第 30 卷第 6 期　　　　　　　　计 算 机 应 用 研 究　　　　　　　　Vol. 30 No. 6
2013 年 6 月　　　　　　　Application Research of Computers　　　　　　　Jun. 2013

蒙古文字母到音素转换方法的研究*

飞　龙，高光来，闫学亮
（内蒙古大学 计算机学院，呼和浩特 010021）

摘　要：针对蒙古文字母到音素的转换(grapheme to phoneme conversion，G2P)问题，提出了基于规则的蒙古文 G2P 转换方法和基于联合序列模型的蒙古文 G2P 转换方法。实验结果表明，利用联合序列模型的蒙古文 G2P 转换方法要明显好于基于规则的蒙古文 G2P 转换方法。并且建立的基于联合序列模型的蒙古文 G2P 转换系统的词误识率为 16.32%，音素误识率仅为 3.37%，能达到实用要求。

关键词：蒙古文；字母到音素的转换；联合序列模型；联合多元；联合分割
中图分类号：TP391.1　　　**文献标志码：**A　　　**文章编号：**1001-3695(2013)06-1696-05
doi:10.3969/j.issn.1001-3695.2013.06.024

Research on grapheme to phoneme conversion for Mongolian

BAO Fei-long, GAO Guang-lai, YAN Xue-liang
(*College of Computer Science*, *Inner Mongolia University*, *Hohhot* 010021, *China*)

Abstract: This paper presented the rule-based Mongolian G2P conversion method and the statistic-based Mongolian G2P conversion method for Mongolian G2P conversion. Experimental results show that Mongolian G2P conversion method based on the joint-sequence model is significantly better than the rule-based Mongolian G2P conversion method. The word error rate is 16.32% and the phoneme error rate is 3.37% for the Mongolian G2P conversion system based on the joint-sequence model, and this system has reached the application requirements.

Key words: Mongolian; grapheme-to-phoneme conversion (G2P); joint-sequence models; joint multigram; co-segmentation

0　引言

字母到音素的转换模型可以应用到自然语言处理的很多领域，如语音合成(text-to-speech,TTS)即生成输入文本的发音信息、语音识别(speech recognition)即构造发音词典、语音关键词检测(keyword spotting，KWS)即把关键词转换成发音音素串等。

针对英语、法语、德语等语言，G2P 技术的研究始于 20 世纪末，先后提出了基于规则的转换方法[1~4]和基于统计的转换方法[5~9]。其中 Bisani 等人[9]将联合序列模型(joint-sequence models)应用到英语、德语和法语等语言的 G2P 转换中，取得了很好的转换效果。然而蒙古文 G2P 转换方法的研究尚未开展。

随着蒙古语语音识别[10]、语音合成和语音检索[11]等领域的深入研究，发现蒙古文 G2P 转换问题非常重要，而且成为一个亟待解决的课题。蒙古文的书面形式和口语形式不是一一对应的，会出现元音和辅音的增加、脱落及变换等现象。这种现象给蒙古文 G2P 转换工作带来了一定的难度。本文深入分析了蒙古文的发音特点，提出了基于规则的蒙古文 G2P 转换方法和基于统计的蒙古文 G2P 转换方法，并通过实验比较了上述两种方法。

1　蒙古文特点

蒙古文是拼音文字，它包含 8 个元音字母和 27 个辅音字母组，如表 1 所示[12]。在蒙古文中，常用的口语发音音素[13]共 63 个，其中元音音素 37 个，辅音音素 26 个，如表 2 所示。在蒙古文口语中，元音音素可分为两类：a)独立元音，这类元音音素发音比较清晰，比较完整，而且往往也是一个词的重音所在，独立元音包括词的第一音节的短元音以及词的任意长元音、复合元音；b)依附元音，这类元音音素通常发音不够清晰，不够完整，它包含非第一音节的所有短元音。

表1　蒙古文字母表

元音	拉丁转写	辅音	拉丁转写	辅音	拉丁转写	辅音	拉丁转写	辅音	拉丁转写
ᠠ	a	ᠩ	N	ᠰ	s	ᠸ	w	ᡀ	L
ᠡ	e	ᠨ	n	ᠱ	x	ᠹ	f	ᡁ	Z
ᠢ	i	ᠪ	b	ᠲ	t	ᠺ	k	ᡂ	Q
ᠣ	q	ᠫ	p	ᠳ	d	ᠻ	K		
ᠤ	v	ᠬ	h	ᠴ	c	ᠼ	C		
ᠥ	o	ᠭ	g	ᠵ	j	ᠽ	z		
ᠦ	u	ᠮ	m	ᠶ	y	ᠾ	H		
ᠧ	E	ᠯ	l	ᠷ	r	ᠿ	R		

