文章编号:1002-8684(2007)02-0051-03

MATLAB环境下的语音识别系统

·论文·

杨 熙,苏 娟,赵 鹏
(湖南大学 电气与信息工程学院,湖南 长沙 410082)

【摘 要】介绍了MATLAB环境下的语音识别系统,阐述了具体的实现过程。采用离散隐马尔科夫模型,为提高识别率采用男女2套参数,对离散隐马尔科夫模型在实际语音识别系统中遇到的问题进行分析,并给出相应的解决办法。

【关键词】MATLAB; 语音识别; 离散隐马尔科夫模型
【中图分类号】TN912 【文献标识码】A

Speech Recognition in MATLAB Environment

YANG Xi, SU Juan, ZHAO Peng
(Institute of Electronics & Information Engineering, Hunan University, Changsha 410082, China)

【Abstract】A digital speech recognition system and its realization course are introduced in the MATLAB environment. The DHMM(Discrete Hidden Markov Model) and two sets of parameters for male and female are used to improve the recognition accuracy. Some problems when DHMM is used in speech recognition system are analyzed and the corresponding solutions are provided.

【Key words】MATLAB; speech recognition; DHMM

1 引言

MATLAB的最大特点是它的数据类型只有一种,即矩阵。它将所有的数据都处理成矩阵,用户不必定义变量和数据类型,且矩阵的大小也可任意改变。

数字化的语音信号可作为一维或二维(双声道立体声数据)矩阵来处理,因此MATLAB很自然地应用到语音处理领域。

2 系统设计

2.1 端点检测

MATLAB本身提供了一定的音频处理能力,如“wavread’函数用来读取语音文件,“soundview’能实现可视化语音输出,“wavrecord’实现录音,这些函数的具体应用可参照相应帮助文件。如果希望在MATLAB环境中实现实时语音信号处理,以上函数就不太适合了,可用ActiveX控件来实现,通常是将它嵌入到GUI界面中的。

系统的录音环境为普通办公室,端点检测算法采用基于短时能量和过零率的双门限法。短时能量和过零率分别有2个门限值:低门限和高门限。另有语音时间限值和最大静音时间限值,前者用来去除突发性噪声,后者防止漏捡。

计算短时能量之前,要先通过一个一阶高通滤波器$1-0.937\,5z^{-1}$,主要用来滤除低频干扰,尤其是50 Hz或60 Hz的工频干扰,还可起到消除直流漂移、抑制随机噪声和提升轻音部分短时能量的效果[1]。MATLAB中可用一行语句来实现:amp=sum(abs(enframe(filter([1-0.937 5],1,x),Len,Inc)),2),其中Len为帧长,Inc为帧移。需要注意的是,用MATLAB进行语音处理时,voice box是个十分有用的工具箱,需要从网上下载并添加到MATLAB搜索路径中。图1为连续语音的端点检测结果。

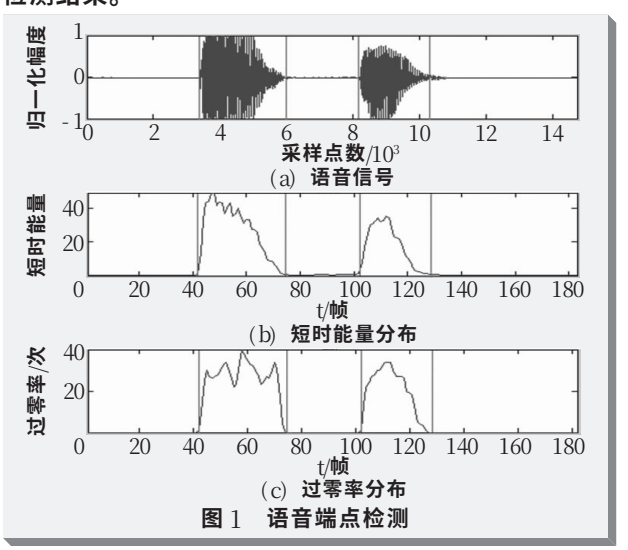

图1 语音端点检测

2.2 特征参数提取

特征参数提取常用2种参数：线性预测倒谱系数(Linear Prediction Cepstral Coefficients, LPCC)和Mel频率倒谱系数(Mel-Frequency Cepstral Cofficients, MFCC)。MFCC在有信道噪声的情况下，能提高识别率，但运算量大。由于录音环境并不恶劣，而LPCC计算非常简便且用于语音识别时效果很好，因此这里采用LPCC。具体计算过程可参考文献[1]。系统中选用的倒谱参数实际上是复倒谱参数。对P个倒谱系数进行窗函数加权可明显改善识别效果[2]，窗函数表示为

