文章编号：

维吾尔语语音检索技术研究

张力文[1]，努尔麦麦提•尤鲁瓦斯[1]，吾守尔•斯拉木[1]

（1.新疆大学 信息科学与工程学院，新疆维吾尔族自治区 乌鲁木齐市 830046）

摘 要 随着大数据时代的到来，各种音频、视频文件日益增多，如何高效地定位关键敏感信息具有非常重要的研究意义，目前研究人员对针对英语和汉语的语音检索技术进行了深入地研究，而针对维吾尔语的语音检索技术还处于起步阶段。针对这一背景本文对维吾尔语语音关键词检索技术进行了研究并采用了大词汇量连续语音识别、利用聚类算法将多候选词图转换为混淆网络、倒排索引、置信度以及相关度的计算等技术和方法，对维吾尔语语音检索系统进行了研究与搭建。最后在测试集上对该系统进行测试，测试结果显示在语音识别正确率为 82.1%的情况下，当检索系统的召回率分别达到 97.0%和 79.1%时，虚警率分别为 13.5%和 8.5%。

关键词 维吾尔语；语音检索；语音识别；词图；混淆网络；倒排索引；

中图分类号：TP391 **文献标识码：**A

Study on Uyghur Speech Retrieval Technology

Zhang Li-wen[1], Nurmemet Yolwas[1] ,Wushour Silamu[1]

(1.College of Information Science and Engineering, Xinjiang University, Urumqi Xinjiang 830046, China)

Abstract With the advent of the age of big data, the number all kinds of audio and video files are increasing, how to efficiently locate key sensitive information is very important . At present，researchers have deeply studied English and Chinese speech retrieval technology， but Uyghur speech retrieval technology is still in its infancy.Based on this background, this paper, by using the technologies of large vocabulary continuous speech recognition, lattice into confusion network, inverted index, Confidence level and similarity calculation to research and set up the Uyghur speech retrieval system. Experimental results show that in the case of recognition accuracy rate is 82.1%，when recall rate reaches 79.1% and 97.0%，false alarm rate is 8.5% and 13.5%.

Key words Uyghur；Speech Retrieval；Speech Recognition；Lattice；Confusion Network（CN）；Inversed Index；

1 引言

语音检索[1]运用大词汇量连续语音识别（Large Vocabulary Continuous Speech Recognition）的技术将语音数据转换为文本[2]，并根据识别结果建立索引，检索系统根据用户输入的包含关键词的查询请求（Query），在文本中搜索与之对应的文件，最后返回相关的语音段。

目前大多数语音检索系统都是基于语音识别技术的，其中有剑桥大学的 Video Mail Retrieval Using Voice[3]，随着语音识别解码技术的不断发展，基于音素或音节网格的语音检索技术也成为语音研究领域中的热点之一，具有代表性的系统有 Google 推出的 Google Voice Local Search[4]。90 年代初我国也开始对语音检索领域进行深入研究其中中科院进行了查询词为语音的汉语语音文件检索任务[5]，哈尔滨工业大学基于关键词检出技术提出了一种音节网格的语音检索技术[6]。

维吾尔语语音识别研究工作开始于 20 世纪 90 年代初。1994 年，吾守尔·斯拉木采用独特的音节训练词识别方法和词汇扩充方法等技术，研制出联想式特定人维吾尔语音识别系统，其识别率达到 95%[7]。2012 年中国科学院新疆理化所对维吾尔语广播新闻连续语音信号进行敏感词检索[8]，该文献的工作是对语音文件中的敏感词汇进行检索，与该文献有所不同的是，本文所研究的维吾尔语语音检索系统目标是快速针对用户的输入信息对语音文件进行检索与定位。

本文所做的工作主要包括：（1）采用大词汇量连续语音识别技术将维吾尔语语音数据转换为文本数据；（2）将多候选的识别结果词图（lattice）转换为对应的混淆网络（CN）；（3）根据混淆网络建立索引，完成对维吾尔语语音检索系统的搭建；最后在实验部分对该系统进行评测并对评测结果

* **收稿日期：** **定稿日期：**

基金项目：973 国家重点基础研究计划 (No.2014CB340506)；新疆维吾尔自治区科技计划项目(No.201312104).