蒙古文是黏着性语言，它没有前缀和中缀，只有后缀，它的构词是在词根或词干后缀接不同后缀生成新的单词[13]。蒙古

收稿日期：2012-10-09；**修回日期：**2012-11-14　　**基金项目：**国家自然科学基金资助项目(61263037;71163029)；内蒙古自然科学基金重大资助项目(2011ZD11)

作者简介：飞龙(1985-)，男(蒙古族)，内蒙古兴安盟科右中旗人，博士研究生，主要研究方向为蒙古文信息处理、语音识别与语音检索(csfeilong@imu.edu.cn)；高光来(1964-)，男，内蒙古兴安盟扎赉特旗人，教授，博导，主要研究方向为蒙古文信息处理、模式识别、人工智能；闫学亮(1984-)，男，天津人，硕士研究生，主要研究方向为蒙古文信息处理、信息检索.

表2　蒙古文音素表

基本元音	长元音	依附元音	复元音	借词元音	基本辅音	借词辅音
ɑ ə ɪ i ɔ ʊ o u æ œ ʏ	ɑː əː ɪː iː ɔː ʊː oː uː æː eː œː	ɑ̆ ə̆ ɪ̆ ɪ̆ ɔ̆ ŏ	ʊɪ ui ʊæ ue ʊɑ	y ɿ ʅ ɚ	b p w m s d t l r ʤ ʧ ʃ j g x ŋ	f k x ɬ ʣ ʦ ʥ ʈʂ ʂ ʐ

文的后缀可以分为构词后缀、构形后缀和结尾后缀。通过这种形式可生成近百万个蒙古文单词,而面向语音应用的发音词典是无法包含这么大规模单词。因此,在蒙古语语音合成、语音识别和关键词检测等领域的研究中,存在大量集外词(out of vocabulary)。解决这些集外词的 G2P 转换工作成为了亟待解决的关键问题。

2　基于规则的 G2P 转换方法

虽然蒙古文的书面形式和口语形式不是一一对应的,会出现元音和辅音的增加、脱落及变换等现象,但是大部分元音和辅音的增加、脱落和变换等现象是有规律可循的。这些转换规律包括蒙古文元音变读规律、辅音结合规律以及元音和谐律三个方面[13]。基于规则的 G2P 转换方法的流程如图 1 所示,首先对蒙古文单词依据元音变读规律进行转换,其次是按照辅音结合规律进行转换,最后进行元音和谐律的转换。具体介绍如下:

a)蒙古文元音变读规律。要看某一个蒙古文单词中有没有长元音、前化元音和复合元音的演变,另外还要注意短元音的相互演变。本文总结了 54 条元音变读规律,部分规律如表 3 所示。

表3　部分蒙古文元音变读规律

序号	书面语对应的拉丁转写	条件	对应的发音
1	aga	前一个音节出现元音 a、v、i 之一,并后一个音节出现元音字母 i	æ:
		前一个音节出现元音字母 w	ɔ:
		其他	ɑ:
2	iga	前一个音节出现元音字母 w	ɔ:
		其他	ɑ:
3	vga	前一个音节出现元音字母 a 或 v	ɑ:
		前一个音节出现元音 w	ɔ:
4	iya	前一个音节出现元音字母 a、v、i 之一	ɑ:
		前一个音节出现元音字母 w	ɔ:
		前一个音节出现元音音素 æ	æ:
		前一个音节出现元音音素 oe 或 oe:	œ:
		…	

b)蒙古文辅音结合规律。主要是说明蒙古文里哪些辅音可以共同在一个音节里直接相连组成一个具有复辅音的音节。具体规则如下:

(a)当词末音节是闭音节,且它前边也是一个闭音节时,则词末音节是稳定的;

(b)词末音节是闭音节,而它前边有开音节时,有两种不同情况:

①如果词末闭音节的两个辅音不能构成复辅音,那么词末音节仍然是稳定的;

②如果词末音节的两个辅音可以构成复辅音,那么它们之间的元音就脱落,两个辅音都依附于前面的开音节,从而组成复辅音,这样音节数减少一个;

(c)词末音节是开音节时,在口语里其元音或者脱落,或者转移:

①前边有开音节时脱落;

②前边有闭音节,而且可以组成复辅音的,也脱落;

