用MATLAB实现维吾尔语广播新闻敏感词检索系统①

布合力齐姑丽•瓦斯力[1], 木合塔尔•沙地克[2], 木特力甫•马木提[3], 李 晓[4]

[1](新疆教育学院数学与信息技术分院, 乌鲁木齐 830043)
[2](新疆教育管理信息中心, 乌鲁木齐 830049)
[3](新疆大学图书馆, 乌鲁木齐 830046)
[4](中国科学院新疆理化技术研究所, 乌鲁木齐 830011)

摘 要: 首先, 用 MATLAB 开发一个敏感词检索系统; 然后, 用该系统对语音信号来自于新疆广播电台网站的维吾尔语新闻 60 分节目语音进行连续敏感词检索; 最后, 对识别结果进行分析并提出提高正识率的思路.

关键词: 语音识别; 关键词识别; 敏感词检索; 维吾尔语; 广播新闻; MATLAB

Implementation of Uyghur Broadcast News Sensitive-word Spotting System on MATLAB

Buheliqiguli Wasili[1], Muhetaer Shadike[2], Mutelifu Mamuti[3], LI Xiao[4]

[1](Xinjiang Education Institute, Urumchi 830043, China)
[2](Xinjiang Education Management Information Center, Urumchi 830049, China)
[3](Xinjiang University Library, Urumchi 830046, China)
[4](Xinjiang Technical Institute of Physics & Chemistry, Urumchi 830011, China)

Abstract: This paper introduces the implementation of sensitive-word spotting system uses MATLAB. This system implements Uyghur broadcast news continues speech Sensitive-word spotting and give the conclusion of the test. After all, this paper gives some advices for this system.

Keywords: speech recognition; key-word spotting; sensitive-word spotting; uyghur; broadcast news; MATLAB

维吾尔语是维吾尔族使用的语言, 在公务活动、社会交际、广播影视、新闻出版、文学艺术、民族教育、科技等各个领域都普遍使用. 现代维吾尔语有 8 个元音, 24 个辅音. 有元音和谐律. 舌位和谐比较严整, 唇状和谐比较松弛. 有元音弱化现象. 构词和构形附加成分很丰富. 名词有数、从属人称、格等语法范畴; 动词有态、肯定否定、语气、时、人称、数、形动词、动名词、副动词等语法范畴. 表示各种情态的动词很发达. 词组和句子有严格的词序: 主语在谓语之前, 限定语在中心词之前. 词汇中除有突厥语族诸语言的共同词外, 还有相当数量的汉语、阿拉伯语、波斯语和俄语的借词[1].

敏感词检索的任务是在给定语音信号 S 时, 找出敏感词 W, 使得 P(W|S)最大, 也就是说, W 最可能是 S 所传达的敏感词:

$$\hat{S} = \arg\max_{s} P(W|S)$$

根据贝叶斯法则

$$\hat{S} = \arg\max_{s} \frac{P(S|W)P(W)}{P(S)}$$

P(S)在 S 给定时是归一化常数, 因而在求时可忽略不计:

$$\hat{S} = \arg\max_{s} P(S|W)P(W)$$

因此敏感词检索系统通常由两个模型组成, 统计语音模型(如HMM)给出了从敏感词W产生语音信号S的概率, 语言模型给出在一个语言中产生敏感词 S 的概率. 敏感词检索就是搜索 S, 使语音模型和语言模型的概率的乘积 P(S|W)P(W)最大[2].

1 系统原理

1.1 识别单元的选取

选择识别单元是语音识别应用的第一步. 语音识

① 基金项目:自治区高校科研计划项目(XJEDU2012S46);新疆多语种信息技术重点实验室项目(049807)
收稿时间:2013-10-23;收到修改稿时间:2013-11-18

别单元包括单词、音节和音素 3 种, 实际中根据具体应用来选取[3]. 本文用单词作为识别单元, 因为连续语音句子中各单词之间不存在明显的边界, 我们实现用两种切词算法对连续语音进行切词.