进行分析。

2 系统框架设计

本文所介绍的维吾尔语语音检索系统以大词汇量维吾尔语语音识别作为前端处理，以词作为识别单元，其识别结果是词或音节的词图（Lattice）多候选结构，识别结果词图再通过聚类算法转换为混淆网络(CN)[9-10], 为了提高检索的速率，根据混淆网络建立倒排索引并将索引存储在文本文件中。检索时对用户输入的维吾尔语查询短语用空格进行分词和预处理，将其转换成可以被检索系统接受的形式，之后利用索引实现检索。最后采用置信度[11-12]评测的方法对结果进行确认和验证，输出所要查询的词语，整个语音检索的设计框架如图 1 所示：

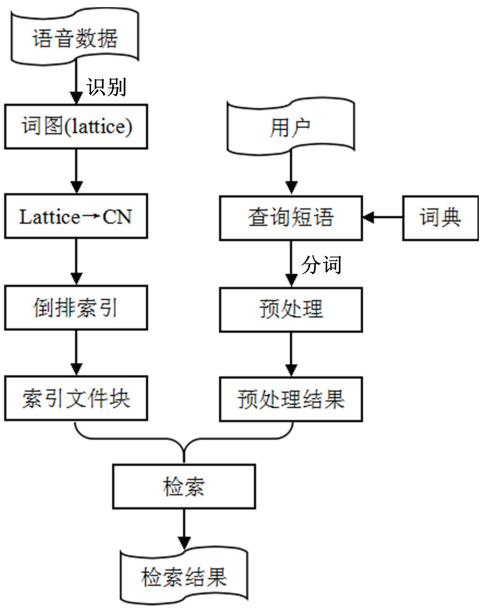

图 1　维吾尔语语音检索系统设计框架

2 维吾尔语语音检索系统关键问题

2.1 维吾尔语大词汇量连续语音识别

维吾尔语 LVCSR 系统[13]与其他语言 LVCSR 系统一样，实质上都属于同一种模式识别，一般情况下系统的主要任务是根据给定的一段语音序列在 HMM 状态空间中找到最优的状态序列，从而找出使这段语音所代表的最有可能的词序列即首选(1-Best)识别结果。而在在语音检索任务中检索系统一般需要识别结果对正确文本的有较高的覆盖率以减小检索结果的漏报率，因此在解码过程中本文利用 HTK[14]工具集中的 HDecode 模块生成多候选（N-Best）的词图（Lattice）识别结果。

2.1.1 声学模型

维语与汉语相似发音都是以音节为单位，但是维语的音节数量非常大常用的就有 3000 多个，同时每个音节之间相互独立没有共享的内容，因此使用音节作为声学建模单元是不理想的。而维吾尔语中音素仅有 34 个（包括 sil 和 sp）非常适合作为声学模型的建模单元，同时考虑到上下文的因素因此采用上下文相关的三音素作为建模单元。那么在理论上就会有 38355 个三音素模型，然而实际在训练集中只出现了 12395 个模型，不同的三音素模型平均有 391 个训练样本，这会导致有些模型不能得到充分训练或某些模型根本没有被训练，为了解决这个问题本文采用基于最大似然决策树的状态共享策略，并且根据维吾尔语语音特征（如元音、辅音、塞音、擦音、塞擦音等）设计了 156 个问题集给决策树在决策分类过程中提供依据[13,15]。

在建立声学模型之前，本文对训练语音数据提取 39 维 MFCC 特征(帧长 25 ms，帧移 10 ms)，其中包括每一帧数据的 12 维倒谱系数和能量及其一阶和二阶差分倒谱，并使用倒谱均值方差归一化方法进行降噪处理。得到 MFCC 特征之后便可采用上面所描述的基于上下文的三音素 HMM 模型进行声学建模，模型训练过程中先利用 HTK[14]工具对其进行 MLE 训练，最后再利用 MLLR 和 MAP 自适应方法对模型进行自适应优化。

