北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis
doi: 10.13209/j.0479-8023.2014.018

维吾尔语大词汇语音识别系统识别单元研究

努尔麦麦提·尤鲁瓦斯†　吾守尔·斯拉木　热依曼·吐尔逊

新疆大学信息科学与工程学院，乌鲁木齐 830046; † E-mail: nurmemet@xju.edu.cn

摘要　维吾尔语是一种黏着语，单词不太适合于作为维吾尔语大词汇连续语音识别系统识别单元。针对维吾尔语大词汇连续语音识别系统中的识别单元选择问题,设计更适合维吾尔语的子词识别单元，提出维吾尔语单词和子词相结合的组合识别单元构建方法，评价了单词、子词和组合识别单元的语言模型和语音识别性能。实验结果表明，所提出的识别单元在单元数量、语言模型复杂度等方面表现出更加优越的性能，并且使识别系统的单词错误率比基于单词的系统相对减少 22%。

关键词　维吾尔语；大词汇；语音识别；识别单元

中图分类号　TP391

Research on Recognition Units of Large Vocabulary Speech Recognition System of Uyghur

Nurmemet Yolwas† , Wushour Silamu, Reyiman Tursun

College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046; † E-mail: nurmemet@xju.edu.cn

Abstract　Uyghur is an agglutinative language and words are not optimal recognition units for Uyghur LVCSR systems. According to recognition unit selection problem in Uyghur LVCSR systems, a more suitable recognition units for Uyghur likes sub-word is designed, and the combining recognition units of word and sub-word are proposed. The performance of language models and speech recognition are evaluated on different recognition units. Experiment results show that the proposed recognition units outperforms word units in terms of unit size, language model perplexity, and can give a relative word error rate reduction of 22% over the word based system.

Key words　Uyghur; LVCSR; speech recognition; recognition unit

维吾尔语属于阿尔泰语系突厥语族，形态结构上属黏着语类型。维吾尔语中一个词根或词干与几个词缀链接后形成几个新单词。当单词作为识别单元时，由于大词汇量连续语音识别系统的识别基元有限，在维吾尔语大词汇量连续语音识别系统中可能会出现较高的未登录词(out of vocabulary, OOV)问题[1]，这会导致汉语和英语大词汇量连续语音识别中使用的方法在维吾尔语中不能表现应有的性能。

针对黏着语连续语音识别任务上的未登录词问题，过去十几年里研究者们提出了将单词切分成能够组合产生大量单词的较小语素(morpheme)单元，并将其作为识别单元的方法。在单词切分语素单元的问题上，有些研究者利用基于语言词法知识的词形态分析工具对单词进行了语素切分[2−3]，还有一些研究者则利用无监督学习方法，从语料库中自动选出最佳的基于统计的子词(sub word)单元[4−5]。Hirsimäki 等[6]对不同识别单元在不同黏着语连续语音识别任务中的性能进行很好的总结。由于基于无监督的统计方法不依赖预先准备的标注数据，在维吾尔语大词汇量连续语音识别中识别单元的选择问题上，薛化建等[7]采用此方法把维吾尔语单词切分成子词单元，并将其作为识别基元进行连续语音识别实验。

国家自然科学基金(61063024, 61363063)和新疆多语种信息处理重点实验室开放课题(049807)资助
收稿日期: 2013−06−14; 修回日期: 2013−08−25; 网络出版时间: 2013-11-06 11:23

虽然基于无监督学习的子词获取方法能够有效解决维吾尔语大词汇连续语音识别系统中的 OOV 问题, 但可能会影响语言模型上下文信息和提升声学混淆。比如, 维吾尔语单词 amerikiğa (ئامېرىكىغا, 意为: 对美国)通过无监督学习进行词切分子词后可能会生成 amerik+i+ğa 切分形式[7], 而且对于维吾尔语很多单词来说这种切分形式普遍存在。从这切分形式中可以看出, 即使创建 3 元语言模型, 也只能描述单词内的上下文关系, 失去单词之间的上下文信息。除此之外, 识别单元中可能会出现“i”之类的音素识别单元, 这会导致声学模型之间混淆。

针对以上问题, 本文从维吾尔语自身特点出发, 设计维吾尔语语音识别单元, 研究其在维吾尔语大词汇量连续语音识别系统中的性能。首先, 采用基于无监督的单词切分及后处理方法将维吾尔语单词切分成子词识别单元, 并且将统计意义上的词干子词作为词首, 后缀子词连接成词尾的方法形成单词的基于统计的词首词尾(stem-ending)识别单元, 减少过短单元可能带来的声学混淆。之后, 将出现频率较高的单词保留原型, 只对频率较低的词进行子词切分, 并且将高频单词与各种子词单元进行组合形成组合识别单元。最后, 评价并分析各种识别单元在测试数据上的语言模型复杂度和连续语音识别任务上的单元识别性能。

