基于噪声信道模型的维吾尔语央音原音识别*

艾山·吾买尔　吐尔根·依不拉音　早克热·卡德尔
（新疆大学信息科学与工程学院，新疆，乌鲁木齐，830046）
Email:Hasan1479@xju.edu.cn

摘　要：该文提出了一种基于噪声信道模型的维吾尔语弱化元音恢复方法。该方法用噪声信道模型来描述维吾尔语词干元音的弱化过程，即词干中的部分元音在信道传输过程中被噪声发生弱化。本文根据维吾尔语的元音和谐、辅音和谐以及音节结构等特点，从词尾提取的二字符、三字符和最后音节等基础上建立语言模型和似然度计算公式，根据实验结果选择表现最好的词干词尾三字符模型。在基于有限状态自动机的维吾尔语名词词干提取系统中采用该方法处理不符合规则的弱化现象时，提高词干提取准确率 15%以上。

关键词：噪声信道；维吾尔语；元音弱化；音节；词干提取；弱化元音恢复

Noisy Channel Based Uyghur Harmonized Vowel Identification

ZAOKERE Kadeer, TUERGEN Yibulayin, AISHAN Wumaier
(School of Information Science and Engineering, Xinjiang University, Urumqi, Xinjiang, China, 830046)
Email:Hasan1479@xju.edu.cn

Abstract: This paper presents a noisy channel model based Uyghur harmonized vowel identification method. This method describes the Uyghur vowel harmonization process by using noisy channel model, which means that some of the vowels of the stem would be weakened to neutral vowels when they are transferred in a noisy channel by noise. In this paper, we build language model and channel probability formula on the last two and three letter and last syllable of the harmonized stem on the basis of the Uyghur vowel and consonant harmony and structure of syllable. Experiment results show that the last three letters has a better result to identify the harmonized vowel. After we combined this method with FSM based Uyghur noun stemmer, the precision of the stemmer improved over 15%.

Key Words: Noisy Channel; Uyghur; Vowel Harmony; Syllable; Stemming; Harmonized Vowel Identification

1 引言

黏着语语言是一种有时态变化的语言类型，通过在单词的词尾粘贴不同的词缀来实现语法功能。维吾尔语、日语、韩语、芬兰语、满洲语、蒙古语、土耳其语、匈牙利语、泰米尔语等为典型的黏着语。维吾尔语具有非常丰富的形态结构，维吾尔语单词根据结构可分为词根和词干。在维吾尔语中，词根是最小有义单位，词根不能分解。词干通常由几个词根互相连接或词根与几个构词词缀的互相连接而产生。维吾尔语中，构词词缀用于造词，构形词缀用于表达语法功能。构形词缀通过连接到词干词尾在句子中表达与其他成分的关系。

维吾尔语具有丰富的形态系统，所以在研究和开发维吾尔语自然语言处理系统时必须解决词干还原或词干提取(stemming)。维吾尔语的词干提取同汉语分词一样很重要，是一项基

资助项目：国家自然科学基金(60663006);国家语委科研项目(MZ115-75)；新疆维吾尔自治区高新技术计划项目(200712109);新疆大学校院联合资助项目(XY080124)。

础课题，又是具有挑战性的课题。词干提取的主要任务是把已发生形态变化的单词还原为词干形式，即分开词干和构形词缀。目前很多语言已经实现了可用的词干提取算法。比如，有Malay [2], Latin [3], Indonesian [4], Swedish [5], German [6] and Turkish [7]等。