$$W(m)=1+(P/2)\sin(m\pi/P),\quad m=1,\cdots,P \tag{1}$$

设$W_n(m)$是第n帧加权窗函数，第n帧的加权倒谱系数表示为

$$\hat{C}_n(m)=W_n(m)C_n(m),\quad m=1,\cdots,P \tag{2}$$

LPCC参数只反映了语音参数的静态特性，而人耳对语音的动态特征更为敏感，为进一步提高识别率，在加权倒谱系数后再增加P个差分倒谱分量

$$\Delta\hat{C}_n(m)=\left[\sum_{k=-K}^{K}k\hat{C}_{n-k}(m)\right]G,\quad m=1,\cdots,P \tag{3}$$

其中，K为方差范围；G为加权系数。实验表明[2]，取K=2，G=0.375。这样就形成了一个2P维的特征矢量。采用P=12，即12维的LPCC和12维的离散线性预测倒谱参数(Discrete Linear Prediction Cepstral Coefficients, DLPCC)组成24维的语音特征参数向量。

2.3 离散马尔科夫模型[1-5]

隐马尔科夫(Hidden Markov Model, HMM)是对语音信号时间序列结构建立统计模型，它是数学上的双重随机过程，能很好地描述语音信号的整体非平稳性和局部平稳性，是一种较为理想的语音信号模型。

假设允许出现的状态为L种，记为$S_l(l=1,\cdots,L)$；记n时刻模型所处的状态为x_n，显然$x_n\in(S_1,\cdots,S_l)$，$\forall n$；若每个运行过程只完成(N-1)状态转移，那么产生的一条有限长度马尔科夫链$x_1,x_2,\cdots,x_N$可用行矢量表示为$X=[x_1,x_2,\cdots,x_N]$；初始状态概率矢量π是一个L维矢量，它的每一个分量π_l表示x_1等于S_l的概率，这可用$\pi_l=p_r[x_1=S_l]\ (l=1,\cdots,L)$表示；矩阵A是状态转移概率的集合，每一个元素用$A_{ij}$表示；B为输出观测值概率的集合；Y为输出序列矢量，$Y=[y_1,\cdots,y_N]$，系统产生任意一个Y的概率记为$P_Y[\pi,A,B]$，可用$P_Y[\pi,A,B]=\sum_X p_r[X]\left[\prod_{n=1}^{N}p_{x_n=S_l}[y_n]\right]$来计算，其中$p_r[X]$表示对认可一特定X出现的概率，$p_r[X]=\pi_{x_1}A_{x_1x_2}A_{x_2x_3},\cdots,A_{x_{N-1}x_N}$，$\sum_X$表示对所有可能出现X进行求和。一个离散隐马尔科夫模型(Discrete Hidden Markov Model, DHMM)[1-5]系统可用π，A，B三项参数来描述。

HMM系统的随机输出矢量y_n具有离散概率分布时，y_n只能取有限多个离散分布矢量中的某一个。假设对这些离散矢量赋予标号j且矢量总数为J，那么可用这些标号来表示每一个输出。标号j的总数J取决于码本的容量，目前码本的容量一般取64，128或256，这里选取J=128。假设在状态S_l时，y_n具有标号j的概率用b_{lj}表示，则构成一个(L×J)维矩阵B，它的第l行第j列元素为b_{lj}。

采用DHMM模型，选取无跨越从左到右模型，模型状态数设为6，在这一模型中状态转移矩阵A中只有主对角元素A_{ii}和右副对角元素A_{ii+1}允许非零，这符合人的语音特点，而且A比较稀疏，因此大大减少了模型参数估值的计算量。

