中国科学: 信息科学 2015 年 第 45 卷 第 10 期: 1328–1340
www.scichina.com info.scichina.com
《中国科学》杂志社
SCIENCE CHINA PRESS
面向感知中国的新一代信息技术专刊 · 论文

面向双语教学的维吾尔语发音质量自动评估

董滨[①], 丁铭[①*], 王磊[②], 颜永红[①]

① 中国科学院声学研究所语言声学与内容理解重点实验室, 北京 100190
② 中国科学院新疆理化技术研究所, 乌鲁木齐 830011
* 通信作者. E-mail: dingming@hccl.ioa.ac.cn

收稿日期: 2015–06–09; 接受日期: 2015–08–06; 网络出版日期: 2015–09–16
新疆维吾尔自治区科技重大专项 (批准号: 201230118-3) 资助项目

摘要 推行新疆地区的双语教学是国家推进民族交流的重要举措, 对维语进行自动发音质量评估能够大大提高双语教学的效率. 然而维吾尔语作为黏着语, 其特殊的构词方法造成大量无法被收入词表的集外词存在, 使得基于传统语音识别系统的发音评估方法难以应用. 为了实现高效的维语发音质量评估, 本文在研究分析了维语的发音规则和发音习惯后, 决定采用子词作为基本的识别单元; 在原有发音质量自动评估系统上, 改用基于双层词法分析的维语子词切分器; 同时引入音素解码器计算后验概率的分母, 让系统可以在子词级别直接计算置信度. 经过实验数据的验证, 基于子词的维语评估系统的性能要优于基于音素的系统.

关键词 双语教学 黏着语 发音评估 子词 后验概率 置信度

1 引言

语言是人类学习、工作、生活和社会交往的重要工具, 民族之间要想有效融合, 必须从语言开始. 新疆少数民族使用的语言以维吾尔语为主, 属于阿尔泰语系突厥语族. 维吾尔语在新疆不仅是人们的交际工具, 也是民族间相互理解、互相帮助、共同进步的基础, 各民族相互学习语言在推动自治区的经济发展与社会进步方面具有重要的意义. 要上传下达路线、方针、政策, 做好基层各项繁杂工作, 维护新疆地区的经济繁荣和社会稳定, 就必须重视维吾尔语的学习和使用. 国家一直以来就很重视新疆的民族交流和融合, 特别是汉族和维吾尔族的相互理解和团结友爱, 因此在 20 世纪 50 年代就开始大力宣传和鼓励汉族干部学习维吾尔语 [1].

双语教学作为一种高效的语言学习方式, 在我国的中高级外语教育中已得到了广泛的应用, 并取得了不错的反响和效果. 因此将这种先进的教学方式运用到新疆干群的维汉语教学当中也是水到渠成的事情. 但是, 传统的双语教学模式却难以满足现阶段大规模推广维汉双语教学的需求. 因为如今的双语教学早就将重点由语法词汇的学习转向了互动交际能力的培养, 越来越强调语言发音的学习和训练, 传统的语言教学模式需要教师特别专注于学习者的学习, 这就要求语言教师拥有足够的耐心和大量的时间, 当学习者增多, 相应的教师数量也必须跟上, 要想维汉语的双语教学能够推广, 就需要大量具有双语能力的合格教师, 然而, 维汉双语人才的缺失是现阶段自治区和国家在推广双语教学政策时都不得不面对的困难. 引入计算机来辅助双语教学成为了解决上述困难的一条可行之计. 随着计算机

引用格式: 董滨, 丁铭, 王磊, 等. 面向双语教学的维吾尔语发音质量自动评估. 中国科学: 信息科学, 2015, 45: 1328–1340, doi: 10.1360/N112014-00327

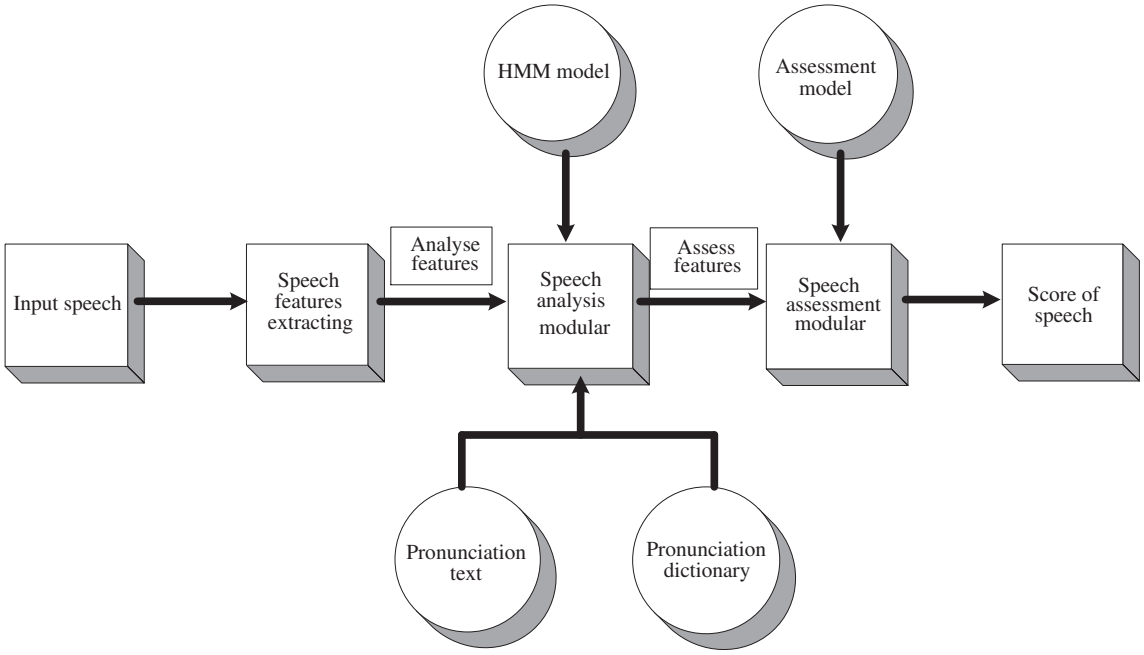


图 1 基于 HMM 的汉语普通话水平客观测试系统的框图

Figure 1 The structure chart of the HMM-based mandarin pronunciation level assessment system