③前边有闭音节,但不能组成复辅音的,元音转移到前边。

前两种情况下音节数减少,后一种情况下音节数不减少。如果词末元音脱落后形成的闭音节前边还有一个开音节,而且该闭音节的前两个辅音是可以构成复辅音的,那么它们就连续进行语音脱落,从而形成一个复辅音音节。这是从((c)①)演变出来的((b)②)现象,见于由三个音节组成的一部分词。

经过上述变读形成的词末音节,都属于稳定的闭音节,它们一般不再发生变化。如果在这种稳定闭音节前只有一个音节,该词也不再发生变化。因此,对所有的双音节词和部分三音节词来说,辅音结合或元音脱落的变化也就到此为止。

关于三个以上多音节词的音节转换,经过上述词末音节元音的脱落和词末稳定闭音节形成之后,如果在它(词末稳定闭音节)前边还有两个以上的音节,则该两个以上的音节继续按上述条件规则演变。这时该两个以上音节中的后边一个音节起末尾音节的作用。

c)蒙古文元音和谐律主要有两个内容:(a)一个词中前边的独立元音与后边的独立元音的关系,也就是独立元音之间的协调关系;(b)一个词中前边的独立元音与后边的依附元音之间的制约关系。具体规律如表 4 和 5 所示。

表4　蒙古文独立元音和谐律

序号	第一音节	第二音节	第三音节(第二音节后的音节)
1	ɑIʊæYɑ:I:ʊ: æ:ʊI ʊ̆I:ʊɑ	ɑ:I:ʊ:æ:ʊIʊæ ʊ̆I:ʊ̆æ:	ɑ:I:ʊ:æ:ʊIʊæ ʊ̆I: ʊ̆æ:
2	ɔɔ:œ oe:	I:ɔ:œ:	I:ɔ:œ:
3	ɔɔ:œœ:	ʊ:ʊIʊæ ʊ̆I: ʊ̆æ:	ɑ:I:ʊ:æ:ʊIʊæ ʊ̆I: ʊ̆æ:
4	əi u ə:i:u:ui ŭi:	ə:i:u:e:ui ue ŭi: ŭe:	ə:i:u:e:ui ue ŭi:ŭe:
5	o o:	o:i:	o:i:
6	o o:	u:ui ue ŭi:ŭe:	ə:i:u:e:ui ue ŭi:ŭe:

表5　独立元音与依附元音和谐律

序号	第一音节	第一音节后的音节
1	ɑ I ʊ	I(dʒ tʃ ʃ j 之后) ɑ̆ 其他辅音之后
2	ə i u	i(dʒ tʃ ʃ j 之后) ĕ 其他辅音之后
3	ɔ	I(dʒ tʃ ʃ j 之后) ɔ̌ 其他辅音之后
4	o	i(dʒ tʃ ʃ j 之后) ŏ 其他辅音之后

举例说明基于规则的蒙古文 G2P 转换过程。例如,蒙古文单词"ᠤᠷᠤᠭᠰᠢᠯᠠᠭᠤᠯᠬᠤ"(拉丁转写为 vrvgsilagvlhv,中文意思为促进、推进)。首先用元音变读规律完成元音变读,结果为:"ʊrʊgʃIʊ:lxʊ";其次完成辅音结合,结果为:"ʊrʊgʃIʊ:lʊx",这里包含三步:第一步 lxʊ→lʊx(根据辅音结合规律的(3.c),并成为稳定音节);第二步是 lʊ:(长元音、复合元音都属于稳定音节);第三步 rʊgʃI→rʊgʃ(根据辅音结合规律的(3.b),并成为稳定音节);最后完成元音和谐,结果为:"ʊrɑ̆gʃlʊlɑ̆x"。因此,单词"ᠤᠷᠤᠭᠰᠢᠯᠠᠭᠤᠯᠬᠤ"的口语读音为"ʊrɑ̆gʃlʊlɑ̆x"。

本文利用基于规则的 G2P 转换方法建立了基于蒙古文正字法词典的发音词典,并人工进行了校正。但发现基于规则的

G2P 转换方法有以下两个缺点:a)部分蒙古文不遵循元音变读规律和辅音结合规律;b)很难把蒙古文所有的发音规律都总结齐全。因此,本文运用了统计方法进一步对研究蒙古文 G2P 转换问题进行了研究。