(1)基于分帧原理的等宽切词算法:

function dkseg(s,Nt,deltaN), 其中, s 为广播新闻语音文件, Nt 为敏感词长度, deltaN 为词移.

(2)基于单词清浊音结构的切词算法[4]:

function vuseg(s,Nt,errTol,vuStruc), 其中, s 为广播新闻语音文件, Nt 为敏感词长度, m_mgcvu 为敏感词清浊音结构, N 为帧长.

本文所有函数代码及说明请看参考文献[5].

1.2 特征提取

对观测的语音信号来说, 可以有很多不同的特征, 包括时域和频域的. 在语音识别中, 常用的方法是利用线性预测编码(LPC)对语音信号进行特征分析. 本文先把语音信号分成长度为 N 的帧, 相邻帧间的帧移为 deltaN. 给每个帧乘 Hamming 窗并进行基于 Levinson-Durbin 递归的 M 阶 LP 分析. 然后用半个正弦波窗对 LPC 系数进行加权处理转成 Q 个倒谱系数. 得到的结果中, 前半是观测的特征矢量, 后半是用于增加动态信息的倒谱系数差. 最后输出结果为 2Q*T 的矩阵, 其中 T 为帧数. 用 LP 参数提取帧的特征矢量序列, 每个帧由 12 个倒谱系数和 12 个分差倒谱系数构成. 特征提取函数定义如下:

function y = hmmfeatures(s,N,deltaN,M,Q)

函数以帧为单位, 对语音信号 s 进行分析, 返回用于 HMM 训练的观测矢量 y.

1.3 矢量量化

为了应用离散概率密度型的 HMM, 需要对观测的特征矢量进行矢量量化, 它的作用是产生一个包含 K 个可能的观测矢量的码本. 本文用 K-均值聚类方法进行矢量量化. 矩阵 Y 代表要聚类的数据, 其中每一行对应一个矢量, K 是聚类的期望值, maxiter 是最大的迭代数, 用 Euclidean 距离的平方对 Y 中的每个矢量进行聚类. 生成码本矢量函数定义如下:

function [cb,K,T,dist] =

hmmcodebook(data,N,deltaN,M,Q,Kmax)

函数用基元数组格式定义的训练序列生成包含特征矢量的原型的码本 cb. 基元数组的长度表示单词数, 每个基元又是一个新的基元数组, 是该单词的文件名. K 是所有序列的特征矢量被连接起来后矢量量化(聚类)K<Kmax 的特征矢量原型, T 是用于生成码本的观测矢量的数目, dist 是 VQ 平均失真.

1.4 模型训练

模型训练是根据一定的准则, 从大量已知模式中获取表征该模式特征的模型参数. 本文用向前向后算法(F-B 算法)来进行模型训练. 训练序列是以基元数组数据格式定义的. 对每一个序列, 进行基于帧的分析, 得到观测矢量, 然后将这些矢量量化成由 cb 给定的可能的码本矢量. 然后使用 F-B 重新估计算法对每个单词(量化)的观测矢量进行 HMM 训练. S 是 HMM 模型状态数. maxiter(训练最大迭代数)和 tol(即 tolerance, 绝对误差限)是向前向后算法的停止准则, 例如, 5000 和 1e-3. 训练函数定义如下:

function [A_m,B_m,pi_m,loglike_m] =

hmmtrain(data,N,deltaN,M,Q,cb,S,maxiter,tol)

HMM 模型的输出概率返回在基元数组 A_m, B_m, pi_m, 和 loglike_m 中.

1.5 语音识别

语音识别是根据一定的准则, 使未知模式与模型库中的某一模型获得最佳匹配. 本文对每个模型进行对数似然估计来进行敏感词识别. 基元数组中存放要识别单词的文件名. 对每个单词, 进行基于帧的分析, 得出观测矢量. 然后对观测矢量进行矢量量化, 生成可能的码本 cb. 最后用得出的观测序列和 HMM 概率计算对数似然, 对数似然为最高的单词视为识别的单词. 这个过程循环到全部单词. 识别函数定义如下:

function [logp,guess] =

hmmrecog(data,A_m,B_m,pi_m,cb,N,deltaN,M,Q)

logp 是对数似然, guess 是识别的敏感词序号.