技术的发展和语音处理技术的不断完善, 计算机辅助语言学习和测试已经成为可能, 基于计算机的语言学习系统已经从最初的只能进行阅读、听力和简单的输入等功能发展到了更高的阶段. 若运用得当, 计算机辅助语言学习系统能够大大减少教师所需付出的教学时间, 也能够大大增加一个教师能够应付的学生数量, 并且能够让学习者在脱离教师的环境中进行发音的学习和矫正. 而且计算机系统能够给出更为详细的学习者发音参数, 让传统的被动模仿式学习通过学习者得到的信息反馈和建议变为主动的自我调整式学习. 为了能够实现计算机辅助维汉双语教学, 就必须研究维语和汉语的发音质量客观评估算法, 因为发音评估是计算机辅助语言学习核心 [2].

自 20 世纪 90 年代以来, 国内外很多研究单位的学者在自动语音识别基础上对发音质量客观评估算法进行了研究. 美国 Stanford 大学 [3]、英国 Cambridge 大学 [4] 等机构的研究人员分别在带有各种口音的英语、法语、西班牙语和日语等语言上进行过研究, 国内的清华大学 [5]、中国科学院声学研究所 [2,6] 等单位在带有口音的汉语和英语上进行了研究. 经过这些单位 20 多年的研究发展, 目前形成了一套采用基于 HMM 自动语音识别技术的自动发音质量评估方法. 如图 1, 基于 HMM 语音识别技术的发音质量客观评测方法分为 3 个主要的部分: 语音特征提取模块、语音分析模块和语音评估模块. 在语音特征提取模块中, 对输入语音即测试者的发音进行预处理, 并提取语音的分析特征; 在语音分析模块中, 借助离线训练的 HMM 模型、发音文本和发音词典的帮助, 计算测试者语音与正确语音之间的接近程度, 作为语音的评估特征; 并在语音评估模块中评估模型的帮助下得到最后输入语音的发音质量得分.

在上述系统中实现维语的发音评估, 需要有两方面的工作, 一是黏着语特性的研究, 另一个是黏着语的语音识别系统搭建.

黏着语通过词的形态变化表示语法意义, 这造成相比于非黏着语它具有大得多的常用词库, 像土耳其语、芬兰语、匈牙利语、维吾尔语都是典型的黏着语. 早在 2000 年德国 Karlsruhe 大学就进行了

土耳其语的自动语音识别研究 [7]. Helsinki 大学的研究者之后提出了分解、统计子词的概念 [8], 对解决黏着语集外词过多的问题具有重大意义. 后来的研究者则通过合并相邻子词 [9], 建立混合词典 [10], 使用高阶语言模型 [11] 等方法, 解决了子词语言模型上下文较短的问题.

在维吾尔语方面, 新疆大学 [12] 结合维吾尔语黏着性的特点, 建立了维吾尔语连续语音语料库, 实现了基于隐含 Markov 模型 (HMM) 的维吾尔语连续语音识别系统. 京都大学则在他们的维语识别系统中进行了语速识别单元 [13] 和混合词典 [14] 的研究. 中国科学院声学研究所的李鑫等 [15] 为了解决维吾尔语语音识别系统集外词过多的问题, 采用统计子词代替词语作为识别系统的词典单元, 并提出了独特的子词切分方法. 综上来看, 目前尚未见有针对维吾尔语的发音评估系统的论述. 维吾尔语和汉语、英语等常见语种不同, 维吾尔语发音时由若干音素拼接而成, 在元音和谐、辅音结合等方面有自己独特的规律. 因此, 有必要专门针对维吾尔语的发音特点, 结合已有的发音评估技术的研究成果, 进行维吾尔语发音质量自动评估技术的研究.

本文首先介绍当前已有的发音评估技术的基本原理和方法, 然后通过分析维吾尔语独特的黏着特性, 进而找到了解决维语发音质量评估难题的有效方法: 即先对维吾尔语进行自动子词切分; 然后采用基于子词的发音质量置信度测度算法得到相应的评分特征; 最后将评分特征输入评分模型得到最终结果. 实验的结果表明了这套方法对评估性能起到了提升作用.

2 维语子词的机器切分

2.1 维吾尔语的黏着特性

维吾尔语是一种黏着语, 可以通过不断在词干后结合附加成分构成新的词语. 词干和附加成分统称为语素. 附加成分按其作用可以分为构词附加成分、构形附加成分和构词 — 构形附加成分 3 类 [16].

(1) 构词附加成分表示词汇意义, 结合在词干后能构成新词. 例如,

Shinjang (新疆)+ liq → shinjangliq (新疆人),

Uyghur (维吾尔族)+ ce → uyghurce (维吾尔语),

Xenzu (汉族)+ ce → xenzuce (汉语),

Weten (爱国)+perwer → wetenperwer (爱国的).