2.3.1 DHMM的3个基本问题

对于DHMM模型$\lambda=[\pi,A,B]$，有3个基本问题需要解决：(1) 若已知HMM系统的3项特征参数，针对系统可能产生的任何Y计算$P_Y[\pi,A,B]$；(2) 已知3项特征参数，若得到了此系统产生的某个Y，估计该系统产生此Y时最可能经历的状态序列X；(3) 若有1个HMM系统，系统根据所给的若干输出Y来确定其特征参数，而这些参数要使系统产生学习样本集合中各个样本的概率平均值达到最大。

2.3.2 3个问题的解决

(1) 第1个问题的解决

常用的解决方法称为前向-后向概率计算。$\alpha_n(l)$表示前向概率，$\beta_n(l)$表示后向概率，它们都是通过递推得到的。在实际语音识别系统中通常递推长度N可达到40~100(笔者设定为40)，甚至更大，这使$\alpha_n(l)$最后变得非常小，以致超出了计算动态范围的下限，即使采用双精度计算，当n相当大时几乎所有$\alpha_n(l)$都趋于零。$\beta_n(l)$的计算也存在类似情况，这就是计算中的下溢问题。为解决此问题，只需要每推算一次便对运算结果乘以一个适当放大因子，使每一步递推维持在相似水平上，然后用此修正递推值作进一步运算，这里用$\hat{\alpha}_n(l)$和$\hat{\beta}_n(l)$表示修正后的值，递推计算过程为

$$\tilde{\alpha}_1(l)=\alpha_1(l)=\pi_l b_{iy_1}, \quad l=1,\cdots,L \tag{4}$$

$$\tilde{\alpha}_n(l)=\sum_{i=1}^{L}A_{il}b_{iy_n}\alpha_{n-1}(i), \quad n=2,\cdots,N, \quad l=1,\cdots,L \tag{5}$$

$$\hat{\alpha}_n(l)=\phi_n\tilde{\alpha}_n(l) \tag{6}$$

其中，

$$\phi_n=\left[\sum_{l=1}^{L}\tilde{\alpha}_n(l)\right]^{-1}, \quad n=1,\cdots,N \tag{7}$$

$$\hat{\beta}_N(l)=\phi_N, \quad l=1,\cdots,L \tag{8}$$

$$\hat{\beta}_n(l)=\phi_n\left[\sum_{h=1}^{L}A_{lh}b_{hy_{n+1}}\beta_{n+1}(h)\right], \quad l=1,\cdots,L, \quad n=(N-1),\cdots,1 \tag{9}$$

已知$\hat{\alpha}_N(l)=\left[\prod_{r=1}^{N}\phi_r\right]\alpha_N(l)$，而$\sum_{l=1}^{L}\hat{\alpha}_n(l)=1\ (n=1,\cdots,N)$，则

$$P_Y[\pi,A,B]=\sum_{l=1}^{L}\alpha_N(l)=\left[\prod_{r=1}^{N}\phi_r\right]^{-1}\sum_{l=1}^{L}\hat{\alpha}_N(l)=\left[\prod_{r=1}^{N}\phi_r\right]^{-1} \tag{10}$$

对式(10)两侧取自然对数，得到

$$\ln[P_Y(\pi,A,B)]=-\sum_{r=1}^{N}\ln\phi_r \tag{11}$$

在实际运算中，利用式(11)的对象运算可避免下溢问题。

(2)第2个问题的解决

一般采用Viterbi算法。实际应用中，通常采用对数形式的Viterbi算法，这样将避免进行大量的乘法运算，真正减少了计算量，同时还可保证很高的动态范围，而不会由于过多的连乘而导致溢出问题[3]。