2 系统设计与实现

本系统是用 Matlab 2011 开发的. 用 Matlab 的 GUIDE 工具设计系统用户界面[6], 其运行界面如图 1 所示. 详细代码请看参考文献[5,7]. 系统简单操作步骤和重要代码段如下:

首先, 打开一个或多个语音文件(本文语音文件来自于新疆广播电台新闻在线网站). 系统会自动计算语音文件频率, 样点数和时长.

(1)打开一个或多个语音文件函数代码如下:

[FileName, PathName]=

```
uigetfile({'*.wav', 'WAV files(*.wav)'},'选择一个或多个(按住 Ctrl 或 Shift)语音文件','MultiSelect', 'on');
```

(2)本文语音语料库语音格式为 wav 格式, 采样率为 8000Hz, 位数为 16 位, 声道数为单声道, 所以对打开的语音文件的这些参数进行判断, 如果与语料库格式不符, 提示用户并停止操作, 其代码如下:

```
m_file=[PathName FileName];
[x, Fn, nbits, opts]=wavread(m_file);
if Fn~=8000 || nbits~=16 || opts.fmt. nChannels~=1
    msgbox('语音文件格式须为: 采样率 8000Hz, 16 位, 单声道！','消息','warn','modal')
    return
end
```

(3)在轴对象上画出打开的语音文件波形图代码如下:

```
plot(handles.ax_plot2,x,'c');
title(handles.ax_plot2,['语音文件(' num2str(m_file_count) '个)'])
```

然后, 选择检索的敏感词(也可以全选), 选择二次确认方式(取中位数、取模数、取平均值等三种)和输出方式(输出到 Excel), 选择切词方式(切词方式有两种, 一种是基于清浊音结构的切词, 另一种是等宽切词. 若选择基于清浊音结构的切词方式, 则要输入切词长度容错倍数, 默认为 2. 等宽切词方式的词长可以设为要搜索敏感词的所有发音长度平均值或某个发音的长度, 等宽切词方式词移默认值为 80).

最后, 设定重复识别次数和其它相关参数后开始检索. 如, 帧长: 默认值为 200；帧移: 默认值为 80；LP 阶次: 默认值为 12；到谱系数: 默认值为 12；码书矢量数: 默认值为 16；HMM 状态数: 默认值为 5；训练最大迭代数: 默认值为 5000；绝对误差限: 默认值为 1e-3.

(1)切词算法的选择代码段如下:

```
%打开敏感词语音语料库和语音文件
load mgc_train;
x=wavread(get(handles.ed_yuyin,'String'));
%获取敏感词清浊音结构
m_mgcvu=mgc.phoneme{i};
%计算敏感词长度
m_mgcL=mgc.label{i};
m_mgcC=mgc.case{i};
m_no=length(m_mgcC);
nt=0;
    for j=1:m_no
    t=load(m_mgcC{j});
    nt=nt+length(t);
end
nt=round(nt/m_no);
if m_segstyle==1 %VU 切词
    m_rongcuo=
    str2double(get(handles.ed_qiecirongcuobei,'string');
    vuseg(x,nt,m_rongcuo,m_mgcvu);
else %等宽切词
    m_ciyi=str2double(get(handles.ed_ciyi,'String'));
    dengk(x,nt,m_ciyi);
end
```