3 基于统计的 G2P 转换方法

3.1 联合序列模型

蒙古文单词是由蒙古文字母串组成。假设 G 为蒙古文字母串集合,对应的口语发音音素串集合为 Φ。蒙古文的 G2P 问题可表述为

$$\varphi(g) = \arg\max_{\varphi' \in \Phi^*} p(g,\varphi') \tag{1}$$

式(1)表示对于蒙古文单词 $g \in G*$ 寻找最有可能的发音 $\varphi \in \Phi^*$。* 表示所有字符串的集合。通过此公式可得到一个蒙古文单词最有可能的发音音素串。

联合序列模型的基本思想是输入和输出序列共同可以生成包含输入和输出符号的联合单位的共同序列。简单情况下,每个单位带有零或一个输入符号和零或一个输出符号。这相当于有限状态转换器(FST)的传统定义。这种可以由多个输入和输出符号组成的单位称之为共同序列(co-sequence)或联合多元(joint multigram)。Bisani 等人[5]把这种字母和音素的联合多元(grapheme-phoneme joint multigram)简称为 graphone。

Graphone 是一对不等长的字母和音素序列的组合 $\hat{q} = \{g, \varphi\} \in Q \subseteq G^* \times \Phi^*$。本文使用 g_q 和 φ_q 分别表示 $\hat{q}$ 的第一和第二部分。如果 graphone 最多包含一个字母和一个音素,则称之为单数 graphone。Q 的列表可以从训练数据中获得,也可以通过手工指定。

在联合多元模型中,假设蒙古文单词对应的字母串和对应的发音生成一个共同的 graphone 序列。例如,蒙古文单词“ ”可以生成有八个 graphone 的序列,如图 2 所示。

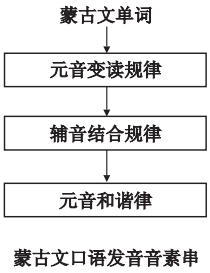

图1 蒙古文基于规则的 G2P转换流程

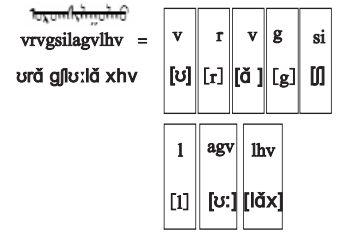

图2 蒙古文单词生成八个 graphone序列的结果图

字母和音素序列被分成相等的段数,这样的分组称为联合分割。对齐项是可以交换使用的。本文把这种特殊的对齐类型称为“m-to-n”。对于一个给定的输入和输出字符串对,分割 graphones 的结果不是唯一的。对于可能有歧义的 m-to-n 对齐,可以对输入的字母串进行自由的组合。例如,把蒙古文单词“ ”分割成 14 个 graphone 同样是有效的,如图 3 所示。

这种模糊的联合概率是由所有相匹配的 graphone 序列的总和来决定的。

$$p(g,\varphi) = \sum_{q \in S(g,\varphi)} p(q) \tag{2}$$

其中:$q \in Q^*$ 是 graphones 的一个序列,$S(g,\varphi)$ 是 g 和 φ 的所有联合分割的集合。

$$S(g,\varphi) := \left\{ q \in Q^* \;\middle|\; \begin{matrix} g_{q_1} \wedge \cdots \wedge g_{q_K} = g \\ \varphi_{q_1} \wedge \cdots \wedge \varphi_{q_K} = \varphi \end{matrix} \right\} \tag{3}$$

其中:$\wedge$ 表示序列的串联,$K = |q|$ 表示 graphone 序列 q 的长度。联合概率分布 $p(g,\varphi)$ 成为 graphone 序列 $q = q_1 \cdots q_K$ 上的概率分布 $p(q)$,它可以用标准的 M-gram 模型近似表示为

$$p(q_1^K) = \prod_{j=1}^{K+1} p(q_j | q_{j-1}, \cdots, q_{j-M+1}) \tag{4}$$

位置 $j<1$ 和 $j>K$ 是被视为特殊的边界符号 $q_j = \perp$,它允许作为蒙古文单词的开始和结束位置的特征现象来建模。下面介绍对于这种模型的估计方法和对应的实验评估。

3.2 模型估计

3.2.1 Multigram 的最大期望值

考虑在不是联合分割的训练数据中对可变长度单元模型的推理问题。给定 N 个蒙古文单词和对应的口语发音的训练样本 $O_1, \cdots, O_N = (g_1, \varphi_1), \cdots, (g_N, \varphi_N)$,但是字母和音素没有水平对齐。首先,由于一个联合分割 S 定义唯一的联合序列,发现如果有一个联合序列模型,就可以计算每个训练样本的任何联合分割概率:

$$p(g,\varphi,S) = p(q) \tag{5}$$

因此,训练数据的对数似然值可以用所有分割的总和来表示。

$$\log L(O_1, O_2, \cdots, O_N) = \sum_{j=1}^{N} \log L(O_j) = \sum_{j=1}^{N} \log\Big(\sum_{S \in S(\varphi j)} P(\varphi_j; S)\Big) \tag{6}$$