维吾尔语具有自己独特的元音和谐系统。在学术界，维吾尔语被认为具有与土耳其语非常相似的元音和谐系统，但实际上维吾尔语的元音和谐与土耳其语的元音和谐相差很大[8]。在国际学术界很多专家试图使用土耳其语和芬兰语等语种的元音和谐模型来解释维吾尔语的元音和谐现象，但都失败[9-11]。维吾尔语的形态系统中存在元音弱化、增音和元音脱落等现象，其中增音现象的发生有一定的规律，可用规律进行还原，原因脱落在极少数的词语发生形态变化时发生，完全可以用词典查询方法来解决，但元音的弱化又是非常普遍，又是非常灵活，尤其是外来词发生形态变化时，很难根据上下文恢复弱化的元音，这种现象使得计算机难以正确提取单词词干，导致维吾尔语词性标注、信息检索以及机器翻译等课题的研究中出现严重的数据稀疏。

研究基于有限状态自动机的词干提取算法过程中发现有相当一部分词干的弱化形式不符合维吾尔语语法中所定义的规则。维吾尔语中的外来词具有很复杂的元音弱化现象，很难定义一套互不冲突的或无歧义的规则来确定发生弱化的元音的原元音。由于元音的弱化不仅与连接词尾的词缀的元音有制约关系，还有一定程度上与元音附近的福音也存在一定的搭配关系[8],[12],[13]，本文提出基于噪声信道模型的维吾尔语弱化元音恢复方法，实验结果表明，该方法较好的解决了此问题。

2 维吾尔语元音弱化和央音恢复规则

现代维吾尔语有ا，ە，ې，ى，و，ۇ，ۆ，ۈ 等八个元音和ب,پ,ت,ج,چ,خ,د,ر,ز,ژ,س,ش,غ,ف,ق,ك,گ,ل,م,ن,ھ,ۋ,ي 等24个辅音。维吾尔语元音根据发音特征分为前元音、后元音、央元音、占唇、圆唇、高元音、低元音等。

表1 现代维吾尔语元音分类表

前后 / 唇状 / 舍位	前		央		后	
	占唇	圆唇	占唇	圆唇	占唇	圆唇
高		ۇ	ى			ۈ
次高		ۆ			ې	و
次低	ە					
低					ا	

维吾尔语中音节是自然能感受的最小语音片断，也是语音的基本构成单位。根据统计，绝大部分维吾尔语音节都符合如下音节规则，即： C+V+C+C。其中，字母“V”代表元音，一个音节中有且仅有一个元音，元音是音节的中心；“C”代表辅音，一个音节中可以有0-3个辅音字母。即可能出现如下几种音节结构：V、VC、CV、VCC、CVC、CVCC。除此之外，还包括一些不符合上述音节规则的词，其形式有：C+C+V、C+C+V+C、C+C+V+C+C、C+V+V、C+V+V+C[14]。一个音节以元音结尾，则称为开音节；音节以辅音结尾，则称为闭音节。音节的结构形式规定了在音节层不允许有元音和元音的组合，只允许有元音与辅音或辅音与辅音的组合，因此，元音在音节层只能以单个元音为单位与（的同时与）辅音结合，维语元音与辅音的相互组合较自由，大都数不受一定的规则制约[15]。

现代维语元音和谐规则主要有以下几条：(1) 在同一词中前元音与前元音共现，后元音与后元音共现，前后元音不能共现。(2) 同一词中的元音在前或后特征一致的基础上，圆唇元音与圆唇元音共现,不圆唇元音与不圆唇元音共现，圆唇元音与不圆唇元音不能共现。(3) 同一词中，中性元音既可以与前元音共现又能与后元音共现，既可以与圆唇元音共现，又可以与不圆唇元音共现。在现代维语中词干内不同音节中的元音按元音和谐规则共现，词干后接加词缀时，词缀元音音形的交替同样遵守元音的和谐的规则。维吾尔语的元音恢复规则如下：

检查词干内部发生弱化的ش，ش之后连接的词缀的元音特征。若有前元音或后元音，则根据词缀的元音性质进行还原。比如：(向某的学校)مەكتىپىگە，这个单词中有词缀گە，词缀中存在前元音ە，所以把مەكتىپ还原成مەكتەپ。