2.1.2 语言模型

本文采用基于统计方法的语言模型，在语言模型生成之前先做训练数据的准备，每个文本文件中的每一句是以<s>开始，以</s>结尾，每个词用空格分开。一般由于训练语料中很难包含所有的可能的词序列组合，因此本文采用正向生成的二元模型和逆向生成的三元模型来解决模型的稀疏问题，其中正向模型依赖于它左侧的上下文，而逆向模型依赖于它的右侧的上下文。语言模型利用 SRILM[16]工具训练。

2.2 词图（lattice）转换为混淆网络

由于在面向大型的语音音频文件时，词图是一种非线性的图形结构，因此语音检索的过程中用词图作为索引就使得索引所占的存储空间较大；同时由于词图包含的每一个候选结果都是基于其后验概率尽可能大而得来的，这就不能保证识别结果中每个词的错误率最小。而由 Mangu[9]提出的混淆网络存储格式从词错误率最小的角度出发对词图进行了优化，这就使得识别结果词图从原来的对整个

待选句子的决策变成了对多个候选词的决策，从而使得识别结果的存储空间也相对减小了许多。因此本文将识别结果的存储格式由词图转换为混淆网络，混淆网络形式的识别结果如下图所示：

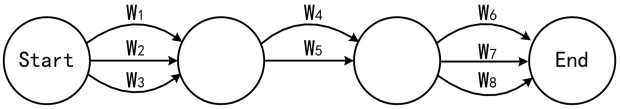

图 2 混淆网络形式的识别结果

本文使用 Mangu、Brill 等人提出的聚类算法（Clustering Algorithm）将词图转化为混淆网络，算法大致有如下几个步骤[9-10]：

（1）lattice 中的弧上都包含了一些得分，采用前-后向算法计算每条弧的后验概率；

（2）对后验概率小于事先设定阈值的弧进行裁减；

（3）对相同词的弧进行合并，并将合并前每条弧上的后验概率进行求和，得到合并弧的后验概率；

（4）对在同一时间间隔内相互竞争且拥有相同语音性质的互不相同的词进行聚类，最终形成混淆网络。

对后验概率较低的弧进行裁剪是为了更好地将相互竞争的词进行对齐，同时可以提高系统的检索速率。然而如果裁剪阈值设定的过高，就很有可能会裁剪掉正确的词，从而降低召回率，这一点将会在后面的实验中得到验证。

2.3 倒排索引

要达到快速检索语音文件的目的，需要对多候选识别结果建立索引。文本检索的相关研究表明，使用倒排索引结构可以有效地提升检索速度，在文本检索中倒排索引的索引项是词（汉语中还有可能是字，本文针对维吾尔语自身特点采用词作为索引项），每个词对应一系列的包含文档 ID 以及该词在文档中的位置信息的索引记录。但是由于语音识别结果与文本不同，识别结果中的每个词还包含时间信息和相应的得分，因此语音索引记录中除了包括索引项所在的所有文档编号以外还包含了起始时间、终止时间以及一些相应的得分信息。基于以上描述本文所采用的倒排索引结构如图 3 所示：

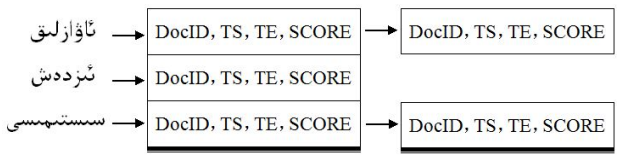

图 3 维吾尔语语音检索倒排索引结构图

图 2 表示了三个维吾尔语词“语音”、“检索”、“系统”（图中从上到下的顺序）的倒排索引结构，其中 DocID，TS，TE，SCORE 分别表示索引项所在文档编号、起始时间、终止时间和相应得分，在维吾尔语语音关键词检索中，当用户输入所要查询短语之后，系统会根据空格进行分词并删除停用词之后再查找查询词所在的文本文件和其对应的语音段。