在联合单位中分割 S 是一个隐藏的变量。最大似然率训练可以采用期望最大化算法(EM)。本文首先考虑上下文独立的 unigram($M=1$)情况,更新参数 θ' 的重估公式如下:

$$p(q;\theta) = \prod_{i=1}^{|q|} p(q_i;\theta) \tag{7}$$

$$e(\hat{q};\theta) = \sum_{i=1}^{N} \sum_{q \in S(g_i,\varphi_i)} p(q | g_i, \varphi_i; \theta) n_{\hat{q}}(q) = \sum_{i=1}^{N} \sum_{q \in S(g_i,\varphi_i)} \frac{p(q;\theta)}{\sum_{q' \in S(g_i,\varphi_i)} p(q';\theta)} n_{\hat{q}}(q) \tag{8}$$

$$p(\hat{q};\theta') = \frac{e(\hat{q};\theta)}{\sum_{\hat{q}'} e(\hat{q}';\theta)} \tag{9}$$

其中:$n_{\hat{q}}(q)$ 是在序列 q 中 graphone $\hat{q}$ 出现的次数。把 $e(\hat{q};\theta)$ 称其为 q 的证据(evidence),它表示在当前的参数 θ 下训练样本中 $\hat{q}$ 出现的期望值。$e(\hat{q};\theta)$ 可以通过前向后向过程计算得到。

对于高阶模型($M>1$),用 h 来表示在前边的联合单元序列 $h_j = (q_{j-M+1}, \cdots, q_{j-1})$;用 $n_{q,h}(q)$ 来表示在序列 q 中 M-gram $q_{j-M+1}, \cdots, q_{j-1}$ 出现的次数。重估公式如下:

$$p(q;\theta) = \prod_{i=1}^{|q|} p(q_i | h_i; \theta) \tag{10}$$

$$e(\hat{q},h;\theta) = \sum_{i=1}^{N} \sum_{q \in S(g_i,\varphi_i)} p(q | g_i, \varphi_i; \theta) n_{q,h}(q) = \sum_{i=1}^{N} \sum_{q \in S(g_i,\varphi_i)} \frac{p(q;\theta)}{\sum_{q' \in S(g_i,\varphi_i)} p(q';\theta)} n_{q,h}(q) \tag{11}$$

$$p(\hat{q} | h;\theta') = \frac{e(\hat{q},h;\theta)}{\sum_{\hat{q}'} e(\hat{q}',h;\theta)} \tag{12}$$

此外,默认序列 q 中已经包含了开始和结束边界标志。

显然,上述公式不允许新的 graphone 出现的概率为零,所以通过人工设置比较满意的长度约束,并均匀分布到所有 graphones 来初始化模型参数。通常只使用一个简单的上限 L,即 $|g_q| \leqslant L$ 和 $|\varphi_q| \leqslant L$,但排除了不会生成的情况 $|g_q| = |\varphi_q| = 0$。

所以,更复杂的约束是可想而知的,如字母和音素序列长度的不同范围或下限设置。一个被公认的初始分布是 graphones 总数的倒数:

$$p_0(q) = [\sum_{l=0}^{L}\sum_{r=0}^{L} |G|^l |\Phi|^r]^{-1} \tag{13}$$

其中:$l = r = 0$ 表示附加的序列结束标志。

Graphone 的长度约束参数 L 对 graphone 的数目有明显的影响。序列模型的其他外部参数是最大极限的历史长度 M。M 和 L 一起规定了模型的有效范围,即在给定的位置字母或音素的数目影响估计的概率值。

一般情况下,用最大似然估计法训练模型时很可能会出现过拟合现象,并且在预测未出现的数据时效果不佳。同样,从训练样本中分析得到的一些单调初始化的 graphone 会达到某个概率聚集,而只有其中的小部分有助于"正确"的模型估计。这两个问题分别通过下面讨论的平滑和裁剪进行处理。

3.2.2　证据裁剪

证据裁剪可以解决过拟合问题。也就是说,修剪低于阈值的证据值,取代在式(12)中的 $p(\hat{q}|h;\theta)$。

$$p(\hat{q}|h;\theta) = \begin{cases} 0 & \text{if} \quad e(\hat{q},h;\theta) < \tau \\ e(\hat{q},h;\theta) & \text{otherwise} \end{cases} \tag{14}$$