(2) 构形附加成分只表示纯粹的语法意义, 结合在词干后构成一个词的不同形态. 构形附加成分结合在名词词干后可以表示数、领属人称和格的语法意义, 结合在形容词词干后可以表示级的语法意义, 结合在动词词干后可以表示式、体、时和人称的语法意义. 例如,

Ishchi (工人)+ lar → ishchi (工人们),

Ishchi (工人)+ lar + miz → ishchilarmiz (我们的工人们),

Ishchi (工人)+ lar + miz + ning → ishchilarmizning (我们的工人们的).

(3) 构词 — 构形附加成分加在词干的后边不能改变词的词汇意义, 但可以改变词性和词在词组或句子中的功能. 例如,

K ü l (笑, 动词)+ ü sh → k ü l ü sh (笑, 名词),

K ü l (笑, 动词)+ ü watqan → k ü l ü watqan (正在笑的, 形容词),

K ü l (笑, 动词)+ ü p → k ü l ü p (笑, 副词).

词干结合构词附加成分形成的词语类似于汉语或英语中的词语, 而结合构形附加成分形成的词语则对应于汉语或英语中的词组. 构形附加成分的存在是维吾尔语中出现大量不同词形的原因, 而这些

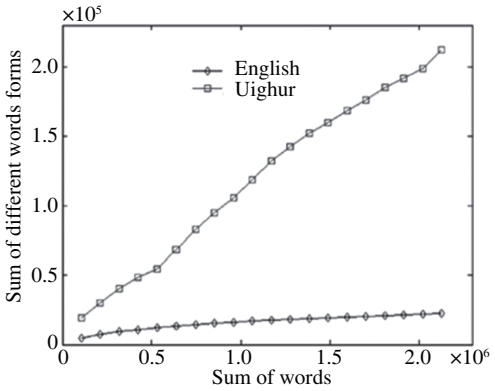


图 2 维语和英语电话谈话文本中不同词形总数比较

Figure 2 Comparison of different word forms' sum in Uighur and English telephone talking text. @Copyright 2013 The Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences

构形成分是不写入维语词典的, 这使得想要在维语自动识别系统的词典中缩减集外词变得几乎不可能. 中国科学院声学研究所的李鑫等统计了不同规模的维吾尔语和英语电话谈话文本中出现的词形总数, 得到的曲线如图 2 所示. 从图 2 中可以看出, 随着语料规模的扩大, 维吾尔语文本中不同词形数目的增长速度明显超过英语. 当文本规模达到 2.13 M 词语时, 维吾尔语文本中包含的不同词形有 212.3 K, 远大于英语的 22.4 K [15]. 在维吾尔语中, 表示同一语法意义的构形附加成分一般具有多种变体. 在词干结合附加成分的过程中, 变体使用要遵循的规则包括元音和谐、辅音和谐和元音弱化. 这些拼写规则使得词干结合附加成分时需要考虑连接边界的发音特点, 增加了形态分析的难度 [15].

2.2 基于双层词法分析的子词切分

本文将维语的每个单词都拆成一个或多个子词的形式, 作为本文中各种评分算法的前提. 为了将维语词汇有效地切分成子词, 我们设计了基于有限状态转录机 (FST) 的双层词法分析算法.

双层词法分析基于语言学上的双层形态学, 这是一种非常适合于黏着语分析的形态分析技术 [17]. 在这项技术中, 一个词语通常被表述为词汇层 (lexical level) 和表层 (surface level) 之间的映射关系. 所谓词汇层是表示组成该词的词干和附加成分的语法含义, 表层则为词语的正常拼写方式. 他们之间通过中间层来描述词干添加附加成分后的动态拼写变化. 以单词 foxes 为例, 它是 fox 的复数形式, 那么在双层词法分析中可以表述为图 3 的样式, 示例中的 +N+PL 就代表该词为复数形式的语法意义, 中间层中的符号 ^则标示了词干和附加成分之间的边界. 为了完成各层之间的转化就需要用 FST.

为了说明有限状态转录机 FST 方法, 需要首先介绍有限状态自动机 (FSM). 有限状态自动机是具有离散输入和输出系统的一种数学模型. 它有有限个内部状态, 随着信号的输入, 内部状态不断地转移. 有限状态自动机可以定义为一个有序五元组 $M = \langle Q, \Sigma, \delta, q_0, F\rangle$, 其中

(1) Q 是非空有限的状态集合;

(2) Σ 是非空有限的输入字母表;

(3) $\delta : Q \times \Sigma \to Q$ 是状态转移函数;

(4) q_0 是初始状态;

(5) F 是终结状态集.

有限状态自动机可以用状态图表示. 状态转移图是一个有向图, 每个结点代表一个状态. 初始状

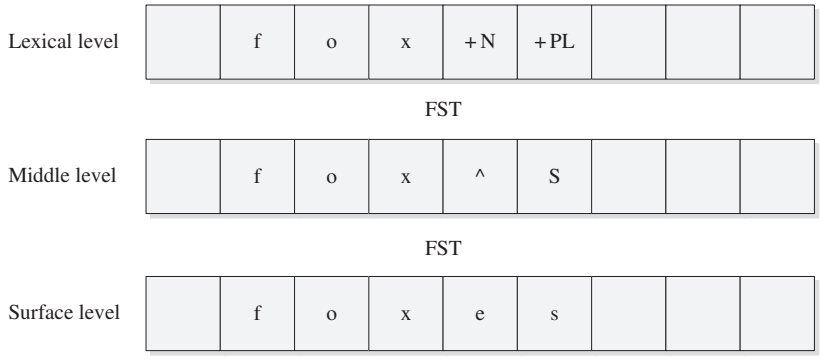


图 3　foxes 的双层词法分析示例

Figure 3　A sample for double-level morphology analysis

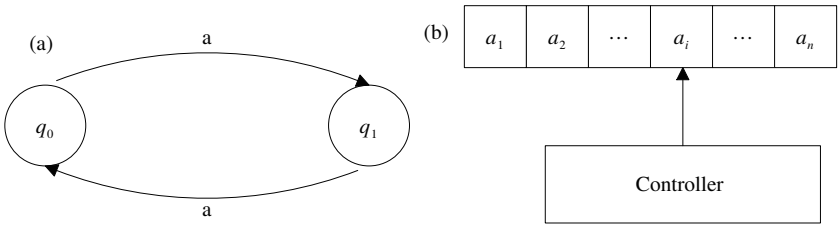


图 4　有限状态自动机

Figure 4　Finite state automata machine. (a) State transition diagram; (b) automata machine operating principle