2.4 置信度的计算

目前的研究中基本是以弧的后验概率作为置信度[11-12]的衡量标准。假设我们的查询词 Query 的音节串为 $l_1l_2...l_k$，后验概率记为 P(Q|O)，计算公式如下：

$$P(Q|O)=P((l_1l_2...l_k)|O)=P(A(l_1l_2...l_k)|O)$$

其中 $A(l_1l_2...l_k)$ 代表包含音节串 $l_1l_2...l_k$ 所有的路径的集合，公式具体推导过程见文献[11-12]。

2.5 相关度的计算

在计算相关度之前先将语音文档 D 分成若干个语音片段（Segment）$S_1,S_2,S_3...S_I$，当用户输入查询短语 Query（简写为 Q）时，查询短语通过分词并删除停用词等处理以后，被分成若干个 Word，分别记做 $W_1,W_2...W_J$，经过 2.3 节中对置信度的计算，能够计算出查询词在各个语音段发生的后验概率 $P(W_j|S_i)$（$1\le j\le J, 1\le i\le I$），之后便可得到查询短语 Q 和语音文档 D 的相关度计算公式如下：

$$SIM(D,Q)=\sum_{j=1}^{J}\sum_{i=1}^{I}P(W_j|S_i)$$

上述公式计算出查询词所发生的频率，索引时是依据计算得到的 SIM(D,Q)值来排序文档，因此可以看出查询词出现的频率越高，查询词与语音文档间的相关度就越大。

2.6 维吾尔语语音检索中的特殊问题

早期的语音检索研究主要针对英语而进行。随着语音检索技术的发展，针对一些其它语言（如汉语、阿拉伯语等）的语音检索技术也被越来越多的人所重视。与英文和汉语相比，维吾尔语有其自身特点，而这些特点也影响到了维吾尔语语音检索系统的设计与实现。

维吾尔语属于阿尔泰语系突厥语族，是黏着性语言，同一词干利用丰富的词缀可产生超大词汇量。就说明要建立覆盖维吾尔语中所有单词的发音词典有一定的难度，而且单词作为语音识别单元时，识别系统中会产生较多的未登录词(Out Of Vocabulary，OOV)，这会影响识别性能。那么在检

索系统无法识别的词时，检索结果就会出现较多的错误。目前 OOV 问题的主要解决方法就是对查询短语中的未登录词进行词干和词缀的切分[17], 然而维语词缀包含较多信息，该方法就会造成信息缺失，对于维吾尔语 OOV 问题还需更进一步的探索和研究。

3 实验及结果分析

3.1 实验配置

3.1.1 训练数据描述

实验中声学模型训练集采用的是 16khz 采样频率，16bit 量化精度，单声道，用 PC 在办公室环境下录制。训练语料包含 356 个人（189 女，167 男）发声的 128 小时的 2465 条语句。频谱特征观察矢量为每帧 39 维向量，包扩 12 阶 MFCC，归一化对数能量，及其一阶、二阶差分。

实验中语言模型训练集采用共有 1,335,000 个句子和 590,000 个不重复单词的维吾尔语文本语料库，内容包含新闻、杂志、政府公文、各种理工科书籍等同时对语料库中的句子以单词为单位进行反向处理，选取 60,000 个高频单词作为识别发音词典和语言模型建模基础单词列表，采用 SRILM[15] 语言模型训练工具分别建立了基于单词的正向 2-gram 和反向 3-gram 语言模型。

3.1.2 测试数据描述

识别阶段的测试语音库包含 10 个说话人（5 男，5 女）发声的 2 小时的约 1000 个语句的 wav 文件，测试集对语言模型的平均 OOV 率为 14.8%。在检索阶段，本文分别对 20 个维吾尔语关键词进行检索，其中 20 个关键词中有两个为集外词。

