Open Journal of Circuits and Systems 电路与系统, 2015, 4, 15-21
Published Online March 2015 in Hans. http://www.hanspub.org/journal/ojcs
http://dx.doi.org/10.12677/ojcs.2015.41003

Hans汉斯

Implementation and Optimization of Interception System Based on Speech Retrieval

Huiting Ni, Xiaoqun Zhao

College of Electronic and Information Engineering, Tongji University, Shanghai
Email: niht90@gmail.com, zhao_xiaoqun@tongji.edu.cn

Received: Mar. 16th, 2015; accepted: Mar. 25th, 2015; published: Mar. 30th, 2015

Abstract

Based on the technology of speech retrieval and combined with the adaptive correlation filter module, an optimized interception system is constructed. The interception system based on speech retrieval is improved and optimized from two aspects. One is using context related initial-final as the acoustic model of keyword recognition. The other is putting forward a correlation judgment module to re-filter the speech documents that have nothing to do with the subject. Confirmed by the experimental data, the great enhancement of system precision can be obtained by small sacrifice of recall rate while using correlation judgment module.

Keywords

Cluster Analysis, Correlation Judgment, Keyword Recognition, Speech Retrieval, Interception

基于语音检索的侦听系统的实现与优化

倪慧婷，赵晓群

同济大学电信学院，上海
Email: niht90@gmail.com, zhao_xiaoqun@tongji.edu.cn

收稿日期：2015年3月16日；录用日期：2015年3月25日；发布日期：2015年3月30日

摘　要

本文以基于关键词识别的语音检索技术为基础，结合自适应的相关性过滤模块，构建了优化的侦听系统。对基于语音检索的侦听系统从两方面进行了改进和优化：一是以上下文相关的声韵母作为关键词识别的声学建模基元，二是添加相关性判决模块对与主题无关的语音进行再过滤。实验数据证实，相关性判决模块能以召回率的小部分牺牲换取系统查准率的大幅度提升。

关键词

聚类分析，相关性判决，关键词识别，语音检索，侦听

1. 引言

作为一种难于浏览和搜索的媒体类型，语音信息的有效提取工具尚比较少，资源再利用率很低。面对离线语音库存储的海量语音文档，亟需一种方法能快速有效地检索出其中符合需求的语音资源。

语音检索技术的研究可追溯至上世纪 60 年代，贝尔实验室开发了一个 10 个数字的英文语音检索系统。随着大词汇量连续语音识别技术的发展，语音检索技术有了长足的进展，不再局限于对离散语音的检索。目前已有的语音检索系统有剑桥大学的 Video Mail Retrieval Using Voice，Google 公司的 Google Voice Local Search 等。但相比于成熟的文本文档检索技术，语音检索技术的鲁棒性较差、准确率较低等问题还有待进一步深入研究。

2. 系统原理和结构

关键词识别技术作为语音识别的一个重要研究方向，其不同于连续语音识别之处在于，它并不试图将输入语流中的每个字词还原出来，而是在语音内容不受限的输入语流中只将使用者感兴趣的词辨认标记出来。由于用户对关键部分的发音通常都是清楚完整的，因此识别语音流中的关键词比识别语音流中的整个句子要容易得多。该项技术适用于只需要了解语音文档的部分关键信息就可解决问题的情景，与**语音检索**的需求正不谋而合。基于关键词识别的语音检索还可应用于军事安全和公共安全领域，作为情报获取、追踪的重要方法。

对于特殊的语音检索系统，如**侦听系统**，直接将仅经关键词识别过滤得到的语音文档作为检索结果输出，其中仍包含大量实际上与侦听主题无关的音频。本文针对此设计了优化策略，提出计分公式并结合文本聚类技术应用于关键词识别后的处理，以进一步过滤无关的语音文档，减少人工审听的工作量。

基于语音检索的侦听系统如图 1 所示。该系统分为关键词识别模块和相关性判决模块，其中关键词识别模块又分为离线训练部分和在线识别部分。

语音识别模块的各项技术相对较成熟。**离线训练部分**用于获取声学模型和语言模型，包含前端处理、特征计算等模块。前端处理模块用于连续语音做预加重、加窗、分段。特征计算模块用于提取语音帧的时域特征或频域特征。语音识别系统的识别率，很大一部分取决于提取出的特征对语音信号描述的准确性。**在线识别部分**首先由离线训练部分得到的声学模型、语言模型和即时定义的关键词表共同生成用作解码的搜索空间，再将待识别语音提取的特征输入关键词识别模块，采用 Viterbi 帧同步解码算法产生关键词的假想命中。后面的置信度计算模块将根据关键词候选结果和其它知识源计算这些假想命中的关键词的置信度，最终由置信度判决模块决策，给出识别结果。