如果词缀中找不到前后原因，则根据现代维语元音和谐规则，从词干内寻找前后元音，并根据所找到的元音的性质进行还原。比如，(从妈妈)ئاپىسىدىن，连接这个单词的词缀不包括前后元音，但在词干内部有后元音ا，所以把ئاپى还原成ئاپا。

虽然维吾尔语有着明确的元音和谐规则，但很多外来词并不服从这一规则，使得维吾尔语的元音恢复出现歧异现象。比如，(的)پارتىينىڭ = (的)نىڭ+(党)پارتىيە，(信)خەت，(支票)چېكى，(头)بېشى，(智慧)پاراسەت，(刀)پىچاق，(主人)ئىگىسى，(许可)ئىجازىتى。以上单词中，有些单词的词干内部也没有前后元音，有些单词虽然有前后元音，但不能按照元音性质进行还原。对 5 万个名词进行统计后发现，发生弱化词语占 20%。维吾尔语元音弱化第一规则是完全可靠，但是第二条规则存在较多的歧异现象，使用规则恢复弱化的元音时正确率徘徊在 45%到 60%，因此本文中采用噪声信道模型消除歧义。

3 元音弱化恢复的噪声信道模型

3.1 信道噪声模型

信道是通信系统中信息传输的通道[16]。由于干扰(即噪声) 的存在，输入 I 经信道以一定的概率转换为输出 O。信道的特性可以用条件概率分布 p (O| I) 描述。信道作为通信系统的一部分，其特性必然会影响信息的传输质量。一个好的通信系统必须充分考虑其信道特性，信道特性可以事先统计得到[17]。

从直觉来看，噪声信道模型可以这样理解：表层形式（弱化的词干）可以看成是词汇形式（词汇就是没有连接词缀和没有发生弱化的形式）通过了噪声信道模型之后得到的一个实例。由于在信道中有“噪声”，使得我们难以辨认出词汇形式的“真实”单词面目。我们的目的就是建立一个信道模型，使得能够计算出这个“真实”单词如何被“噪声”改变面目，从而恢复它的原本面目。

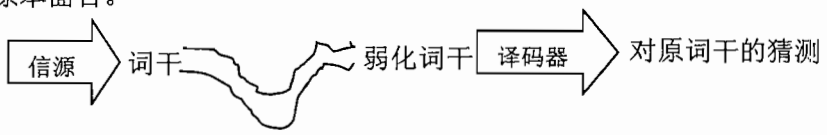


图 1 噪声信道模型

3.2 元音弱化信道噪声模型

维吾尔语的央音恢复是把发生弱化的词干最后一个音节的ش恢复成ش,ە,ا中之一。比如，噪声信道传输过程中发生变化的پارتىيىنى的三个可能候选پارتىسى,پارتىيە,پارتىيا中选出最有可能的那

一个词干。换句话说，在发生弱化的词干 V 的所有可能的词干，我们只想使得 p(词干|发生弱化的词干)为最大的那个词干。我们用 $\hat{s}$ 表示“对正确词干的估计”，用 O 来表示“观察序列”（把每个字母看成一个观察）。因此候选词干中挑选出最优词干的等式为：

$$\hat{S} = \arg\max_{S \in V} P(S \mid O) \tag{1}$$

式(1)能保证挑选出最优的词干。对于给定的弱化词干 S 和观察 O，怎么算出 p(S|O)。根据贝叶斯公式有：

$$\hat{S} = \arg\max_{S \in V} p(S \mid O) = \arg\max_{S \in V} \frac{p(S)p(O \mid S)}{p(O)} = \arg\max_{S \in V} p(S)p(O \mid S) \tag{2}$$

式(2)中有两个概率分布模型需要考虑，一个是 p(S)，称之为语言模型(language model)或先验概率(prior probability)，是指在输入语言中“词”序列的概率分布；另一个是 p(O|S)，称之为信道概率(channel probability)或似然度(likelihood)。

