第 37 卷 第 2 期
Vol.37 No.2

计 算 机 工 程
Computer Engineering

2011 年 1 月
January 2011

• 人工智能及识别技术 • 文章编号：1000—3428(2011)02—0197—03 文献标识码：A 中图分类号：N945

基于 DDBHMM 的维吾尔语音声学识别

王飞飞[1a]，吾守尔·斯拉木[1a]，那斯尔江·吐尔逊[1b,2]

(1. 新疆大学 a. 信息科学与工程学院；b. 数学与系统科学学院，乌鲁木齐 830046；2. 西安交通大学电子与信息工程学院，西安 710049)

摘 要：在维吾尔语连续语音识别试验的声学层建模基础上，引用 DDBHMM 模型将上下文相关的三音子作为基本识别单元，并提出一种状态绑定的思想，对状态进行优化。为得到更充分的训练模型，提高识别效率，对语料库进行扩充，在多组对比试验的基础上，分析扩充前后对声学层识别速度、准确率等各个方面的影响。

关键词：语料库；维吾尔语；DDBHMM 模型理论；三音子

Uyghur Speech Acoustics Recognition Based on DDBHMM

WANG Fei-fei[1a], Wushour Silamu[1a], Nasirjan Tursun[1b,2]

(1a. Information Science and Engineering College; 1b. Mathematics and Systems Science College, Xinjiang University, Urumqi 830046, China;
2. Electronic and Information Engineering College, Xi'an Jiaotong University, Xi'an 710049, China)

【Abstract】DDBHMM(Duration Distribution Based HMM) is adopted as the acoustic model for Uyghur continuous speech recognition, and the context-dependent triphone model is selected as the best recognition unit, the Uyghur speech recognition system is optimised by using the state-binding method. In order to make the models be trained more sufficiently to improve the recognition performance, the corpus is enlarged, the emphasis is on analysis of the effect that the speech database's enlargement brings to the recognition rate and accuracy and so on based on several groups of contrasted experiments.

【Key words】corpus; Uyghur; DDBHMM model theory; triphone

DOI: 10.3969/j.issn.1000-3428.2011.02.068

1 概述

维吾尔语属于阿勒泰语系突厥语族，是黏着性语言，其语音的上下文联系很紧密，同一字母的发音会因其在词中所处位置的不同而不同。故去年笔者所在工作组引入清华大学王作英教授在 20 世纪 90 年代初提出的 DDBHMM(Duration Distri- bution Based HMM)模型理论[1]，在此理论基础上进行维吾尔语连续语音识别试验，由于部分模型训练不是很充分，因此这次主要扩大语料库进行进一步试验。

2 维吾尔语语音特征

维吾尔语隶属阿尔泰语系突厥语族西匈语支，属于黏着语类语法结构。维吾尔语的音素是从音色角度划分出来的，可以划分为元音和辅音两大类。其中，元音有 8 个，辅音有 24 个。在连续语流中，音段音素要互相结合以形成更大的语音单元如音节、短语、句子等。和汉语一样，维吾尔语中能够感受的最小语音片段是音节，它是维吾尔语音的基本构成单位。最基本的维吾尔语音节由元音和辅音构成，每个音节有且只有一个元音，构成方式有单个元音、元+辅、辅+元、辅+元+辅、元+辅+辅、辅+元+辅+辅等，但是后来由于一些外来词汇如汉语、俄语等的引入新增了 5 种音节模式：辅+辅+元、辅+辅+元+辅、辅+辅+元+辅+辅、辅+元+元、辅+元+元+辅等。在 2 个音节或 2 音节以上的词中，一般最后一个音节为词重读音节。维吾尔语音节结构是：(起音)+领音+(收音)。其中，领音是不可缺少的，音节中可以没有起音和收音，但不能没有领音，而且，领音必须是元音。

3 基于 DDBHMM 的声学模型的建立

3.1 DDBHMM 模型简介

DDBHMM 以状态驻留度的概率分布为基本参数，解决了经典 HMM 的齐次性假设导致的段长指数分布问题，同时也克服了经典 HMM 对于转移概率的独立估计这些实际语音发音特点的缺陷，DDBHMM 模型的状态驻留概率和转移概率可以完全由状态段长分布确定，不仅能更好地描述各帧之间的相关性而且在计算上由于无需对状态驻留概率和转移概率进行独立估计，只需从训练数据中估计出段长分布即可，因此使得模型训练量大大减少。