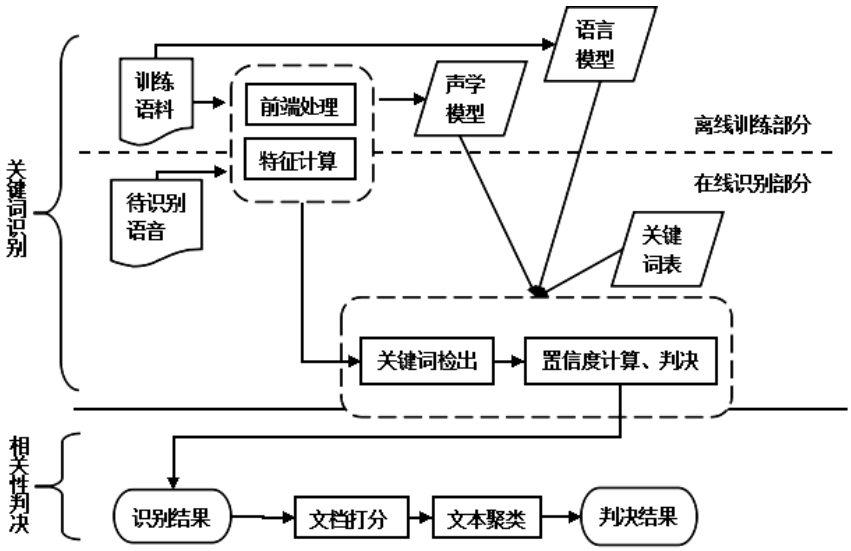

Figure 1. Structure diagram: the principle of system
图 1. 系统原理结构图

相关性判决模块为本文根据侦听系统特性而设计。在获取关键词识别结果后，将识别结果(包括关键词流及关键词的时间位置信息)进行打分，这个分数一方面用作提供给监控人员直观了解语音文档的相关性，另一方面用作文本聚类选择簇中心的依据。文档聚类后，将相关簇和无关簇分别输出至判决结果。

3. 关键词识别

在构建声学模型方面，采用文献[1]提出的多尺度声学模型建模思想，分前景模型和背景模型分别对关键词和非关键词进行声学建模。与之不同的是，文献[1]采用的声学建模单位为音素，本文经过实验并结合文献[2] [3]分析比较，最终根据汉语的发音特色采用声韵母作建模基元。

声学建模基元的选择决定了模型的精度和复杂度。选用较小的声学单元如音素，灵活性好，模型数量少(采用上下文无关策略建模，模型数量仅为 32 个)，但由于某些音素语音段长度过短，其识别效果很不稳定[4] [5]。选用较大的声学单元如音节，能很好地描述音节内部的发音变化，模型数量较多(采用上下文无关策略建模的有调音节模型数量为 1300 多个)，但由于音节作建模基元并不能描述汉语音节间普遍存在的协同发音现象，因此必须做上下文相关扩展，此时建模基元的数目将迅速增大，造成训练数据不足的问题。在这两种声学单元之外，声韵母作为适合汉语发音特点的一种单元，其模型数量适中(标准声韵母集合采用上下文无关策略建模时模型数量为 58 个)，上下文相关信息比较确定(如声母的上下文只可能是静音或韵母)，又有很多语言学知识可以利用来提高建模基元的性能。

本文的前景模型采用上下文相关的扩展声韵母基元，每个基元用自左向右无跳转的 3 状态 HMM 来描述，每个状态的混合高斯数为 32；背景模型则由上下文相关的扩展声韵母基元作聚类后产生，每个基元用自左向右无跳转的 3 状态 HMM 来描述，每个状态的混合高斯数为 16。较细致的前景模型和较粗略的背景模型结合在一起，构成了多尺度声学模型的系统。