此过程不可能在迭代过程中使 graphones 逐渐消失。证据裁剪同时有效地控制了 graphone 列表的大小。在训练数据上,阈值 τ 需要进一步调整。

3.2.3　减值证据

比较估计式(12)和典型的 N-gram 语言模型注意到,除了用证据值替代传统的 N-gram 计数值面临着本质上是相同的建模问题。众所周知,有效的平滑技术对建立好的语言模型是至关重要的。实证研究表明,用插值和边缘保留回退分布作绝对减值,也被称为 Kneser-Ney 平滑,比所有其他已知的平滑方法的效果都要好。不同于传统语言模型的计数值,证据值是一个小数。所以采用从传统的语言模型获取的结果时必须谨慎,因为它们的推导可能依赖于整数计数的假设。绝对减值和插值估计方程如式(15)所示。

$$p_M(\hat{q}|h) = \frac{\max\{e(\hat{q},h) - d_M, 0\}}{\sum_{q'} e(\hat{q}',h)} + \lambda(h) p_{M-1}(\hat{q}|\bar{h}) \tag{15}$$

为清楚起见,添加了一个下标 M 表示分布的阶数。$d_M \geqslant 0$ 是减值参数;$p_{M-1}(\hat{q}|\bar{h})$ 是广义的,低阶(M-1)-gram 的分布取决于减少的历史 $\bar{h}_i = (\hat{q}_{i-M+2}, \cdots, \hat{q}_{i-1})$;$\lambda(h)$ 为归一化参数,它使得所有的分布总和到 1。

在语言模型中的最小计数值为 1(除了未见过的事件),然而证据值可以变得任意小,实际上小于减值。所以减值的证据估计包含证据裁剪的一种形式:用低于减值参数的证据值的 graphones 拒绝进入模型。证据裁剪这种形式和明确的形式式(14)之间的一个显著区别是在减值里对未见过的事件分配了减值证据,而在其余的证据里有效地分配到了所有可见的事件。

仍然需要指定回退分布 p_{M-1}。对减少历史 $\bar{h}$ 本文利用一致性约束:

$$\sum_{h\in\bar{h}} p_M(\hat{q}|h) \sum_{q} e(\hat{q}',h) = \sum_{h\in\bar{h}} e(\hat{q},h) \tag{16}$$

替代式(15),并对 $p_{M-1}(\hat{q}|\bar{h})$ 在 p_{M-1} 是标准合格率的约束下解决:

$$p_{M-1}(\hat{q}|\bar{h}) = \frac{\hat{e}(q,\bar{h})}{\sum_{q'} \hat{e}(q',\bar{h})} \tag{17}$$

用 $\hat{e}$ 表示减少的证据:

$$\hat{e}(q,\bar{h}) := \sum_{h\in\bar{h}} \min\{e(q,h), d_M\} \tag{18}$$

当然,式(17)中的 $p_{M-1}(\hat{q}|\bar{h})$ 也需要平滑处理。平滑 p_{M-1} 的两个方法看上去比较合理:a)在式(15)中"插入"减少的证据值式(18);b)平滑约束条件。

事实证明,除了对减值参数不同的解释,这两种方法会得到相同的结果。绝对减值递归地应用于低阶分布 p_{M-2},$p_{M-3}, \cdots, p_0$。零元分布 p_0 与所有潜在的 graphones 式(13)相同。由于小数的证据值不适合它自己的运算,所以本文依靠 Powell 方法[14]在持有集(the hold-out set)上优化减值参数 d。

3.2.4　自底向上的模型建立和减值期望最大化

迭代过程中,本文用单调的概率分布初始化 unigram 模型式(13),即所有可能的 multigrams 有相同的初始概率。在训练集上,用不受约束的计数 $c(q)$ 选择性地初始化,即在每个词中不管相邻的 graphones 的重复,计 graphone 的出现次数。

$$c(q) := \sum_{i=1}^{N} \sum_{l_1=1}^{|g_i|} \sum_{l_2=l_1}^{|g_i|} \sum_{r_1=1}^{|g_i|} \sum_{r_2=r_1}^{|g_i|} \delta((g_{l_1} \wedge \cdots \wedge g_{l_2}, \varphi_{r_1} \wedge \cdots \wedge \varphi_{r_2}) = q) \tag{19}$$