(2)矢量量化、训练、识别函数调用代码段如下:

```
%打开语音文件并获取相关参数
x=wavread(get(handles.ed_yuyin,'String'));
N=str2double(get(handles.ed_zhenchang,'String'));
deltaN=str2double(get(handles.ed_zhenyi,'String'));
M=str2double(get(handles.ed_jieci,'String'));
Q=str2double(get(handles.ed_daopuxishu,'String'));
Kmax=str2double(get(handles.ed_mashushiliangshu,'String'));
S=str2double(get(handles.ed_zhuangtaishu,'String'));
maxiter=str2double(get(handles.ed_maxiter,'String'));
tol=str2double(get(handles.ed_tol,'String'));
%生成码本矢量
[cb,K,T,dist]=hmmcodebook(train,N,deltaN,M,Q,Kmax);
%模型训练
[A_m, B_m, pi_m, loglike_m]=hmmtrain(train, N, deltaN, M, Q, cb, S, maxiter, tol);
load seg_test; %打开语音文件切词结果
data_test=test.case;
%敏感词识别
[logp,guess1]=hmmrecog(data_test,A_m,B_m,pi_m,cb,N,deltaN,M,Q);
```

(3)确定二次确认方式代码段如下:

```
m_moshi1=median(m_yuce);
m_moshi2=mode(m_yuce);
m_moshi3=round(mean(m_yuce));
```

```
m_moshi=get(handles.ed_jiansuomoshi_value,'string');
switch m_moshi
        case '1'
            m_queren=m_moshi1;
        case '2'
             m_queren=m_moshi2;
        case '3'
             m_queren=m_moshi3;
end
```

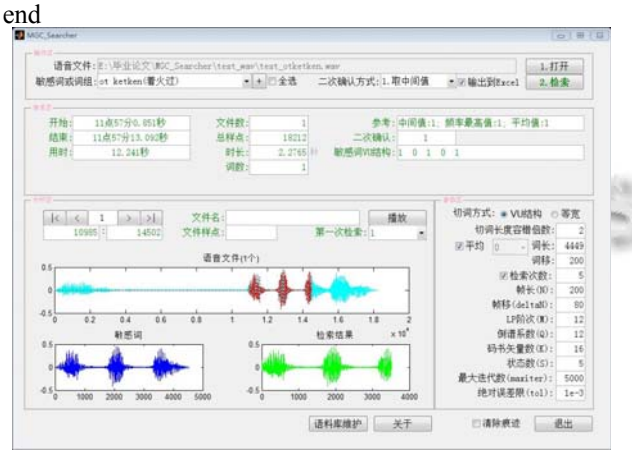

图 1　系统运行界面

检索结果是第一次估计词序列, 第二次确认结果和三种第二次确认方式参考结果并显示敏感词和检索结果波形图. 如果选择“输出到 Excel”选项, 可以把检索结果存放在 Excel 表里面. Excel 表存放结果如表 1 所示(检索敏感词为“ot ketken”, 重复识别次数为 5). 我们可以看到, Excel 表中的信息比程序界面显示的信息多一点, 不仅包含第一次估计词序列、第二次确认结果和三种确认方式参考结果外, 还包含每次估计词的对数似然、开始样点和结束样点和开始时间和结束时间.

3　实验结果及分析

我们用本系统对语音信号来自于新疆维吾尔广播电台网站的维吾尔语新闻 60 分节目语音进行连续敏感词检索. 首先, 从新闻 60 分节目连续语音中手工截取同一个播音员、在不同的语音段发音的敏感词, 共截取 8 个敏感词, 每个敏感词重复发音两次(其中有的只有发音一次), 建立训练语音语料库, 语料库建立详细过程请看参考文献[3,5]. 然后, 对同一个播音员的含有敏感词的另外新闻 60 分节目的连续语音(不是截取训练敏感词的语音段)进行敏感词检索. 实验结果如表 2 所示.

从表 2 我们可以看到, 第二次确认方式为“取中位数”的正识率为 75%, “取模数”的正识率为 37.5%, “取平均值”的正识率为 62.5%. 所有第二次确认方式的正识率为 37.5%. 另外, “取中位数”和“取平均值”的误识均能找到敏感词边缘.

我们又发现, 有的误识结果波形与敏感词波形倒反, 比如敏感词“yarlinish”的第三次识别结果波形(词序为 59)与敏感词波形倒反, 如图 2 所示.

