音识别中的研究进展综述[J]. 计算机应用研究，2017，34(8)：2241-2246.

[5]　黄哲杉. 语音机器人隐马尔可夫算法探究[J]. 现代信息科技，2018(4)：95-98.

[6]　赵力，邹采荣，吴镇扬. 基于连续分布型HMM的汉语连续语音的声调识别方法[J]. 信号处理，2000，16(1)：20-23.

[7]　曹阳，黄泰翼，徐波. 基于统计方法的汉语连续语音中声调模式的研究[J]. 自动化学报，2004，30(2)：191-198.

[8]　DO J H，KANG O. Automatic prosodic tone choice classification with Brazil's into nation model[J]. International Journal of Speech Technology，2016，19(1)：95-109.

[9]　刘超. 语音识别中的深度学习方法[D]. 北京：清华大学，2015.

[10]　余尤好. 神经网络在通信系统回音对消中的应用[J]. 武汉工程大学学报，2012，34(9)：70-74.

[11]　张仕良. 基于深度神经网络的语音识别模型研究[D]. 合肥：中国科学技术大学，2017.

[12]　张登岐. "十五"高教版《现代汉语》的语法系统[J]. 阜阳师范学院学报(社会科学版)，2005(6)：60-68.

[13]　YOUNG S V，EVERMANN G N，GALES M，et al. The HTK book version 3.5 [EB/OL]. (2018-07-25) [2018-08-15].http://htk.eng.cam.ac.uk/

[14]　张强，陶宏才. 基于HTK的语音识别语言模型设计及性能分析[J]. 成都信息工程学院学报，2009，24(2)：142-146.

[15]　周盼. 基于深层神经网络的语音识别声学建模研究[D]. 合肥：中国科学技术大学，2014.

[16]　捷亚. 谈谈汉语拼音的字母教学[J]. 语文建设，1985(3)：45-46.

[17]　黄中伟，杨磊，徐明，等. 普通话语音识别中的基本音素分析[J]. 深圳大学学报(理工版)，2006，23(4)：356-357.

[18]　卢偃. 现代汉语音节的数量与构成分布[J]. 语言教学与研究，2001(6)：28-34.

本文编辑：陈小平