通过应用平滑方法式(15),这些计数(受 graphone 长度约束的限制)被用于计算初始的概率分布。高阶的 M-gram 模型使用以前生成的(M-1)-gram 模型进行初始化。这意味着在低阶模型中只允许与不减值的 M-gram 相符合的历史。

本文要解决的是证据减值怎样与 EM 算法相互进行交互的问题。首先,优化减值需要数据集,该数据集独立于计算证据值时的数据集。不分离这些数据集导致减值的总值会低估。为此,本文从训练数据中分离训练集 O_t 和典型的较小的持有集 O_h。训练集用于计算证据值,而持有集用于调整减值参数。

在每个迭代中原始的 EM 算法确实提高了样本出现的可能性,但这通常会导致过度拟合和在某个点上持有集的可能性将开始减小。因此,在减值 EM 算法中为了确保持有集的可能性不降低,会更新减值。

3.3　解码

估计模型之后,式(1)可以用于集外词的音素转换。从正字形式转换成音素形式时,通常用极大值来近似式(2)中的总和。

$$p(g,\varphi) \approx \max_{q\in S(g,\varphi)} p(q) \tag{20}$$

具体来说,对给定的字母串寻找最有可能对应的 graphone 序列,并转换成对应的音素串。

$$\varphi(g) = \varphi(\max_{q\in Q^* | g(q)=g} p(q)) \tag{21}$$

4　实验

本文以蒙古文正字法词典为实验对象。该词典包含了 38 278个蒙古文单词。为了实验的方便性,将所有蒙古文单词都转换成对应的拉丁转写。本文用到的蒙古文字母为 35 个(元音 8 个,辅音 27 个),音素为 63 个(元音 37 个,辅音 26 个)。性能评价标准为词误识率(the word error rate,WER)和音素误识率(the phoneme error rate,PER)。

$$\mathrm{WER} = 1 - \frac{N_{\mathrm{correct}}}{N_{\mathrm{total}}} \tag{22}$$

$$\mathrm{PER} = \frac{N_{\mathrm{ins}} + N_{\mathrm{del}} + N_{\mathrm{sub}}}{N_{\mathrm{phtotal}}} \tag{23}$$

其中：N_{correct}为转换正确的蒙古文单词数目；N_{total}为所有需要转换的蒙古文单词数；N_{phtotal}为所有需要转换的蒙古文单词对应的音素个数总合；N_{ins}为转换时出现的插入错误个数；N_{del}为转换时所有出现的删除错误总合；N_{sub}为转换时所有出现的替换错误总合。

本文首先通过基于规则的 G2P 转换方法对蒙古文正字法词典进行 G2P 转换，并且人工进行了校正。对于正字法词典的所有单词用基于规则的 G2P 方法进行转换后，词误识率(WER)达到了 32.3%，音素误识率(PER)达到了 7.6%。

利用联合序列模型训练时，本文选择的训练集含 3.4 万个词，测试集含 4 278 个词。持有集的大小为训练集的 5%，平滑算法采用 Kneser-Ney。由于在联合序列模型中 graphone 的长度上限 L 和 M-gram 的阶数 M 的大小会直接影响模型的复杂度和实验结果，因此，本文通过一系列对比实验以获得最优参数。

图 4 为在不同的 graphone 长度上限 L 和 M-gram 的阶数 M 的情况下蒙古文 G2P 转换的音素误识率结果。从图中可以看出，当 $M=1$ 时，$L=4$ 的效果最好，$L=1$ 的效果最差；但是当 $M>4$ 时，$L=1$ 的实验结果明显要好于其他情况，并且随着 M 的增加会越发明显。

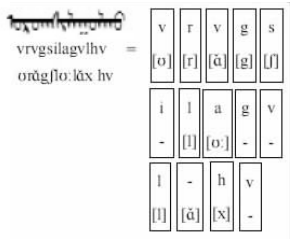

图3 蒙古文单词生成14个graphone序列的结果

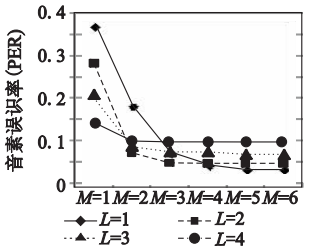

图4 在测试集上L和M不同值时G2P转换结果比较

当 $L=1$ 时，在测试集和训练集上对 M 作了进一步的实验，实验结果如表 6 所示。从表中可以看出，当 $M=9$ 时实验结果为最优，测试集上的词误识率达到了 16.32%，音素误识率达到了 3.37%，训练集上的词误识率达到了 2.66%，音素误识率达到了 0.5%；当 $M>9$ 时，模型会越来越大，但是它的词误识率和音素误识率则不会再降低。