先验概率 p(S)，可以根据单词或弱化音节在语料库中出现的频率次数来计算。但是，似然度 p(O|S)的精确计算至今还是一个未有解决的研究课题。一个词干的弱化与词干元音特征、最后一个音节是否开音节或闭音节、最后一个元音周围的辅音的发音部位和发音方法等因素有一定的关系。我们从 55707 个变形的维吾尔语名词中选出了 10917 个不符合弱化还原规则的弱化词，占 19.60%。根据维吾尔语单词的特征得知，维吾尔语的同一词内存在音节之间的制约关系。据统计结果得知维吾尔语有 2600 多个音节，但在 10917 个单词词尾出现的音节有 402 个。另外，我们从词尾取三个和两个字母进行统计后发现，元音字母和与某些辅音搭配出现，词尾出现的两个字符和三个字符组合分别有 95 种和 625 种。

为了获取最佳语言模型和似然度，我们使用词尾二字符、三字符和音节等三个层面计算建立语言模型和计算似然度。语言模型为：

$$\mathrm{p(s)} = \frac{\mathrm{C}(c) + 0.5}{\mathrm{N} + 0.5\mathrm{V}} \tag{3}$$

其中，C(c)分别为词尾二字符、三字符以及音节在词干语料库中出现次数，N 为语料库单词数目，V 为语料库的词汇量。

似然度 p(O|S)的定义如下：

$$\mathrm{p(O \mid S)} = \frac{\mathrm{Replace(OV, HV)}}{\mathrm{Count(HV)}} \tag{4}$$

其中，Replace(OV,HV)为弱化元音 HV(Harmonized Vowel)在正确词干中应为 OV(Original Vowel)的次数，Count(HV)为 HV 在弱化词干语料库中出现的次数。

4 实验与分析

4.1 实验库结构设计与建设

实验中采用新疆大学多语种信息技术重点实验室正在研制的现代维吾尔语 100 万词语料库的名词库进行训练和测试。我们从 55707 个变形维吾尔语名词进行人工和自动结合的词干提取，从中选出 10917 个不符合弱化还原规则的弱化词作为实验语料使用。

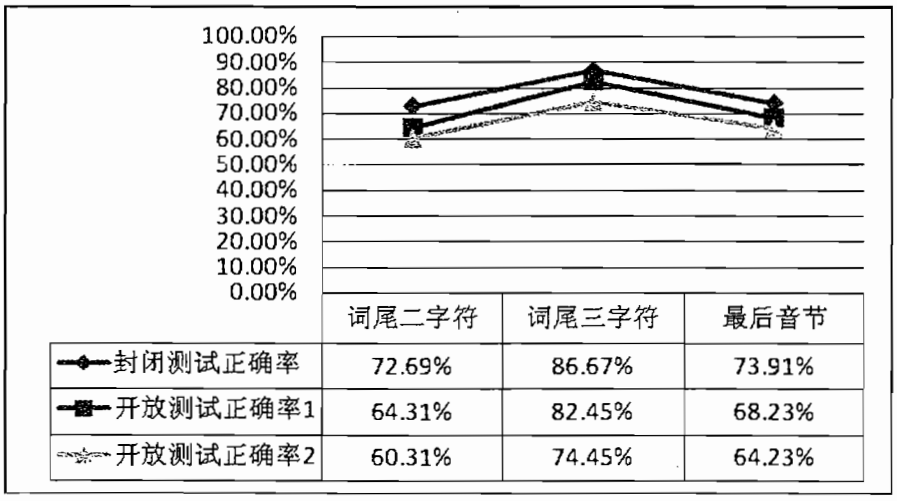
我们开发《有限状态自动机和词典相结合的维吾尔语词干提取和校对软件》，利用该软件自动提取词干和词缀，然后语言学人员进行校对，分析结果包括弱化词干、词干以及连接词缀原形等。完成人工校对后，根据弱化词干、词干以及连接词缀原形提取弱化类型、弱化词干词尾二字符和三字符等生成弱化词语料库。该语料库的结构如表 2 所示。