1 维吾尔语语音识别单元

1.1 音节识别单元

维吾尔语是构词和构形附加成分比较丰富的语言。维吾尔语中单词由词干和词缀组成, 而且词干与词缀, 词缀和词缀之间可以互相连接, 一个单词有若干个音节组成。比如, 将 oquğučilarni(把学生们) 切分成音节和词干词缀后有下面的形式:

oquğučilar =o+ qu+ğu+či+ lar +ni (音节序列),

oquğuçilar =oquğuči (词干)+ lar (词缀)+ni(词缀)。

维吾尔语的音节有一定规则, 维吾尔语固有的音节结构是 (起音)+领音+(收音)。领音必须是元音, 音节中可以没有起音和收音, 但是不能没有领音。所以可以通过规则方法对维吾尔语单词进行音节划分。本文采用规则方法对包含 1.335×10^6 个句子和 2.85×10^5 个不重复单词的维吾尔语文本语料库进行分音节处理。通过对文本语料进行音节统计发现约有 6465 个音节识别单元。

1.2 子词识别单元

维吾尔语单词切分词干词缀对维吾尔语自然语言处理研究至关重要。目前, 基于语言规则的维吾尔语词切分方法[8]和基于监督式的统计方法是维吾尔语词切分研究[9]的主流方法。规则方法需要语言学家制定规则库, 或者需要收集较大的维吾尔语词库, 而基于监督式的统计方法则需要手工切分的训练语料库。

维吾尔语中词干词缀是语言词法上的子词形式, 由于上述规则或基于监督的统计方法来处理维吾尔语单词切分问题都面临成本较高的问题, 因此, 本文主要关注统计意义上的词干词缀, 也叫做子词(sub word)。本文先采用基于无监督的统计方法[4]对维吾尔语文本语料库进行子词切分, 并且子词切分过程中对每一个子词赋予统计意义上的词干词缀属性。通过此方法获取的词干词缀属性不一定等于语言词法上的词干词缀。然后, 对切分结果进行以下后处理: 首先, 假如单词分解后子词中有单音素词缀单元, 将其添加到其后子词, 如果其后没有子词, 则将单音素子词添加到其前子词, 并将新子词属性与被连接的子词属性保持一致。其次, 虽然维吾尔语中前缀数量有限, 但切分结果中出现了很多前缀。观察发现这些前缀大部分是词干, 因此, 我们将每个词的分词结构中的前缀属性转换成词干。本文后面提到的子词识别单元经过这样处理产生。最后, 将统计意义上的词干子词作为词首, 后缀子词连接成词尾, 形成单词的基于统计的词首词尾识别单元。

为了能够在基于单词的训练语料库中训练子词、词首词尾, 利用单词和各识别单元映射词典对语料库单词进行子词替换, 生成了可训练的文本语

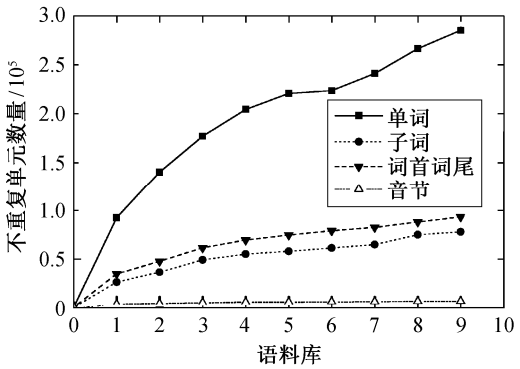

图 1 训练数据中不重复单元数量

Fig.1 Number of distinct tokens versus training data sets

料。通过以上处理发现文本语料中约有 7.8×10^4 个子词, 9.33×10^4 个词首词尾识别单元。图 1 给出了将语料库分成互相重叠的 9 个部分(每个部分新增 1.4×10^5 个句子)后对每一个部分进行统计得到的不重复单词、子词、词首词尾和音节识别单元数量。从图 1 可以看出, 随着语料库规模的不断增加, 子词、词首词尾和音节识别单元的数量比单词明显下降, 新单元增长率比较平稳。

1.3 组合识别单元

当子词作为识别单元创建 3 元语言模型时, 很有可能描述单词内的上下文关系, 失去单词之间的上下文信息。采用组合识别单元主要是为了利用单词和子词识别单元各自特点, 即减少系统未登录词率, 增强语言模型上下文信息。组合识别单元是由上述词首词尾识别单元和单词识别单元组成。在组合识别单元中, 将一定数量的高频词直接作为识别单元, 另一部分由与高频单词识别单元不重复的高频子词和词首词尾等单元组成。本文从不同训练语料库中选取出现频率较高的约 2.5×10^4 个单词、4.5×10^4 个子词、2.5×10^4 个词首词尾和 6465 千个音节, 去除重复单元后形成了 6.5×10^4 个不重复的组合单元。