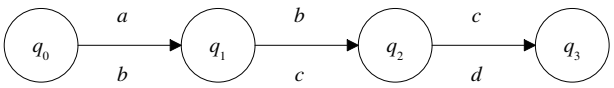


图 5　有限状态转录机示例

Figure 5　A sample for finite state transducer

态用一个指向该节点的箭头标明, 终结状态用双圈标明. 如图 4 所示, 可以把有穷自动机看作一个具有有限个状态的控制器, 它有一个读头. 控制器对给定的输入序列 s 的工作方式如下: 开始时, 控制器处于初始状态 q_0, 读头扫描第一个元素, 在每一步, 控制器根据当时的状态 q 和扫视的元素 a, 把它的状态转移到 $\delta(q,a)$, 同时读头向右移动一步, 如此一步一步地进行, 直到扫描完 s 的所有元素为止. 如图 4(b) 所示, 对于一个输入的序列, 自动机逐个扫描输入元素, 并且其状态不断转换, 如果最终自动机进入终结状态, 并且输入序列扫描结束, 则称自动机接受该序列或识别该序列, 否则称拒绝该序列或无法识别该序列.

有限状态转录机是有限状态自动机的扩展. FST 和 FSA 基本相同, 包括状态和状态转移函数, 不同之处在于, 在状态转换弧上标有一对符号, 其中一个符号是输入符号, 另外一个符号是输出符号, 这样, 当 FST 按照自动机的机制进行字符串的识别时, 除了进行状态之间的转换外, 还输出一个符号序列, 从而实现了输入符号到输出符号的一种映射. 以图 5 为例, 其中初始状态为 q_0, 结束状态为 q_3, 这样, 当输入一个符号串 abc 的时候, FST 接受该符号串, 并输出一个符号串 bcd. 基于有限状态技术的双层词法分析模型的基本思想是根据其词法规律建立词法规则库, 词法分析程序依据规则库中的规则对输入进行还原处理, 并形成相应的词法特征信息, 即生成一个含有词干和其形态特征的中间层, 以

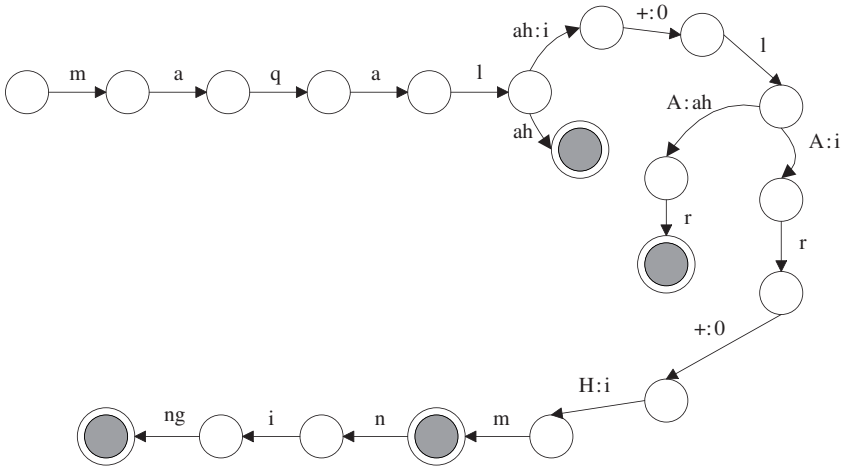


图 6 基于有限状态转录机的维语子词切分

Figure 6 FST-based Uygur subword segmentation

供后续处理机制使用. 这些词法规则采用正则表达式来描述, 有限状态自动机是作为正则表达式的一种识别装置, 采用这种词法分析方法, 可以把语言中的形态变化规律作为系统数据进行处理, 使得词法分析系统具有良好的可维护性和可修改性, 而且实现了分析功能和生成功能. 双层词法分析模型把一个词表示为词汇层和表层之间的对应, 词汇层表示组成该词的子词之间的毗连关系, 表层表示该词实际拼写的最终状况. 形态剖析要建立映射规则, 把在表层上的字母列映射为词汇层上的语素和特征的序列. 转录机实现两个符号集合之间的映射, 有限状态转录机通过有限自动机来实现这种转录.

图 6 展示了用基于有限状态技术的双层词法分析模型实现的维语子词切分的例子. 图中, 每条弧都用拉丁记法标记了表层和词汇层, 表层和词汇层用冒号隔开, 没有冒号的表示表层和词汇层记法一致. 比如, “m” 表示表层和词汇层都是 “m”, 而 “ah:i” 表示表层为 ah, 词汇层为 i. 图中的 “+:0” 符号用于表示两个子词之间的分隔符. 通过将维语单词和该模型相匹配, 即可根据 “+:0” 符号的位置得出子词切分信息. 为了构造图 6 所示的有限状态转录机, 我们总结了 3 部分知识:

(1) 词干词缀词典. 建立包含约 100000 条词干和 200 条常用词缀的词表.

(2) 不同种类语素的连接顺序, 例如, 对于维语名词, 词缀连接顺序为: 词干 [数][人称][格].