表2　弱化词实验库实例

单词	弱化词干	弱化词干词尾二字符	弱化词干词尾三字符	弱化词干最后音节	词干	词干词尾二字符	词干词尾三字符	词干最后音节	弱化类型
ۆھىمدىن	ۆھىمى	مى	ىمى	مى	ۆھىمە	مە	ىمە	مە	ە
ئائىلىسىدىن	ئائىلى	لى	ىلى	لى	ئائىلە	لە	ىلە	لە	ە
ئابدىسىنى	ئابدى	دى	ىدى	دى	ئابدە	دە	ىدە	دە	ە
ئاپىسىنى	ئاپى	پى	اپى	پى	ئاپا	پا	اپا	پا	ا
ئاپتېپىدا	ئاپتېپ	پ	تېپ	تېپ	ئاپتاپ	اپ	تاپ	تاپ	ا
ئاپىتىنىڭ	ئاپىت	ت	پىت	پىت	ئاپەت	ەت	پەت	پەت	ە
قوشنىسىغا	قوشنى	نى	شنى	نى	قوشنا	نا	شنا	نا	ا
ئادىتىنىڭ	ئادىت	ت	دىت	دىت	ئادەت	ەت	دەت	دەت	ە
ئادىمىگە	ئادىم	م	دىم	دىم	ئادەم	مە	دەم	دەم	ە

4.2 实验

我们利用公式 (3)训练三种语言模型，利用 (4)计算似然度或信道概率，最后使用公式(2)选出具有最高概率值弱化词干原形或弱化类型。为了观察该方法和三种模型的鲁棒性，分别设计封闭测试和开放测试。为了估计语料规模的影响，设计两次开放测试，开放测试 1 在80%的语料库上进行训练，然后对 20%的语料库进行测试；开放测试 2 在 60%的语料库上进行训练，然后对 40%的语料库进行测试。因为本文提出的方法只需要对每一个弱化词干词尾包含的央音进行弱化元音识别，所以此方法性能以准确率为标准进行测试。准确率为正确识别的弱化元音词干数对人工识别的弱化元音词干的比例，实验结果如图 2 所示。



	词尾二字符	词尾三字符	最后音节
封闭测试正确率	72.69%	86.67%	73.91%
开放测试正确率1	64.31%	82.45%	68.23%
开放测试正确率2	60.31%	74.45%	64.23%

图 2　封闭和开放测试结果

4.3 分析

从直觉或语言特点来看，使用维吾尔语最小语音单位音节进行的测试结果应该为最佳，

但实际实验结果表明，使用词干词尾三字符进行训练模型时的弱化元音识别性能最佳。根据实验结果，基于三种统计方法的性能排序为：词尾三字符、词尾音节、词尾二字符。从这个排序可得，基于音节的方法的性能处于其他两种模型的中间，主要元音在于音节是可变长的单位，音节长度可为 1 到 4，而基于三个字符的方法具有较高的准确率是因为三个字符具有较强的约束能力，在有些情况下能利用词尾音节前一个音节提供的信息。另外，从两次开放测试实验结果可得，语料库规模对该方法的鲁棒性具有较大的影响。经分析错误处理的词干发现，导致错误的原因有拼写错误占 9%，元音脱落占 12%，方法错误处理占 79%。该方法错误处理出现的主要词语为从阿拉伯语接受的外来词，这些单词的结构与从其它语言接受的外来词又有较大的区别。

5 结论

元音弱化、脱落和增音等现象中元音弱化恢复是维吾尔语词干提取系统必须要克服的难点之一。我们在有限状态自动机的基础上开发的维吾尔语名词词干提取算法由于未能有效处理不符合规则的弱化词干准确率降为 80%左右，其中导致错误提取的主要原因为不符合规则的弱化词。为了解决此问题，我们根据维吾尔语的特点，提出了基于词干词尾二字符、三字符和音节的信道噪声的弱化元音恢复方法，并进行了不同规模的封闭和开放测试。该方法的准确率虽然未能达到 90%以上的高准确率，但是在一定程度上较好地处理了弱化词，结合基于有限状态自动机的词干提取方法后，词干提取准确率从 80.04%提高到 96.91%。

参 考 文 献

[1] Kernighan,M.D., Church,K.W., and Gale,W.A.(1990). A spelling correction program based on an noisy channel model. In COLING-90, Helsinki, Vol II ,pp.205-211.