4. 相关性判决

一方面，关键词识别模块过滤掉了一部分不含关键词的语音文档，但在含关键词的文档中仍有**大量同关注主题无关的语音文档**。以足球比赛为例，解说词中提到“球进了”“球传到了 1 号队员脚下”，这些语音是确实和足球比赛相关的；亲子教育音频中提到“红红的太阳像圆圆的球挂在天上”，这些语

音却是和足球比赛无关的。另一方面，关键词识别模块的输出为每个语音文档对应的关键词流，如表 1 所示，并**不直观**，需由监控人员根据个人经验借助语音时长、关键词出现次数等来判断被测音频与关注主题的相关程度，费时费力且其结果也不稳定。本文设计的相关性模块用于过滤那些含关键词但与关注主题无关的语音文档并对文档打分，为监控人员选择审听提供直观依据快速反应。

4.1. 计算关键词得分

将经关键词识别过滤后的语音文档记为 Set，对每篇文档按本文提出的式(1)计算关键词得分。此得分包含了关键词的集中度得分和绝对数量得分，能较好地反映文档属性。

$$\text{Score}=\begin{cases}\sum_{i=0}^{K}\lg\left[\left(F-\sum_{j=1}^{K}KW_j\right)\Big/GAP_i\right] & K\geq 1\\ 0 & K=0\end{cases} \tag{1}$$

式中 F 为语音文档的帧数，K 为关键词个数，KW_j 为每个关键词所占的帧数，GAP_i 为关键词与关键词之间的垃圾帧数。公式中，累加项 $\left(F-\sum KW_j\right)\Big/GAP_i$ 表征了关键词的**集中度**，其尺度通常较大，因此取对数缩小；累加操作对得分的影响体现了关键词的**绝对数量**对文档相关性的影响。

图 2 所示为四类不同的语音文档，横轴代表时间，灰色方块代表关键词。其中 A 类语音文档中关键词数量相对较少，但较集中；B 类语音文档中关键词相对较分散，但数量较多；C 类语音文档关键词数量少，且分布稀疏；D 类语音文档则没有关键词出现。

A、B、C、D 四段语音时长均为 25 s，以 25 ms 为一帧，得 F = 1000 帧。以帧为单位，其它数据如下：

$$\text{A}\begin{cases}GAP_0=100\\ GAP_1=GAP_2=20\\ GAP_3=200\\ GAP_4=50\\ GAP_5=300\\ KW_1=KW_2=KW_3=KW_4=KW_2=62\end{cases}\qquad \text{B}\begin{cases}GAP_0=GAP_7=30\\ GAP_1=GAP_4=GAP_6=50\\ GAP_2=GAP_3=GAP_5=100\\ KW_1=KW_2=KW_3=KW_4=KW_2=70\end{cases}$$

$$\text{C}\begin{cases}GAP_0=250\\ GAP_1=690\\ KW_1=60\end{cases}\qquad \text{D}\{GAP_0=1000$$

对如图 2 所示的四类语音文档打分。

$$\text{Score}_\text{A}=\sum_{i=0}^{5}\lg\left[\left(F-\sum_{j=1}^{5}KW_j\right)\Big/GAP_i\right]=5.628\qquad \text{Score}_\text{B}=\sum_{i=0}^{7}\lg\left[\left(F-\sum_{j=1}^{7}KW_j\right)\Big/GAP_i\right]=7.113$$

$$\text{Score}_\text{C}=\sum_{i=0}^{1}\lg\left[\left(F-KW_1\right)/GAP_i\right]=1.008\qquad \text{Score}_\text{D}=0$$

可见，A 和 B 的得分较接近，C 和 D 的得分较接近，AB 和 CD 之间的分差则很大。这一结果符合直观预期。后文将以此得分为依据对语音文档进行进一步处理。

4.2. 文档聚类

将文档按公式(1)得分排序，分别取得分最高和最低的文档作为聚类之初的中心。得分最高的文档代表了符合关注主题的文档集合的中心，得分最低的文档则代表其补集的中心。

Table 1. Results of keyword recognition
表 1. 关键词识别结果

检测音频文件名	内容
20131228_1.wav	会议 表决 会议 委员会……
20131228_2.wav	-
20131228_3.wav	表决 会议 人大 会议 委员会 会议 会议 会议 代表……
……	……