(3)第3个问题的解决

常用的解决方法称为Baum-Welch算法。经典的Baum-Welch算法中，参数重估公式是在只有一个观察序列条件的假设下推导出来的。而实际应用中，有大量观察序列参与训练。设有一个HMM系统，它产生了K个相互独立的输出$Y^{(k)}\ (k=1,\cdots,K)$，设系统参数为λ，系统产生的每个输出的概率用$P_Y^{(k)}[\lambda]$来表示。修正参数计算方法为

$$\hat{A}_{ijmax}=\frac{\sum_{k=1}^{K}P_k^{-1}\sum_{n=1}^{N_k-1}\hat{\alpha}_n^{(k)}(i)A_{ij}p_{X_{n+1}=S_j}[y_{n+1}^{(k)}]\hat{\beta}_{n+1}^{(k)}(j)}{\sum_{k=1}^{K}P_k^{-1}\sum_{n=1}^{N_k-1}\sum_{j=1}^{L}\hat{\alpha}_n^{(k)}(i)A_{ij}p_{X_{n+1}=S_j}[y_{n+1}^{(k)}]\hat{\beta}_{n+1}^{(k)}(j)} \tag{12}$$

$$\hat{b}_{ljmax}=\sum_{k=1}^{K}P_k^{-1}\sum_{\substack{n=1\\(y_n=j)}}^{N}\gamma_n(l)\Big/\left[\sum_{k=1}^{K}P_k^{-1}\sum_{n=1}^{N}\gamma_n(l)\right] \tag{13}$$

其中，$\gamma_n(l)=\hat{\alpha}_n(l)\hat{\beta}_n(l)$。

HMM的初值选择也非常重要。π设定为均匀分布之值或非零随机值组成的矢量。第1个状态的初始概率为1，其余取0(由于采用从左到右无跳转HMM模型，因而无须重新设初始状态概率矢量π)；A_{ij}设定为状态数的倒数，在选用的模型中，对于每个词条v的任何状态S_i只涉及A_{ij}^{v}和$A_{i,i+1}^{v}$ 2个转移概率，且两者之和为1。当$A_{i,i+1}^{v}$初值为1/6时，可得到最好的结果[2]；B采取分段K-均值算法。整个系统的仿真界面如图2所示。

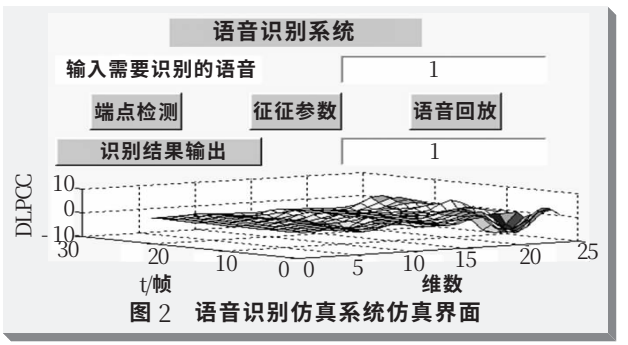

图2 语音识别仿真系统仿真界面

3 结束语

在语音识别系统中，语音的采样频率采用8 kHz，16 bit采样，帧长为160，帧移为80；训练人数为40，采样样本为男女各20个人的数码语音资料，男女分用2套不同参数；训练系统中当概率的相对变化值小到一定数值(设定为5×10^{-4})或迭代次数超过40时，结束迭代。实验表明，系统达到了较好的实时性和较高的识别率。由于MATLAB功能强大，在处理中可直接利用许多现成的函数，编程方便，结果可视化也容易实现。

参考文献

[1] 赵力. 语音信号处理[M]. 北京：机械工业出版社，2003.

[2] 杨行峻，迟惠生. 语音信号数字处理[M]. 北京：电子工业出版社，1995.

[3] 何强，何英. MATLAB扩展编程[M]. 北京：清华大学出版社，2002.

[4] 赵鹏，苏娟. 基于DSP的两级连接数码语音识别系统[J].湖南大学学报(增)，2005，5：51-53.

[5] JOSHI R L, POONACHA P G. A new MMSE encoding algorithm for vector quantization[C]// Proceedings of International Conference on Acoustics, Speech, and Signal Processing. [S.l.]: IEEE Press, 1991: 645-664.

作者简介

杨熙，硕士研究生，主要研究方向为语音增强、语音识别等；

苏娟，副教授，硕士生导师，主要研究方向为单片机、嵌入式系统应用、语音识别等。

[责任编辑] 潘浩然 [收稿日期] 2006-10-24