(3) 语素连接时书写的变化规则. 例如, 当附加词缀后词干最后一个音节是开音节时, 音节中的 a 或 ε 弱化成 i 或 e.

使用这 3 部分知识, 构建相应的正则表达式, 再将正则表达式编译成有限状态转录机, 从而实现了维语的子词切分.

3 子词级发音置信度

3.1 音素级发音置信度的计算

所谓的发音置信度又可以称为发音质量测度, 能够量化地描述说话人发音质量好坏的物理量都可以叫发音置信度. 而要在音素级别量化地表征发音质量, 音素后验概率则是一个好的选择. 音素后验概率描述待测音素语音段和声学模型之间的相似程度. 在基于 HMM 的语音识别中, 声学模型一般由大量的标准语音数据训练而成, 其具有给定声学层面发音标准的作用, 也就是说被测试的音素与声学模

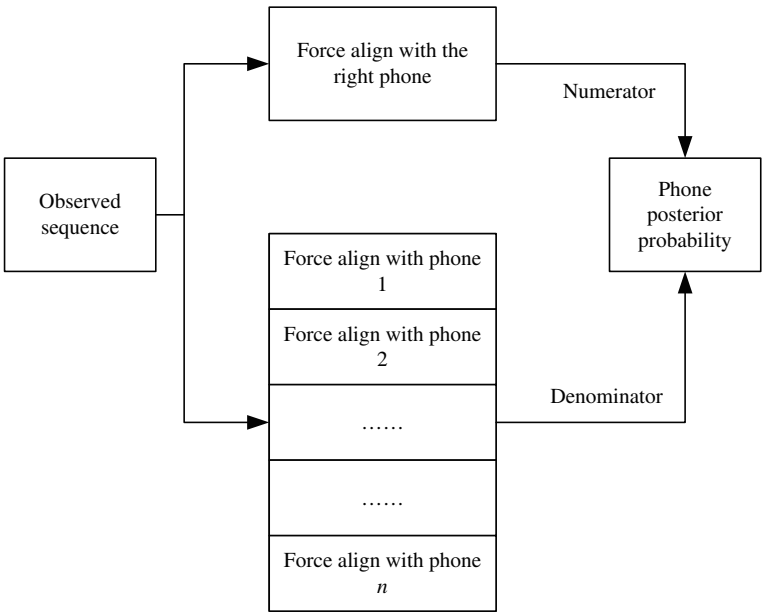


图 7　音素后验概率的计算

Figure 7　Calculation of the phone posterior probability

型的相似度越高那么受测者的发音就越好. 单个音素的相似度可以通过音素解码时的声学模型概率得到, 但是当一个音素放在说话人的整段语音中考虑时就得计算音素的后验概率了. 在给定观察序列 O 的情况下, 用段长进行归一化后的正确音素 q_1 的后验概率可以表述为

$$\mathrm{Poster}(q_i) = |\log(P(q_i|O))| \,/NF(O) = \left|\log\left(\frac{p(O|q_i)p(q_i)}{\sum_{j=1}^{J} p(O|q_j)p(q_j)}\right)\right| /NF(O), \tag{1}$$

其中, $P(O|q_i)$ 为在音素 q_i 下观察到 O 的概率; $p(q_i)$ 为 q_i 的先验概率; J 为系统所有音素的个数; $NF(O)$ 为观察序列的段长. 在实践中, 正确音素的先验概率要远远出现大于其他不正确音素的概率和, 所以可以假设 $p(q_i) = p(q_j)$, 则上式可推导为

$$\mathrm{Poster}(q_i) = \left|\log\left(\frac{p(O|q_i)}{\sum_{j=1}^{J} p(O|q_j)}\right)\right| /NF(O). \tag{2}$$

上式可以直观地理解为, 音素后验概率的分子为观察序列在正确发音音素的解码累积概率, 分母为观察序列在全部发音音素上的累积概率之对数和, 如图 7 所示.

音素后验概率的应用在汉语等语言的发音质量评估上取得了不错的效果, 但是通过上一章的分析得知这种方法直接应用于维语效果并不理想.

3.2　子词级的发音置信度计算

解码阶段采用的维特比算法的局限性, 使音素级的发音置信度不足以准确量化描述维语单词和整句的发音质量. 维特比搜索算法的准则是令路径的累积概率最高, 它从语音段整体最优的角度为观察序列提供了最佳匹配的状态转移序列. 但是, 在音素级别计算观察概率, 强制对齐的观察序列只被限

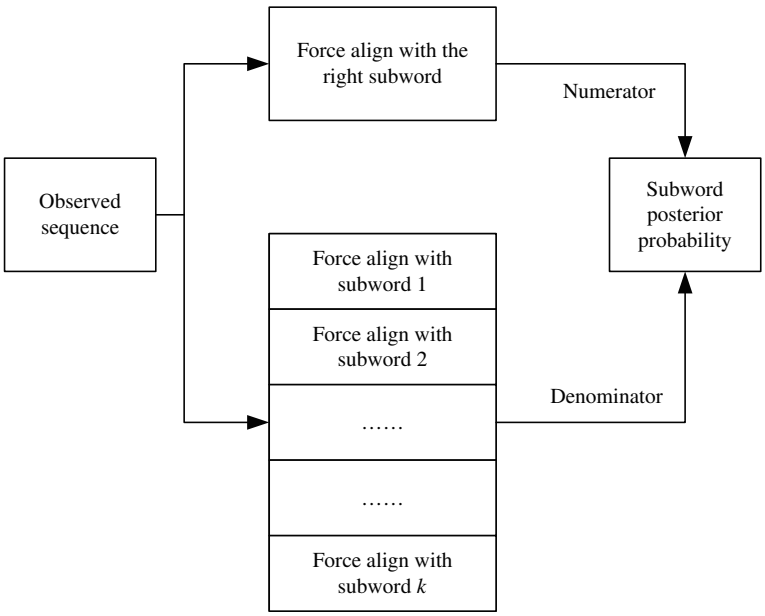


图 8 子词后验概率的计算

Figure 8 Calculation of the subword posterior probability

于当前音素的语音段, 维特比搜索仅达到了在该局部语音段内最优, 而没有达到整个单词以及句子上的最优. 因此, 如果仅在音素级别考察观察序列与声学模型是否匹配, 就可能使待测样本整体上的发音质量的评价有所偏差.