本文是在 DDBHMM 基础上采用最大似然训练和识别算法[2]，训练和识别最终都归结于最优状态路径的获得，也就是对语音序列的分割。在 DDBHMM 中，识别网络中路径的积累距离就是路径所经历的状态与语音的特征矢量的匹配距离的累加值加上段长匹配距离的和。MLSS 算法就是在识别网络中找出一条积累距离最小的作为最佳路径，MLSS 训练算法训练先是语音分段确定语音特征序列的状态分割点，接着便进行模型参数重估即根据语音样本的分割点训练各个状态的模型参数，这样不断进行迭代得到最终的码本进行识别。无论是训练还是识别，其关键问题都是要快速地搜索出各个模型 λ 下的 MLSS 路径。

3.2 识别基元的选择

维吾尔语是黏着性语言，协同发音的现象也比较普遍，每个音素的发音受其左右临近音素间的强烈影响，同一个音

基金项目：国家自然科学基金资助项目(60762006, 60863008)；国家语委基金资助重点项目(MZ115-75)

作者简介：王飞飞(1984－)，女，硕士研究生，主研方向：语音信息处理；吾守尔·斯拉木，教授、博士生导师；那斯尔江·吐尔逊，副教授、博士研究生

收稿日期：2010-06-10 **E-mail：**lhc006@126.com

素在不同的上下文中发音会发生变化，显然不能仅简单采用音素作为识别基元。

音节虽是维吾尔语中最基本的发音单位，但音节数量很大，常用的就有 3 000 多个，再考虑到发音时音节间的影响，则模型的数量就太多，所以，以音节作为识别基元也不是理想的选择。因此本文在 DDBHMM 模型基础上选取三音子作为识别基元，该模型考虑了上下文的相关性。

三音子是一种较理想的音子模型，它充分考虑上下文的协同发音的影响，非常贴近维吾尔语的发音特点，可以解决维吾尔语在发音时，部分音节在语流中产生语流音变现象，以及常见的同化、弱化、脱落以及元音和谐等现象[3-4]。

3.3 声学模型的建立

声学模型是识别系统的底层模型对应于语音到音节概率的计算。它的目的就是提供一种有效的方法计算语音的特征矢量序列和每个发音模板之间的距离，为每个声学单元建立一套声学模型参数。

声学模型的设计与语言发音特点密切相关，这里结合维吾尔语的发音特点，主要讨论模型训练问题。对于大词汇量、连续语音识别系统来讲，大量的训练是必需的，在 DDBHMM 模型训练中关键问题就是要快速地搜索出各个模型下的 MLSS 路径。文献[2]给出了获得最佳分割点 $t_1,t_2,\cdots,t_N$ 快速算法，此算法采用帧同步算法搜索最优路径。在对每一帧进行剪枝时，保证最优路径不会被错误地删除。在连续语音识别中采用此方法进行最优路径搜索的计算量可以比全搜索的方法下降 3 个数量级以上。

本文初始模型参数是没有经过任何工具处理的最原始的参数值，通过 DDBHMM 训练算法进行循环往复的参数估计，通过训练不断地调整模型参数，使系统中的所有模型彼此间的距离尽可能达到最大，最终得到一个可以用于识别的最终模型。

3.4 声学模型的优化-状态绑定

目前，从 16 740 个词中已经提取的全部三音子模型是 8 672 个，每一个模型包含 5 个状态，实验时根据语音特点省略掉了前后的静音部分剩下 3 个状态，但待训练的总状态数=8 672×3+1=26 016 个(加一个静音)仍很大，会导致识别计算量大耗时长。本文根据维吾尔语的特征，进行状态绑定来优化声学模型。所谓状态绑定就是分析不同三音子所包含的状态是否属于相同状态，是则看成同一状态，这样减少总的状态数，而绑定前是根据三音子顺序一个三音子包含 3 个状态(实则 5 个，只使用了 3 个)逐一编号，这样虽然状态精度相对高一些，但是状态量太大严重影响训练识别速率。状态绑定的关键问题是绑定的原则，即什么样的状态可看成同一状态，这便要根据具体的维吾尔语的发音特点：一个维吾尔语三音子模型可能由,3 个音素组成，比如 G-E+l，这种情况下中间音素为主，前后的音素按照维吾尔语音素的连接规则连接起来；也可能由 2 个音素组成，比如是 E+w 或 H-E，在这种情况下前面或者后面是静音；也可能单独一个音素作为一个模型，它的前后都是静音(这个情况只有一个，就是 u)。

本文暂时采用下面的状态绑定的准则(注：以下均是在中间音素相同的情况下，中间音素若不同没有这 2 个模型中不存在相同状态)：

(1)若前面都是静音或是同一个音素，则它们第 1 个和第 2 个状态分别属于同一个状态如 E+H 和 E+K，再如 E-E+j 和 E-E+l。

(2)若后面都是静音或是同一个音素，则它们的第 2 和第 3 状态分别属于同一状态如 H-E 和 K-E，再如 E-E+y 和 G-E+y。

(3)前后都不相同则这 2 个模型只有第 2 状态属于同一状态如 E-E+y 和 G-E+x。

通过该方法，进行状态绑定最后得到的状态数 1 847 个，大大减少了总的状态数。状态绑定以后不仅给试验带来方便使得识别速度加快，而且试验结果证明采取状态绑定后识别率也有所提高。