表 6 在 $L=1$ 时不同 M 值的 G2P 转换结果

	测试集		训练集	
	WER/%	PER/%	WER/%	PER/%
$M=1$	99.28	36.99	99.19	36.94
$M=2$	72.33	18.27	71.53	17.98
$M=3$	33.24	7.15	28.17	5.85
$M=4$	20.69	4.3	11.33	2.22
$M=5$	17.15	3.53	6.06	1.15
$M=6$	16.57	3.42	4.13	0.79
$M=7$	16.42	3.38	3.49	0.67
$M=8$	16.35	3.37	3.38	0.66
$M=9$	16.32	3.37	2.66	0.5
$M=10$	16.32	3.37	2.66	0.5

5 结束语

本文详细描述了蒙古文的口语发音规则，并用这些规则构建了基于规则的 G2P 转换系统。基于规则的 G2P 转换系统的词误识率为 32.3%，音素误识率为 7.6%。但是基于规则转换方法有很多不足之处，如蒙古文的发音规则很难总结齐全；很多蒙古文单词不遵守发音规则等。所以，本文进一步研究了基于统计的蒙古文 G2P 转换方法，详细介绍了联合序列模型的相关算法和蒙古文 G2P 转换中的应用，并建立了基于统计方法的蒙古文 G2P 转换系统。为了获得联合序列模型相关参数的最优值，本文进行了一系列实验。实验表明，持有集为训练数据的 5%，graphone 的长度上限 $L=1$，M-gram 的阶数 $M=9$ 时，蒙古文 G2P 转换的误识率为最低。如表 6 所示，基于统计的 G2P 转换方法对集外词转换时词误识率的最低为 16.32%，音素误识率的最低为 3.37%。所以从结果上看，基于统计的 G2P 方法要明显优于基于规则的 G2P 转换方法。

本文结合蒙古文的特点，建立了基于规则的 G2P 转换方法和基于统计的 G2P 转换方法，并得到了很好的实验效果，这对蒙古文语音合成、语音检索和语音识别的研究有着非常重要的意义。

参考文献：

[1] MENG H M, SENEFF S, ZUE V W. Phonological parsing for bi-directional letter-to-sound / sound-to-letter generation [C]//Proc of Workshop on Human Language Technology. 1994: 289-294.

[2] TORKKOLA K. An efficient way to learn English grapheme-to-phoneme rules automatically[C]//Proc of IEEE International Conference on Acoustics, Speech, and Signal Processing. 1993:199-202.

[3] BAGSHAW P C. Phonemic transcription by analogy in text-to-speech synthesis: novel word pronunciation and lexicon compression [J]. Computer Speech & Language, 1998, 12(2): 119-142.

[4] MENG H. A hierarchical lexical representation for bi-directional spelling-to-pronunciation/pronunciation-to-spelling generation [J]. Speech Communication, 2001, 33(3): 213-239.

[5] BISANI M, NEY H. Multigram-based grapheme-to-phoneme conversion for LVCSR[C]//Proc of INTERSPEECH. 2003: 933-936.

[6] BELLEGARDA J R. Unsupervised, language-independent grapheme-to-phoneme conversion by latent analogy[J]. Speech Communication, 2005, 46(2):140-152.

[7] WANG Dong. Out-of-vocabulary spoken term detection[D]. Edinburgh: University of Edinburgh. 2010.

[8] TAYLOR P. Hidden Markov models for grapheme to phoneme conversion[C]//Proc of INTERSPEECH. 2005: 1973-1976.

[9] BISANI M, NEY H. Joint sequence models for grapheme-to-phoneme conversion[J]. Speech Communication, 2008, 50(5):434-451.

[10] BAO Fei-long, GAO Guang-lai. Improving of acoustic model for the mongolian speech recognition system [C]//Proc of Chinese Conference on Pattern Recognition. 2009: 616-620.

[11] BAO Fei-long, GAO Guang-lai, SU Xiang-dong. The research on mongolian spoken term detection based on confusion network[C]// Proc of Chinese Conference on Pattern Recognition. 2012: 606-612.

[12] 确精扎布. 蒙古文编码[M]. 呼和浩特：内蒙古大学出版社. 2000:4-8.

[13] 清格尔泰. 蒙古语语法[M]. 呼和浩特：内蒙古人民出版社, 1991:77-133.

[14] PRESS W H, TEUKOLSKY S A. Numerical recipes in C[M]. Cambridge: Cambridge University Press, 1992.