因此, 理想的后验概率计算方式应该是在整个样本上计算, 这样才能达到全局最优. 但这需要在全部的发音空间上计算后验概率的分母, 而这是不可能实现的. 由于维语各子词的发音有着相对独立性, 因此使用子词为单位计算后验概率.

子词级后验概率的计算思路与音素级后验概率相似, 理论上应为

$$\mathrm{Poster}(w_i) = \left| \log \left(\frac{p(O|w_k)}{\sum_{k=1}^{K} p(O|w_k)} \right) \right| / NF(O), \tag{3}$$

其中, $P(O|w_i)$ 为在子词 w_i 下观察到 O 的概率; K 为系统所有音素的个数; $NF(O)$ 为观察序列的段长. 和音素的后验概率类似, 上式可以用图 8 来直观地理解. 但是, 维语子词的数量要远远大于维语音素的数量. 维语中共有 32 个音素, 因此做 32 次强制对齐即可完成音素后验概率的计算. 而维语的子词约有 10 万个, 做 10 万次强制对齐是不可接受的. 为了解决这个问题, 我们希望避免公式分母的大量运算.

由于在和各子词强制对齐时, 得到最大累积概率的子词的对数值远大于其他子词, 因此, 可以使用如下的公式计算子词后验概率:

$$\mathrm{Poster}(w_i) = \left| \log \left(\frac{p(O|w_k)}{\sum_{k=1}^{K} p(O|w_k)} \right) \right| / NF(O) \approx \left| \log \frac{p(O|w_k)}{\max P(O|w_k)} \right| / NF(O). \tag{4}$$

要实现该式的计算, 我们使用循环音素解码器来求得 $\max P(O|w_k)$. 由于维特比解码的本身就是一个求累积概率最大值的过程, 因此一次解码后就可以直接读出 $\max P(O|w_k)$. 在使用循环音素解码

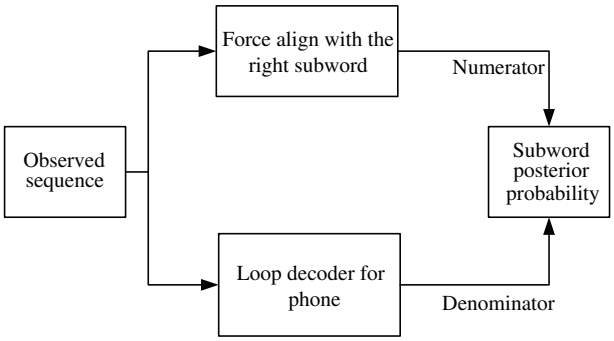


图 9　子词后验概率的快速计算

Figure 9　The fast calculation of the subword posterior probability

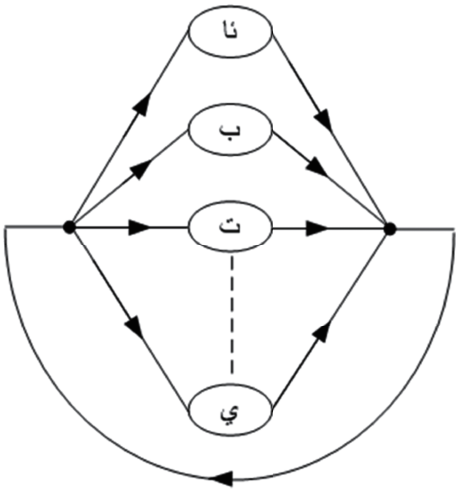


图 10　循环音素解码器

Figure 10　Loop decoder for phone

器的维特比解码过程中, 剪枝机制会不断地避免不需要的计算, 这样就大大降低了计算复杂性, 使计算时间远小于对所有子词依次做强制对齐所需要的计算时间. 这样就可以用一次解码代替约 10 万次解码, 从而使子词级置信度的计算成为可能, 如图 9 所示. 上述算法中的循环音素解码器, 如图 10 所示. 该网络是可重入的, 可以不限次数的循环, 因此可模拟全部发音空间上的样本.

4　实验结果与分析

4.1　系统评价方法

为了对发音质量评估系统进行评价和测试, 我们设计并建立了评价数据库. 评价数据库包括 1000 个不同测试者的录制样本, 测试者的母语为汉语, 普通话标准. 每人朗读一句包括 4 至 5 个词的维语短句. 所有数据均为语音数据, 使用 16000 Hz 采样率, 16 bit 的 wav 格式存储. 评价数据库被分为两部分, 其中 80% 作为试验的训练集, 另外 20% 作为试验的测试集. 训练集和测试集中的男女比例均为 1:1. 为了评价系统性能, 需要人工对试验数据进行评分. 应该注意的是, 由于发音质量评估的主观

性, 不同人对同一批数据的评估打分结果会有差异, 甚至同一个人在不同时间或不同环境下对同样数据的评估结果也会有所不同. 因此, 为了使评估的结果尽量准确客观, 往往同时请数个专家对被测数据进行评估, 然后综合他们的得分作为人工评分的最终结果. 本文聘请了两位维吾尔语专家, 对评价数据库中的每个单词进行了人工打分, 取其平均分作为最终评分. 评测专家对评价数据库的评分是相互独立的, 以尽量做到客观公正. 打分采用三分制, 即把评价数据按照发音质量分为错误、缺陷和正确 3 个等级, 并对应相应的分数.