4 语料库

语料库(Corpus)[5]是大量语音数据组成的数据库，对于 DDBHMM 模型和整个识别器的性能有着极其重要的作用，包括文本语料库和语音语料库。语料库的建立主要经过前期准备、文本语料的收集整理、发音人的选择、录音等步骤。

本文文本语料主要来源于维吾尔文新闻、网站、小说等，初步筛选出 25 000 多个句子覆盖了多种题材及维吾尔语语言中重要的句法和语调变化，语音现象，包含了大部分维吾尔语音素、音节、词根、词缀。旧语料库是在这 25 000 个句子中进一步通过自动挑选和人工补充相结合的方法得到 1 018 个代表性的句子包含 5 501 个词 1 587 个状态。但考虑到训练更加充分，必须使所选语料尽可能覆盖维吾尔语中所有的三音子，并解决数据稀疏的问题，具体做法如下：

(1)统计词频，反复统计语料库中单词的出现率，筛选词频为 10 次以上的单词，称作高频词。

(2)因语料设计是针对连续语音识别系统，句子是一个基本单位，自动切分长句并挑选出那些包含高频词的所有句子。

(3)对实际语料进行预处理，去掉太长和太短的句子。

(4)对各句计分，评估每个句子对三音子的覆盖贡献，按此贡献排序。

这次扩大后的语料库又增加了 1 438 句，现在的语料库总共是 2 456 句，包含了大约 16 740 个不重复的单词，形成了维吾尔语连续语音识别系统的文本语料。

扩大语料库前后各数据情况如表 1 所示。

表 1 相关数据

语料	句子	词	三音子	绑定后状态
旧语料	1 018	5 501	5 130	1 587
新语料	2 456	16 740	8 374	1 847

5 试验与分析

本文试验是在非特定人连续语音数据集上进行的。语音的帧长为 20 ms，每帧语音的归一化能量及 14 维 MFCC 系数连同它们的一阶差分、二阶差分系数共同构成了 39 维的特征矢量。特征的观测概率采用 39 维全协方差阵的 Gauss 分布，段长分布采用 1 维的 Gauss 分布，每个维吾尔语采用 5 个状态的 HMM 表示。

将初始码本先进行训练，利用充分训练后的码本进行识别。由于扩大语料库后数据较多，本文主要是对女声数据方面的试验(限于篇幅仅选取部分数据罗列)。

5.1 扩大语料库后的试验

扩大语料库后，在 CPU 为 Intel Pentium 4 3.4 GHz，内存为 1.25 GB 的硬件条件下进行多次训练和识别试验，如表 2 所示(表中的新数据的初始码本是没有经过任何工具处理的，即将所有状态都赋一样的初始参数值，而旧数据是将采用 HTK 的 HCompV 工具估计后的模型参数作为初始模型参数再进行训练)。

表 2 针对女声的扩大语料库后的识别率

组别	码本	识别文件	识别率/(%)
1	码本 3	1 017 句	58.34
2	码本 3	训练过的 588 句	73.45
3	码本 3	1 438 句	62.58

本组试验是看同样的训练码本下不同数据的识别情况。第 1 次是识别旧数据未参加训练的，第 2 次还是识别旧数据中已经训练过且只随机抽取了一部分，只有 588 句，第 3 次是识别新数据中的 1 438 句。从这 3 组数据可看出，识别训练过的比未训练过的识别率有很大提高，同样识别训练过的但识别新数据的结果不如旧数据，这次新数据的录音质量较前面的旧数据应该有所下降，且新录音中对新数据如三音子的覆盖次数以及训练次数都相对没有旧数据高。

从表中可以明显看到识别新句子比识别旧句子花时长，主要是因为这次新句子普遍都比旧句子长，旧句子大多 15 个词以内，而新句子多是几十个词一句。

5.2 与旧语料库下的试验对比

以下几组试验的识别数据均是旧语料中的 F37-F41 的 1 017 句。结果如表 3、表 4 所示。

表 3 第 1 组对比试验

组别	语料库	初始码本	训练文件	识别文件	识别率/(%)	时间/句
1	旧	旧	6 265 句	1 018	60.67	1.93 min
2	旧	新	同 1	同 1	57.86	1.97 min
3	旧	新	6 265	—	60.67	1.93 min
4	新	新	3 133 句	同 3	54.30	2.15 min