[2] Tai, S. Y., Ong, C. S., and Abdullah, N. A, "On designing an automated Malaysian stemmer for the Malay language(poster)"

In Proceedings of the fifth international workshop on information retrieval with Asian languages, Hong Kong, 2000. pp. 207-208.

[3] Greengrass, M., Robertson, A. M., Robyn, S., and Willett, P, Processing morphological variants in searches of Latin text, Information research news, 1996, 6 (4), pp. 2-5.

[4]Berlian, V., Vega, S. N., and Bressan, S, "Indexing the Indonesian web: Language identification and miscellaneous issues", Presented at Tenth International World Wide Web Conference, Hong Kong, 2001.

[5]Carlberger, J., Dalianis, H., Hassel, M., and Knutsson, O, "Improving precision in information retrieval for Swedish using stemming", In Proceedings of NODALIDA '01 - 13th Nordic conference on computational linguistics, Uppsala,Sweden, 2001.

[6]Monz, C. and de Rijke, M. "Shallow morphological analysis in monolingual information retrieval for German and Italian." In Cross-language information retrieval and evaluation: Proceedings of the CLEF 2001 workshop, C. Peters, Ed.: Springer Verlag, 2001.

[7] G. Eryiğit & E. Adalı, "An Affix Stripping Morphological Analyzer for Turkish", Proceedings of the IASTED International Conference ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 2004, Innsbruck, Austria.

[8] Disharmony and derived transparency in Uyghur Vowel Harmony, Bert Vaux,Harvard University,January 2001

[9] Clements, G. Nick. 1976. The autosegmental treatment of vowel harmony. In W. Dressler and O. Pfeiffer, eds., Phonologica 1976. Innsbruck.

[10] Clements, G. Nick. 1987. Toward a substantive theory of feature specification. NELS 18, 79-93.

[11]Clements, G. Nick and Engin Sezer. 1982. Vowel and consonant disharmony in Turkish. In vander Hulst and Smith, eds., The Structure of Phonological Representations II. Dordrecht: Foris, 213-255.

[12] Alling, Emily. 1999. Uyghur vowel harmony. Manuscript, Harvard University

[13] Lindblad, Vern. 1990. Neutralization in Uyghur. MA thesis, University of Washington.

[14] 维吾尔语文语转换系统文本分析模块初探 马欢，吾守尔·斯拉木 2006 年 8 月 计算机工程 第 16 期 32 卷 267-268

[15] 现代维吾尔语元音格局分析 易斌 2006 年 1 月新疆大学学报(哲学·人文社会科学版)，2006 第 34 卷第 1 期 140-144p。

[16]傅祖云. 信息论基础[M]. 北京: 电子工业出版社, 1989.

[17]一种基于噪声信道模型的汉字识别后处理新方法 清华大学学报(自然科学版) 2001 年第 41 卷第 1 期 24-28

[18]哈密提·铁木尔 现代维吾尔语语法[M] 北京：民族出版社，1987.

[19] 阿依克孜·卡德尔. 开沙尔·卡德尔. 吐尔根·依布拉音. 面向自然语言信息处理的维吾尔语名词形态分析研究[J].中文信息学报,2006,(3):43-48.

[20] 力提甫·托乎提.电脑处理维吾尔语语音和谐律的可能性[A].中央民族大学学报，2004,(5):108-113.