另外前述的发音置信度虽然是一种量化的特征, 但其存在不直观的缺点, 对应错误、缺陷、正确 3 个评分等级时并不能直接应用. 因此我们需要把它转化为具体的等级分数. 从机器评分到发音质量等级的转换是用事先设定的阈值作线性映射得到的. 那么要完成三分制的发音评价, 就需要两个阈值完成映射:

$$S_{\text{phone}} = \begin{cases} 0, & P < \text{Threshold}_{\min}, \\ 1, & \text{Threshold}_{\min} \leqslant P \leqslant \text{Threshold}_{\max}, \\ 2, & P > \text{Threshold}_{\max}. \end{cases} \tag{5}$$

式 (5) 中, P 表示机器评分; S_{phone} 表示发音质量等级得分; $\text{Threshold}_{\max}$ 和 $\text{Threshold}_{\min}$ 分别是一大一小两个映射阈值. 映射阈值在训练集上根据机器评分和人工打分的结果训练得到, 用阈值把机器评分映射为发音质量等级后与人工打分的结果进行比较, 根据两者的差异对阈值进行调整, 取 “人 — 机” 评估相关系数最大时的阈值为训练结果.

而对于系统性能的评价, 我们采用了两种系统评价参数, 分别是相关系数以及分差. 相关系数是考察两列元素一致性的良好测度, 用这种方法测量人工打分和机器打分之间相似程度, 根据这个相似程度对机器打分的准确性做出评价的方法就是相关系数评价方法. 假设 $F = f_i$ 和 $G = g_i$ 分别是对同一发音样本序列的两种不同的评分结果矢量, 则它们的相关系数定义为

$$R = \frac{\langle \overline{F}, \overline{G} \rangle}{\|\overline{F}\| \, \|\overline{G}\|} = \frac{\sum f_i g_i}{\sqrt{\sum f_i^2} \sqrt{\sum g_i^2}}, \tag{6}$$

本文使用的相关系数指标是先在每个测试者 k 的所有发音样本上计算相关系数 R_k, 再计算所有测试者的相关系数的均值:

$$R = \frac{1}{M} \sum_{k=1}^{M} R_k. \tag{7}$$

其中 M 为测试者的人数.

相关系数法是评价发音质量评估系统的一种通用研究方法, 但是这种评价方法的指标不直观. 对于一般用户来说, 更直观的方法是考察人工打分和机器打分之间的差值, 差值越小表明机器打分和人工打分越接近, 也就表明系统的评估准确性越高, 这就是分差评价方法. 本文在每个被测者 k 的所有音节得分之和的基础上计算分差:

$$\text{diff}_k = \frac{\left| \sum_{i=1}^{N} h_{k,i} - \sum_{i=1}^{N} m_{k,i} \right|}{\text{norm}}, \tag{8}$$

其中, N 为测试者所读词个数; $h_{k,i}$ 是人工评分; $m_{k,i}$ 是机器评分; norm 为归一化因子, 可根据 N 的大小进行调整, 计算得到的分差可以看作是分数差异的百分比. 计算所有被测者分差的均值作为衡量

表 1 基于音素和基于子词的评分性能比较

Table 1 Comparison of the performance between phone-based and subword-based on pronunciation assessment

	Correlation coefficient	Score deviation
Phone posterior probability	0.763	0.579
Subword posterior probability	0.772	0.571
Merge two posterior probability	0.785	0.563

系统性能的指标:

$$\mathrm{diff} = \frac{1}{M}\sum_{k=1}^{M}\mathrm{diff}_k. \tag{9}$$

其中, M 为被测者人数.

4.2 实验结果

分别单独使用基于音素后验概率的发音准确度, 基于子词后验概率的发音准确度, 把结果与最终评分比较, 计算相关系数和分差. 之后, 同时使用两者融合进行测试. 融合的方法如下式所示:

$$S_{\mathrm{merge}} = \lambda S_L + (1-\lambda)S_S. \tag{10}$$

式中, S_{merge} 是融合后的分数, S_L, S_S 分别为基于基于子词后验概率的发音准确度的分数和基于音素后验概率的发音准确度的分数, λ 是融合系数, 其取值在 0 到 1 之间.

测试结果如表 1 所示. 由表 1 可见, 子词后验概率的评分性能强于基于音素后验概率, 这说明基于子词的后验概率更客观地反映了发音质量. 当两种后验概率融合时, 系统的性能得到了更大的提升, 这说明这两种后验概率有一定互补性, 同时使用有利于更准确地反映发音质量.

5 结论和展望

本文通过在原有发音质量自动评估系统的基础上进行改进, 首次实现了将维语作为受测语言的发音评估系统, 这对于维汉双语教学的使用和推广都具有积极的意义.

期间研究分析了维语的发音规则和发音习惯, 采用了子词作为基本评估单元, 改进了在常规发音评估系统中通常采用音素作为评估单元的方法; 实现了基于双层词法分析的维语子词的机器切分器; 引入了音素解码器来计算后验概率的分母, 使得子词级别直接计算置信度成为可能, 从而得到基于子词后验概率的置信度, 提高了评分的全面性和准确性.

后续的研究将把注意力集中到维语在子词声学模型上, 针对维语的黏着语特性改进评分所使用的声学模型, 以达到更准确的评分效果.