为了能够训练组合识别单元的语言模型, 每次从语料取一个词, 判断该词是否在发音词典中, 如果是, 保留这个词, 如果否, 从子词、词首词尾列表中获取该词切分形式, 依次判断每一种切分形式中的子词单元是否全部出现在发音词典中, 如果是, 替换成该词相应切分形式, 如果否, 对词进行音节划分, 词替换成音节形式。

2 发音词典

维吾尔语中有字形与音位一一对应的特点, 因此发音词典的生成比较简单。本文分别利用语料库中出现频率较高的 6.0×10^4 个单词、6.5×10^4 个子词、词首词尾和组合识别单元创建发音词典。图 2 给出了维吾尔语各识别单元在发音词典中的音素个数分布情况。从图 2 可以看出, 一些形态比较复杂的单词被切分成音素分布概率比较集中的子词识别单元, 这种分布有助于子词单元声学得分上公平竞争。由于音节识别单元音素个数主要分布在 3 个音素左右, 因此, 组合识别单元中增加了相应音素个数分布概率。

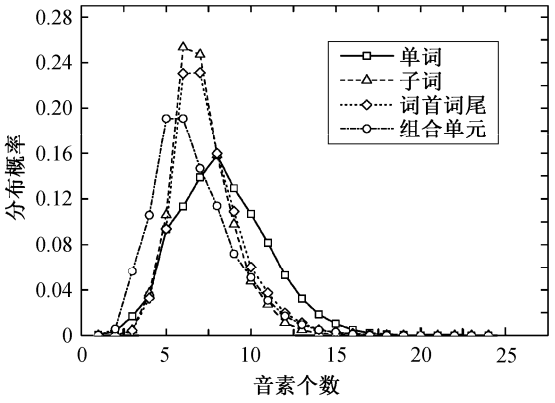

图 2 识别单元音素分布概率

Fig. 2 Distribution of phoneme in recognition units

3 语言模型

可训练文本语料生成以后, 采用 SRILM[10]语言模型训练工具, 将本文语料库中各识别单元作为基元, 分别建立各识别单元的 2 ~ 5 元语言模型, 并采用 Katz 语言模型平滑技术解决数据稀疏的问题。

语言模型性能一般采用交叉熵和复杂度来进行评估。对于不同基元的语言模型, 交叉熵比复杂度能更好地描述语言模型性能[11]。图 3 中给出了基于不同识别单元的维吾尔语语言模型在包含 2.5×10^4 个句子、3.9×10^5 个单词和 6×10^4 个不重复单词的测试文本语料库上的交叉熵。从图 3 可以看出, 子词语言模型的不确定性低于单词语言模型, 这说明子词可以作为比单词更有效的语言模型基元。其中, 由于音节基元少, 能充分训练, 因此音节语言模型交叉熵最低。另外还可以看出, 由于本文文本语料数据还不足于训练更高阶次的语言模型, 在这两个识别单元上 4 元以上语言模型交叉熵没有发生太多变化。

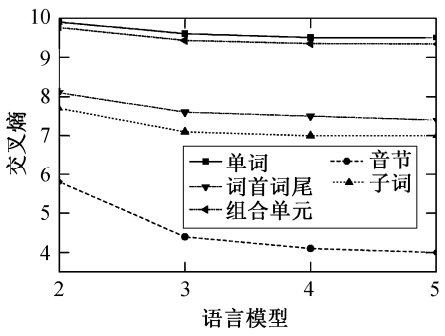

图 3 各识别单元语言模型交叉熵

Fig. 3 Cross entropy of different language models

4 实验与分析

4.1 数据集

本文使用采样率 16 kHz, 采样位 16 bit, 约 128 小时的 356 人(189 女, 167 男)的维吾尔语标准口音且以朗读式语音数据。其中将 346 人的语音数据作为训练数据提取 39 维 MFCC 特征, 并使用倒谱均值方差归一化方法进行降噪处理。声学模型利用 HTK[12]工具将维吾尔语 32 个(包括静音和停顿模型)音素作为声学模型基元使用 MLE 准则训练, 其中每个模型用 3 个输出状态表示, 然后扩展三音素模型并绑定到 9955 个状态, 每一个状态用 16 个独立的高斯混合分布表示。静音模型采用 5 个状态的 HMM 模型, 停顿模型采用 3 个状态的 HMM 模型, 模型中每个状态包含 24 个独立的高斯混合分布。