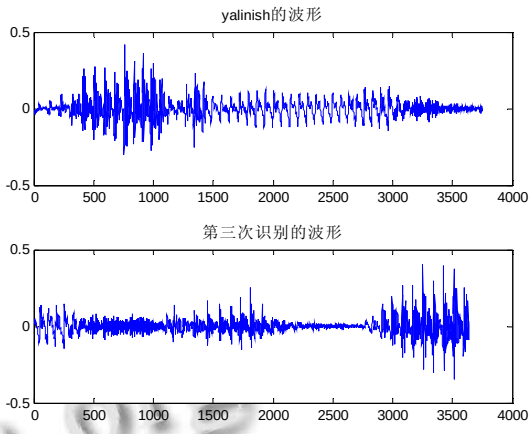

图 2　倒反识别波形

虽然, 目前的实验环境下, 系统呈现较高的识别率, 但是为了更加提高系统的实用性, 一、尽量选择较长的敏感词; 二、选择合适的确认方式要进行第三次确认; 三、程序中阻止识别结果中出现倒反情况; 四、充分利用对数似然提高识别率.

表 1 检索 Ecxel 存放结果

	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	中位	模数	平均	确认
估计词序	136	137	129	157	136	136	136	139	136
对数似然	47.5701	45.9061	40.6629	47.6672	33.6457				
开始样点	10801	10881	10241	12481	10801				
结束样点	15250	15330	14690	16930	15250				
开始时间	1.350125	1.360125	1.280125	1.560125	1.350125				
结束时间	1.90625	1.91625	1.83625	2.11625	1.90625				

表 2 实验结果

敏感词	数据	重复识别次数					第二次确认方式			结果
		第 1 次	第 2 次	第 3 次	第 4 次	第 5 次	中位	模数	平均	
ot	估计	13	27	32	27	27	27	27	25	27
	似然	10.122	7.4673	9.64129	8.6221	7.5708				
ot ketti	估计	105	102	106	65	106	105	106	97	105
	似然	41.014	43.311	43.5636	52.091	41.192				
ot ketken	估计	130	136	136	130	136	136	136	134	136
	似然	40.847	47.081	41.7652	39.883	46.102				
olush	估计	17	93	17	67	94	67	17	58	67
	似然	26.371	29.375	25.4329	30.285	22.815				
yarlinish	估计	13	27	59	12	13	13	13	25	25
	似然	32.904	33.951	26.7553	32.961	30.403				
koyup	估计	17	53	94	44	49	49	17	51	49
	似然	9.2657	12.063	7.0714	10.076	12.628				
orulup	估计	69	74	5	24	5	24	5	35	35
	似然	20.714	18.852	11.8734	13.317	8.9905				
orulup chushken	估计	28	27	4	3	25	25	3	17	25
	似然	61.203	41.672	52.5087	52.729	48.437				

参考文献

1 哈力克•尼亚孜.基础维吾尔语.乌鲁木齐:新疆大学出版社,1997:1–56.

2 翁福良,王野翊.计算语言学导论.北京:中国社会科学出版社,1998:122–136.

3 王炳锡,屈丹,彭煊等.实用语音识别基础.北京:国防工业出版社,2005:26–127.

4 木合塔尔•沙地克,布合力齐姑丽•瓦斯力,李晓.基于维吾尔语单词清、浊音组成结构特征的连续语音单词切分算法.西北师范大学学报(自然科学版),2013, 49(4): 34–37.

5 木合塔尔•沙地克.维吾尔语广播新闻敏感词检索系统的研究[博士学位论文],北京:中国科学院大学,2013.

6 刘焕进,王辉,李鹏等.Matlab N 个实用技巧.北京:北京航空航天大学出版社,2011.

7 木合塔尔•沙地克,李晓,布合力齐姑丽•瓦斯力.维吾尔语广播新闻敏感词检索系统的研究.中文信息学报,2011,25(4): 3–10.