参考文献

1 张瑜. 新疆基层汉族干部学习掌握维语现状的分析与对策研究 —— 以克州地区为例. 中国外资 (下半月), 2012, (10): 276–277

2 Dong B. Research on computer aided standard Chinese learning and objective evaluation of the pronunciation. Dissertation for Ph.D. Degree. Beijing: the Institute of Acoustics in Chinese Academy of Sciences, 2006 [董滨. 计算机辅助汉语普通话学习和客观测试方法的研究. 博士学位论文. 北京: 中国科学院声学研究所, 2006]

3 Kim Y, Franco H, Neumeyer L. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proceedings of Eurospeech'97, Rhodes, 1997. 649–652
4 Witt S M. Use of speech recognition in computer-assisted language learning. Dissertation for Ph.D. Degree. Cambridge: The University of Cambridge, 1999
5 徐明星, 宋战江, 郑方, 等. 汉语语音水平评价方法的研究. 第五届全国人机语音通信学术会议, 哈尔滨, 1998. 174–177
6 Pan F P. Research on evaluation algorithm of Mandarin pronunciation in computer aided language learning. Dissertation for Ph.D. Degree. Beijing: the Institute of Acoustics in Chinese Academy of Sciences, 2007 [潘复平. 计算机辅助汉语普通话发音质量评估算法研究. 博士学位论文. 北京: 中国科学院声学研究所, 2007]
7 Çarkin K, Geutner P, Schultz T. Turkish LVCSR: towards better speech recognition for agglutinative languages. In: Proceedings of ICASSP, Istanbul, 2000. 3688–3691
8 Creutz M, Lagus K. Unsupervised Morpheme Segmentation and Morphology Induction From Text Corpora Using Morfessor 1.0. Technical Report A81. 2005
9 Hacioglu K, Pellom B, Ciloglu T, et al. Onlexicon creation for Turkish LVCSR. In: Proceedings of Eurospeech'03, Geneva, 2003. 1165–1168
10 Arisoy E, Dutagaci H, Arslan L M. A unified language model for large vocabulary continuous speech recognition of Turkish. Signal Process, 2006, 86: 2844–2862
11 Hirsimäki T, Pylkkönen J, Kurimo M. Importance of high-ordern-gram models in morph-based speech recognition. IEEE Trans Audio Speech Lang Process, 2009, 17: 724–732
12 Tao M, Wushour S, Nasirjan T, et al. The Uyghur acoustic model based on HTK. J Chnese Inform Process, 2008, 22: 56–59 [陶梅, 吾守尔 · 斯拉木, 那斯尔江 · 吐尔逊, 等. 基于 HTK 的维吾尔语连续语音声学建模. 中文信息学报, 2008, 22: 56–59]
13 Ablimit M, Neubig G, Mimura M, et al. Uyghur morpheme based language models and ASR. In: Proceedings of IEEE-ICSP, Beijing, 2010. 581–584
14 Ablimit M, Kawahara T, Hamdulla A. Discriminative approach to lexical entry selection for automatic speech recognition of agglutinative language. In: Proceedings of ICASSP2012, Kyoto, 2012. 5009–5012
15 Li X, Hou W, Ji Z, et al. Lexicon design for Uyghur conversational telephone speech recognition. J Chongqing Univ Posts Telecommun (Nat Sci Edit), 2013, 25: 391–396 [李鑫, 侯炜, 计哲, 等. 面向维吾尔语电话交谈式语音识别的词典设计方法研究. 重庆邮电大学学报 (自然科学版), 2013, 25: 391–396]
16 Li X. Spoken term detection for Uyghur conversational telephone speech. Dissertation for Ph.D. Degree. Beijing: the Institute of Acoustics in Chinese Academy of Sciences, 2013 [李鑫. 针对维吾尔语电话交谈式语音的关键词检测技术研究. 博士学位论文. 北京: 中国科学院声学研究所, 2013]
17 Koskenniemi K. A general computational model for word-form recognition and production. In: Proceedings of 10th International Conference on Computational Linguistics and 22nd Annual Meeting of the Association for Computational Linguistics, California, 1984. 178–181

Bilingual teaching - facing automatic assessment for Uygur pronunciation

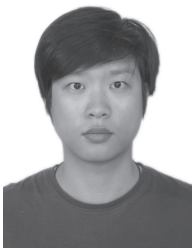
DONG Bin[1], DING Ming[1*], WANG Lei[2] & YAN YongHong[1]

1 *The Key Laboratory of Speech Acoustics and Content Understanding, Chinese Academy of Sciences, Beijing* 100190*, China*;
2 *The Xinjiang Technical Institute of Physics & Chemistry, Chinese Academy of Sciences, Urumqi* 830011*, China*
*E-mail: dingming@hccl.ioa.ac.cn

Abstract Pursuing Chinese and Uygur bilingual teaching in XinJiang Province is an important government policy for improving communication between different language speakers. An automatic assessment system can

yield many advantages in such bilingual teaching. However, as an agglutinative language, Uygur's particular way of word-building results in leaving many words out of the lexicon. Therefore, performance of the old pronunciation assessment system, which is based on a traditional ASR system, is poor. For building a high-performance system, we decide to use subwords as basic recognition units after analyzing the rules and habits of Uygur pronunciation. Experimental results indicate that the accuracy of the subword-based system is greatly improved by the implementation of machine segmentation based on double-level morphology analysis of Uygur subwords, calculating the posterior probability's denominator with a phoneme decoder, and calculating confidence at the subword level.

Keywords bilingual teaching, agglutinative language, pronunciation assessment, subword, posterior probability, confidence



DING Ming was born in China in 1986. He received a B.S. degree in acoustics from Nanjing University, Nanjing, China, in 2009 and went on to graduate work in CALL at the Institute of Acoustics at the Chinese Academy of Sciences, which is currently underway.