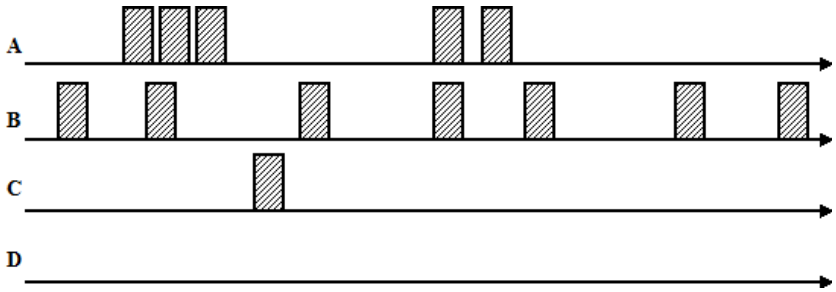

Figure 2. Keywords distribution map: different types of speech documents
图 2. 不同类型语音文档的关键词分布图

可选用的文本聚类算法很多，有基于划分的的 K 均值算法，基于层次的 BIRCH 算法、ROCK 算法、Chameleon 算法，基于密度的 DB-SCAN 算法和基于网格的 STING 算法、CLIQUE 算法、WAVE-CLUSTER 算法等[6]。本文采用基于划分的 K 最邻近(K-Nearest Neighbor，KNN)聚类算法对文档进行分类。KNN 分类算法的基本思想是：当判断一个新的对象类别时，从现有训练集合中寻找最优对象，根据最优对象的类别来判断新对象的类别[7]。

文档聚类中，判断聚合方向的依据为文档相似度，常用文档特征向量夹角的余弦值表示[6]，如下：

$$Sim(D_1, D_2) = \mathrm{Cos}\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)}} \tag{2}$$

公式中 W_{1k}， W_{2k} 分别表示文档 D_1 和 D_2 第 k 个特征项的权值，本文中特征项即为关键词识别模块里选择的关键词。$Sim(D_1, D_2)$ 的值越大，则说明两个文档越相似，越可能归为一簇；反之，则说明越不相似，越可能划分到不同的簇。

5. 实验结果及分析

测试集为 120 篇 16 kHz，16 bit 采样的新闻类语音文档，选取与人民代表大会(后文简称人代会)相关的 15 个词作为关键词。将测试集 Set 中包含的三类语音文档以符号标记，Y 为含关键词且符合关注主题的语音文档，N 为含关键词但不符合关注主题的语音文档， CYN 为 Y 和 N 的合集即含关键词的语音文档，O 类为不含关键词的语音文档。经人工确认，在测试集中有 52 篇人代会相关报道，包含至少 1 个关键词的有 81 篇，即：

$$Y_{set} = 52, N_{set} = 29, O_{set} = 39$$

实验结果将以召回率(Recall)、查准率(Precision)两个指标来评价。召回率是指对于某个类别，被系统正确划分到这个类别中的文档数量与这个类别实际包含的文档数量的比值；查准率是指被系统正确划分

到这个类别中的文档数量与所有被系统划分到这个类别中的文档数量的比值[8]。两个指标的计算公式如下：

$$\text{Recall} = \frac{a}{a+b} \tag{3}$$

$$\text{Precision} = \frac{a}{a+c} \tag{4}$$

公式(3) (4)中，a 表示被正确划分到所属类别的文档数量，b 表示属于该类别但被错误划分为不属于的文档数量，c 表示被错误划分到该类别中的文档数量。在本文实验中，将以上两个指标再细分为关键词文档召回率、关键词文档查准率、关注文档召回率、关注文档查准率，其计算公式如下：

$$\text{Recall}_{\text{Keyword}} = \frac{CYN_{\text{System\&Set}}}{CYN_{\text{Set}}} \tag{5}$$

$$\text{Precision}_{\text{Keyword}} = \frac{CYN_{\text{System\&Set}}}{CYN_{\text{System}}} \tag{6}$$

$$\text{Recall}_{\text{Relation}} = \frac{Y_{\text{System\&Set}}}{Y_{\text{Set}}} \tag{7}$$

$$\text{Precision}_{\text{Relation}} = \frac{Y_{\text{System\&Set}}}{Y_{\text{System}}} \tag{8}$$