3.2 实验结果分析

3.2.1 系统性能评价

语音检索系统的性能评价分为语音识别模块性能的评价和语音检索模块性能的评价：语音识别模块的评价性能采用单词正确率进行评价，在 3.1.2 小节所介绍的测试集上该模块的单词正确率为 82.1%；语音检索的性能评价准则采用接收机工作特性（Receiver Operating Characteristics: ROC）曲线。ROC 曲线以虚警率为横轴，召回率为纵轴，绘制在改变阈值θ时检索系统的工作特性。在 2.2 节中词图转换为混淆网络过程中，阈值θ用于控制词图弧的裁减，当词图弧的置信度低于阈值θ时，就对该弧进行裁减则该弧将不参与混淆网络的转换即不参与建立索引。

3.2.2 实验结果及分析

如表 1 所示为 20 个关键词在不同裁剪阈值θ下的召回率和虚警率并根据表中结果给出召回率和虚警率的 ROC 曲线如图 3 所示：

表 1　不同阈值下系统的虚警率和召回率

阈值 θ	虚警率	召回率
-0.5	13.5%	97.0%
-0.05	12.5%	95.5%
-0.005	10.0%	92.5%
-0.0005	9.5%	89.6%
-0.0001	9.0%	86.6%
0	8.5%	79.1%

从表 1 中的结果和图 4 的 ROC 关系曲线可以看出当阈值θ增大时，相应的召回率和虚警率就会随之降低，这是由于当阈值增大时识别结果词图的弧的相应裁减量就会增多，那么识别结果对正确文本的覆盖率就会降低，自然检索出来的结果就会减少，最终就有可能会导致召回率和虚警率的降低。而在现实的应用当中，检索系统的召回率越高越好而相应的虚警率越低越好，因此如何根据不同的需求来选择阈值的大小使二者达到一个比较好的平衡是一个值得考虑的问题。

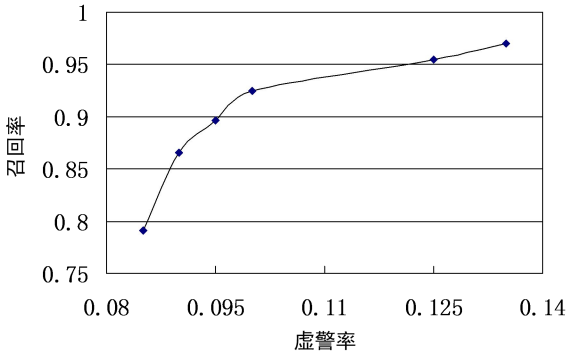

图 4　虚警率与召回率关系曲线（ROC）

4 结论与展望

本文研究并搭建了基于大词汇量维吾尔语连续语音识别技术的语音检索系统，该系统以维吾尔语连续语音识别系统作为识别模块将识别结果以词图格式输出并转化为混淆网络，最后根据混淆网络生成索引，用户在使用时输入查询串，系统先对查询串进行分词然后根据分词结果定位所要搜索

的语音文件。经过评测，发现该系统在识别正确率为 82.1%的情况下，当虚警率为 13.5%和 8.5%时，召回率分别为 97.0%和 79.1%，但是由于测试数据和查询关键词数量较少该评测数据仅仅只能提供一个参考，在具体应用过程中还需要另外讨论。除此之外，目前建立的维吾尔语语音和文本语料库规模还是比较小, 而且没有统一、共享的评测数据，无法对研究结果进行客观的评价也无法与其他系统进行对比实验，因此需要加强评测数据的建立和共享。

该维吾尔语语音检索系统是新疆多语种信息处理重点实验室研发的第一个语音检索系统，为实验室后期的研发奠定一个基础但仍有许多可以改进的地方如：

（1）训练语料库的扩展，可以收集各种各样的语料，覆盖更广的语音现象和更广的领域；

（2）对维吾尔语连续语音识别模块的声学模型和语言模型进一步优化以提高识别正确率（例如利用目前深度神经网络的方法进行优化）；

（3）将维吾尔语音的韵律特征与词图进行融合，充分的运用维吾尔语语言的特点，提高维吾尔语语音检索的性能；

（4）对集外词的问题要进一步深入研究；

（5）优化维吾尔语语音检索系统，界面更加友好，在视觉效果上尽可能适应用户需求，尽可能满足用户的需求。

参考文献

[1] A.Hauptmann,H.Wactlar.Indexing and Search of Multimodal Information[A].Proceedings of IEEE International Conference of Acoustics Speech and Signal Processing,Munich,Germany,1997[C]: 195-198P.