表 4 第 2 组对比试验

组别	语料库	初始码本	训练文件	识别文件	识别率/(%)	时间/句
1	旧	新	3 133 句	一样	54.30	2.15 min
2	新	新	11 143 句	—	58.34	6.38 min

在第 1 组试验中，1,2 是看码本对识别的影响，可知未经任何处理的初始码本训练后的码本去识别，识别率有所下降，若采用更好的初始码本或更合适的训练算法，会得到更好的码本，识别率也会有所提高；在第 1 组试验中，3,4 是看在旧数据下，看训练数据的多少对识别的影响，可以看出识别数据多的识别率明显高一些；第 2 组也是比较训练数据的多少对识别的影响，和第 2 组的结果一样也是训练数据越多识别率越高，但是由于使用新语料库随着语料的增多，词数，三音子数目的增多直接导致状态数的增多，使得识别同样的句子时间将近是原来的 3 倍。而且从表中可以看到尽管第 3 组的试验 2 比第 2 组的试验 1 的训练数据多很多但识别效果反而不如后者，这是因为语料库虽然增加了，但是都是新句子，并没有增加已有句子的训练。

6 结束语

本文主要针对语料库扩大后从声学层方面介绍基于 DDBHMM 的维吾尔语连续语音识别。在状态绑定的基础上，扩大语料库后的识别率较之以前是有所改善的，基本提高 5 个百分点，但是由于新增的录音全部是针对 1 438 个新句子的，新录音里每个句子平均由 8 个不同的人各读过一遍，而旧语料每个句子被 3 个~5 个女性各读一遍，所以，扩大语料库后识别旧句子的正确率不如新句子的正确率；新语料库句子较之旧语料库句子明显长很多，再加上语料库的增加带来的词数、三音子、状态数的增加都导致识别时间明显变长；另外由于这次采用的码本是没有经过任何处理的，各个状态都给的同样的初始值，完全依靠试验训练，且新语料库的录音质量较之旧语料的录音有明显下降，这些都使得识别的正确率没有达到最佳。下一步工作可从以下 4 个方面着手：(1)对于新句子再增加录音，改善个别状态训练不充足的问题。(2)最好选定一个更好的初始码本。(3)增加端点检测，可以尝试利用 LDA+MLLT+CMS 算法[6]组合提取语音特征降低噪音干扰，采用多高斯混合模型，以期提高识别效率。(4)尽快增加语言模型。

参考文献

[1] 王作英, 肖 熙. 基于段长分布的 HMM 语音识别模型[J]. 电子学报, 2004, 32(1): 46-49.

[2] Wang Zuoying, Gao Hongge. An Inhomogeneous HMM Speech Recognition Algorithm[J]. Chinese Journal of Electronics, 1998, 7(1): 73-77.

[3] 哈力克·尼亚孜. 基础维吾尔语[M]. 乌鲁木齐: 新疆大学出版社, 1997.

[4] 李 春, 王作英. 基于语音学分类的汉语三音子识别单元的算法[J]. 清华大学学报: 自然科学版, 2003, 43(1): 16-19.

[5] 那斯尔江·吐尔逊, 吾守尔·斯拉木, 麦麦提艾力. 维吾尔语大词汇量连续语音识别研究——语音语料库的建立[C]//第十一届全国民族语言文字信息学术研讨会议论文集. 北京: [出版者不详], 2007.

[6] 王安娜, 王勤万. 改进的语音特征提取方法及其应用[J]. 计算机工程, 2008, 34(5): 196-197, 200.

编辑 陈 文

(上接第 196 页)

参考文献

[1] 中华人民共和国建设部. CJ/T215-2005 城市市政综合监管信息系统地理编码[S]. 2005.

[2] 陈细谦, 迟忠先, 金 妮. 城市地理编码系统应用与研究[J]. 计算机工程, 2004, 30(23): 50-52.

[3] Clodoveu A, Davis J, Fonseca F T. Assessing the Certainty of Locations Produced by an Address Geocoding System[J]. GeoInformatic, 2007, 11(1): 103-129.

[4] Agnew D C. A Compact Global Geocoding Suitable for Sorting[J]. Computers and Geosciences, 2005, 31(8): 1042-1047.

[5] Rahul B, Craig A K, Snehal T. Exploiting Online Sources to Accurately Geocoding Addresses[C]//Proc of the 12th Annual ACM International Workshop in Geographic Information Systems. Washington D. C., USA: ACM Press, 2004: 194-203.

[6] 江 洲, 李 琦. 地理编码(Geocoding)的应用研究[J]. 地理与地理科学, 2003, 19(3): 22-25.

编辑 陈 文