公式(5)~(8)中，CYN_{System} 表示经关键词识别模块(KWR)判定为含关键词的文档数量，CYN_{Set} 表示测试集中含关键词的文档数量，$CYN_{\text{System\&Set}}$ 表示被 KWR 判定为含关键词且确实含关键词的文档数量；Y_{System} 表示经 KWR 或相关性判决模块(CJ)判定为关注文档的数量，Y_{Set} 表示测试集中的关注文档的数量，$Y_{\text{System\&Set}}$ 表示被 KWR 或 CJ 判定为关注文档的数量。需要特别说明的是，对 KWR 系统，由于含关键词即被判定为相关，因此下式成立：

$$Y_{\text{System}} = CYN_{\text{System}} \tag{9}$$

实验中，经关键词识别模块识别后，检出的关键词流如表 1 所示。

如表 1 所示，一些语音文档含有部分关键词，一些则不含任何关键词。关键词识别模块可过滤如后者这样的无关文档。该模块结果显示，在 79 篇有关键词识别输出的文档中，Y 类 52 篇，N 类 23 篇，O 类 4 篇。此时系统对含关键词的文档查准率较高(94.94%)，但对关注语音文档的查准率仅为 65.82%。关键词识别过滤得到的文档中仍有将近半数是无关语音文档，这表明仅以关键词识别模块过滤，仍将浪费许多人力物力。

将识别结果的关键词流及时间信息送交相关性判决模块判决，所得结果如表 2 所示。

Table 2. Results of correlation judgment
表 2. 相关性判决结果

检测音频文件名	相关程度得分	归簇	检测音频文件名	相关程度得分	归簇
20131228_1.wav	9.65	相关	20131228_5.wav	0	无关
20131228_2.wav	0	无关	20131229_1.wav	0	无关
20131228_3.wav	16.31	相关	20131229_2.wav	1.13	无关
20131228_4.wav	0	无关	……	……	……

在表 2 中显示，相关簇的文档数量为 56 篇，其中 Y 类 51 篇，N 类 11 篇，O 类 0 篇。此时系统对关注语音文档的召回率为 98.08%，查准率高达 82.26%。本实验中，查准率的显著提升主要是因为测试集中含关键词的无关语音文档数量较多。

6. 结语

增加了相关性判决模块的侦听系统较原系统相比，其对关注语音文档的召回率略有减少(不到 2%)，但查准率却明显增加(超过 16%)。如果测试集中含关键词的无关语音文档的比重加大，则对关注语音的查准率会有更好表现。以牺牲一小部分召回率换取查准率的大幅提升，对减轻监控人员的工作量非常有益。同时由于关注语音文档的比重增加，监控人员审听无关语音文档的时间比重减少，侦听系统的实时性和可靠性得到显著改善。此即本文设计的优化策略的意义所在。

参考文献 (References)

[1] 韩疆, 刘晓星, 颜永红, 张鹏远, 潘接林 (2005) 一种任务域无关的语音关键词检测系统. *全国网络与信息安全技术研讨会*2005 *论文集*(*下册*), 信息产业部互联网应急处理协调办公室, 7.

[2] 李曜, 刘加 (2007) 基于汉语语音学决策树构建语音识别声学建模方法研究. *全国网络与信息安全技术研讨会论文集*(*下册*), 信息产业部互联网应急处理协调办公室, 6.

[3] 李净, 徐明星, 张继勇, 郑方, 吴文虎, 方棣棠 (2001) 汉语连续语音识别中声学模型基元比较：音节、音素、声韵母. *第六届全国人机语音通讯学术会议论文集*, 5.

[4] Lee, C., Rabiner, L., Pieraccini, R. and Wilpon, J. (1990) Acoustic modeling for large vocabulary speech recognition. *Computer Speech and Language*, **4**, 127-165.

[5] Youngand, S.J. and Woodland, P.C. (1994) Tree-based state tying for high accuracy acoustic modeling. *Proceedings of Human Language Technology Workshop*, March 1994, 307-312.

[6] 李春青 (2015) 文本聚类算法研究. *软件导刊*, **01**, 74-76.

[7] 孙建旺, 吕学强, 张雷瀚 (2013) 基于语义与最大匹配度的短文本分类研究. *计算机工程与设计*, **10**, 3613-3618.

[8] 熊大康 (2014) 中文短文本分类技术的研究与实现. 安徽大学, 合肥.