[2] 郑铁然，韩记庆，李海洋．基于词片的语言模型及在汉语语音检索中的应用[J]．通信学报．2009,30(3).

[3] G J.E Jones,J.T.Foote,K Sparck Jones,S.J.Young.Video mail retrieval:the effect of word spotting accuracy on precision[A].International Conference on Acoustics,Speech,and Signal Processing 1995[C]. ICASSP'95,1995,1(1):309-312P.

[4] GOOG-411[DB/OL],http://en.wikipedia.org/wiki/GOOG-411, 2008,12.

[5] Hsin-min Wang.Mandarin spoken document retrieval based on syllable Lattice matching[J].Pattem Recognition Letters.2000：615-624P.

[6] 郑铁然，韩纪庆．基于音节Lattice的汉语语音检索技术及其索引去冗余方法[J]．声学学报．2008，33(6)：526．533 页.

[7] 那斯尔江·吐尔逊,吾守尔·斯拉木．基于隐马尔可夫模型的维吾尔语连续语音识别系统[J]．计算机应用．2009,29(7).

[8] 木合塔尔·沙地克，李 晓，布合力齐姑丽·瓦斯力．维吾尔语广播新闻连续语音敏感词检索系统[J]．计算机系统应用．2012,21(3).

[9] L. Mangu, E. Brill, and A. Stolcke. Finding consensus in speech recognition: word error minimization and other applications of confusion networks[J]. Computer Speech And Language,2000,14:373–400P.

[10] Ville T.Turunen,Mikko Kurimo.Indexing confusion network for morph-based spoken document retrieval[A], SIGIR[C].2007,631-638p.

[11] F.K. Soong,W.K.Lo,and S.Nakamura.Generalized Word Posterior Probablity(GWPP) for Measuring Reliability of Recognized Words[A].Proceeding of SWIM2004[C],2004 :127-128P.

[12] F.Wessel,R.Schluter,K.Macherey,H.Ney.Confidence Maesures for Large Vocabulary Continuous Speech Recognition[A].IEEE Transactions on Speech and Audio Processing[C],2001,9(3):288-298P.

[13] 努尔麦麦提·尤鲁瓦斯，吾守尔·斯拉木．面向大词汇量的维吾尔语连续语音识别研究[J]．计算机工程与应用．2013,49(9):115 页.

[14] Young S.The HTK book[EB/OL].[2012-03-031].http://htk.eng.cam.ac.uk/.

[15] 陶梅，吾守尔·斯拉木，那斯尔江·吐尔逊．基于HTK的维吾尔语连续语音声学建模[J]冲文信息学报，2008，22(5)：56—59．

[16] Andreas Stolcke.SRILM—AN EXTENSIBLE LANGUAGE MODELING TOOLKIT.Speech Technology and Research Laboratory,SRI International, Menlo Park, CA, U.S.A.[EB/OL].[2004,07].http://www.speech.sri.com.

[17] 米成刚，王磊，杨雅婷，陈科海．维汉机器翻译未登录词识别研究[J]．计算机应用研究．2013,4,30(4).

作者简介：

作者一张力文（1991—），男，硕士研究生，主要研究领域为语音识别。Email:lwzhang9161@gmail.com；

作者二努尔麦麦提•尤鲁瓦斯（1980—），男，讲师，博士主要研究领域为自然语言处理，语音识别。
Email: y.nurmemet@gmail.com；

作者三吾守尔•斯拉木（1942—），男，院士，博导主要研究领域为多语种信息处理。Email:wushour@xju.edu.cn。