从语音数据库中挑选一组有 10 个说话人(5 男, 5 女)语音数据, 共 1018 个语句, 9805 个单词, 约 2 小时, 作为测试集。

4.2 实验结果

为了与基于单词的大词汇量连续识别系统进行比较, 本文将基于音节、子词、词首词尾和组合识别的连续语音识别系统输出的识别单元序列采用最大匹配算法转换成单词序列, 然后评价对应识别单元环境下连续语音识别系统的单词错误率(word error rate, WER)。除此之外, 还对单元错误率(unit error rate, UER)、字母错误率(letter error rate, LER)和平均识别效率(xRT)进行评价。

表 1 给出单词、音节、子词、词首词尾以及组合识别单元在测试集上连续语音识别的性能。所有实验中激活模型数、状态最大对数概率和语言模型概率阈值设置相同。通过几组实验确定让各识别单元表现出最佳的识别性能的语言模型因子和单词插入惩罚度。从表 1 可以看出, 由于子词、词首词尾和组合单元的语言模型性能较好, 因此, 单元错误率、字母错误率和单词错误率比单词有明显下降。与此同时, 虽然音节语言模型交叉熵最低, 但由于最大匹配分词算法在音节序列转换单词序列错误率较高, 因此音节识别单元没有表现出很好的单词识别性能, 反而最大匹配分词算法效率较高的子词、词首词尾及组合识别单元上得到较低的单词错误率。从表中还可以看出, 基于组合识别单元的识别系统单词错误率有所上升, 但识别效率比子词和词首词尾识别单元得到提升。

表 1 识别性能

Table 1 Recognition performance

识别单元	xRT/%	UER/%	LER/%	WER/%
单词	8.2	20.6	6.6	20.6
音节	7	9.8	4.8	27.7
子词	8.2	11.1	3.7	16.0
词首词尾	8	12.5	3.8	16.4
组合单元	7.7	14.2	4.3	17.0

5 结束语

本文从维吾尔语自身特点出发研究了维吾尔语大词汇连续语音识别系统中的识别单元选择问题, 提出了几种维吾尔语语音识别单元, 详细分析了几种识别单元的单元增长率、语言模型交叉熵、发音词典音素个数分布情况以及语音识别性能。从实验结果中可以看出, 维吾尔语子词、词首词尾和组合识别单元可以有效解决维吾尔语大词汇量连续语音识别系统中的 OOV 率问题。除此之外, 子词、词首词尾语言模型交叉熵低于单词语言模型。从连续语音识别性能来看, 子词、词首词尾和组合识别单元使语音识别系统的单词错误率比基于单词的系统相对减少。因此在一些应用任务上, 如语音检索, 可以考虑子词、词首词尾作为识别单元。

参考文献

[1] 张小燕, 宿建军, 薛化建, 等. 维吾尔语语音识别语料库中的 OOV 研究. 计算机工程与设计, 2012, 33(2): 772−776

[2] Arısoy E, Dutağacı H, Arslan L M. A unified language model for large vocabulary continuous speech recognition of Turkish. Signal processing, 2006, 86(10): 2844–2862

[3] Tanel A. Phonological and morphological modeling in large vocabulary continuous Estonian speech recognition system // Proceedings of Second Baltic Conference on Human Language Technologies. Tallinn, Estonia, 2005: 89–94

[4] Creutz M, Lagus K. Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing, 2007, 4(1): 3–36

[5] Creutz M, Hirsimäki T, Kurimo M, et al. Analysis of morph-based speech recognition and the modeling of

out-of-vocabulary words across languages // Proceedings of NAACL HLT. Rochester, NY, 2007: 380–387

[6] Hirsimäki T, Pylkkönen J, Kurimo M, et al. Importance of high-order n-gram models in morph-based speech recognition. IEEE Transactions on Audio, Speech and Language Processing, 2009, 17(4): 724–732

[7] 薛化建, 董兴华, 周喜, 等. 基于子字单元的维吾尔语语音识别研究.计算机工程, 2010, 37(20): 208–210

[8] 古丽拉•阿东别克, 米吉提•阿布力米提. 维吾尔语词切分方法初探. 中文信息学报, 2004, 18(6): 61–65

[9] 早克热•卡德尔, 艾山•吾买尔, 吐尔根•依布拉音, 等. 混合策略的维吾尔语名词词干提取系统. 计算机工程与应用, 2013, 49(1): 175–179

[10] Stolcke A. SRILM — an extensible language modeling toolkit // Proc ICSLP2002. Colorado, 2002, 2(1): 901–904

[11] Goodman J T. A bit of progress in language modeling. Computer Speech and Language, 2001, 15(4): 403–434

[12] Young S, Evermann G, Gales M, et al. The HTK book [Z/OL]. http:// htk.eng.cam.ac.uk/