文章编号:1674-7070(2011)01-0062-05

基于平均路径长度的语音识别算法的研究与仿真

张艳萍[1]　张延盛[1]

摘要

提出了一种基于平均路径长度的语音识别算法. 采用的识别方法属于小词汇量孤立词语音识别,主要包括端点检测、特征提取和模式识别. 首先,在对语音信号预处理的基础上,采用梅尔频率倒谱系数(MFCC)为特征参数提取算法,动态时间规整(DTW)作为识别算法;然后,结合基于平均路径长度的模板训练方法,即采用少量样本,通过计算平均路径长度得到参考模板;最后,采用实验室环境下采集的语音信号进行实验. 仿真结果表明:改进后的算法与单个样本训练相比,提高了识别率及鲁棒性;同时,相对于矢量量化(VQ)技术,只需较少的训练样本,降低了算法的复杂度. 实验得到了较好的识别效果.

关键词

语音识别;动态时间规整;模板训练;平均路径长度

中图分类号 TN912.34

文献标志码 A

0　引言

语音识别是语言链中的一环,它研究的是使机器能准确地听出人的语音内容的问题,即准确地识别所说的话. 语音识别的最终目的是使计算机能够听懂任何人、任何内容的讲话. 关于语音识别技术,常用的有动态时间规整(Dynamic Time-Warping,DTW)算法、隐马尔科夫模型(Hidden Markov Models,HMM)和人工神经网络三种方法[1]. 在孤立词识别中,动态时间归整算法是把时间归整和间距测量计算结合起来的一种非线性归整技术,在相同条件下,它与另两种算法识别效果相差不大,但 DTW 运算量较少,所以在孤立词语音识别系统中,DTW 算法得到了更广泛的应用. 而在模板训练过程中,使用单个样本时系统鲁棒性不强,使用矢量量化(Vector Quantization,VQ)技术容易出现量化误差,并且需要大量的训练样本,计算量很大[2]. 针对上述方法存在的缺陷,本文采用基于平均路径长度的训练方法,可以有效提高系统鲁棒性,同时只需要使用少量的训练样本,计算量小.

1　语音识别系统的设计

1.1　语音识别的原理

语音识别系统本质上是一种模式识别系统,包括特征提取、模式匹配、参考模式库等3 个基本单元. 语音识别系统的基本结构如图1 所示.

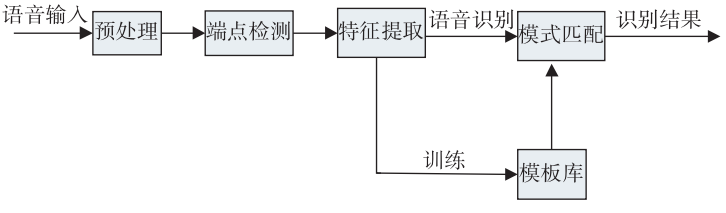

图1　语音识别系统

Fig. 1　Speech recognition system

未知(待识别)语音经过话筒变换成电信号(即语音信号)后加在识别系统的输入端,先经过预处理,再根据人的语音特点建立语音模型,分析输入的语音信号,并抽取所需特征,从而建立语音识别所需的模板. 在识别过程中,根据语音识别的模型,计算机将对已存放好的语音模板和输入的语音信号的特征进行比较,根据一定的搜索和匹配策

收稿日期 2010-07-10

资助项目 江苏自然科学基金(BK2009-410)

作者简介

张艳萍,女,教授,主要从事水声通信的研究. zypgjc@163.com.

1　南京信息工程大学　电子与信息工程学院,210044

略,找出一系列最优的与输入语音匹配的模板. 而这种最优的结果与特征的选择、语音模型的好坏、模板是否准确都有很大的关系[3].

1.2 预处理

预处理包括对语音信号的预滤波、预加重、采样与量化、分帧加窗、端点检测等. 其中预滤波是指滤除高于1/2采样频率的信号成分或噪声,使信号的带宽限制在某个范围内[4].

预加重的目的是为了提升语音信号中的高频成分. 语音信号的高频分量幅度较低,因此,高频部分的频谱比低频部分的难以辨认. 为了提高高频分量的作用,就需将其提升,使得整个信号的频谱比较平坦,同时也能抑制随机噪声. 方法是将语音信号通过一个高通滤波器. 这里采用一个一阶数字滤波器:

$$H(Z) = 1 - \mu z^{-1}.$$

分帧函数采用汉明窗[5]:

$$w(n) = \begin{cases} 0.54 - 0.46\cos\left[\dfrac{2\pi n}{N-1}\right], 0 \leqslant n \leqslant N-1, \\ 0, \quad \text{其他.} \end{cases} \tag{1}$$

窗形状如图2所示.

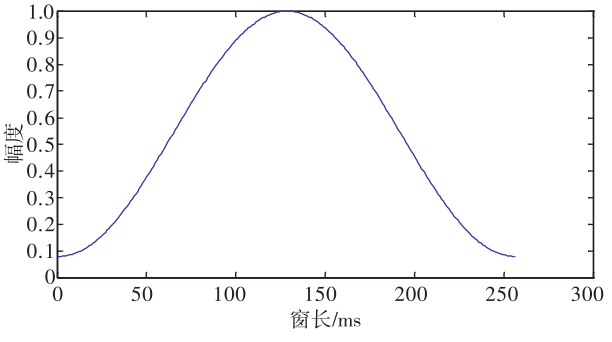

图2 汉明窗形状

Fig. 2 Hamming window

端点检测就是从说话人的语音命令中,正确判断每个语音的起点和终点,端点检测的准确性对识别率有很大的影响. 本文采用短时平均过零率和短时能量来进行端点检测. 短时过零率即单位时间内过零发生的次数:

$$Z_n = \sum_{-\infty}^{\infty} |\operatorname{sgn}[x(n)] - \operatorname{sgn}[x(n-1)]| w(n-k). \tag{2}$$

短时能量分布是指语音数据各帧之间的能量分布:

$$E_n = \sum_{-\infty}^{\infty} [x(k)w(n-k)]^2, \tag{3}$$

其中$x(k)$为语音序列[6].

1.3 特征向量提取

本系统采用梅尔频率倒谱系数(Mel-Frequency Cepstrum Coefficients,MFCC)作为特征值. MFCC考虑了人耳特征,具有很高的抗噪性和鲁棒性. MFCC的提取过程如下.

1) 对输入语音帧预加重和加汉明窗后作FFT(快速傅氏变换)得到其频谱,将时域信号转化为频域信号.

2) 求出频谱平方,即能量谱,并用M个Mel带通滤波器进行滤波,将每个滤波器频带内的能量进行叠加[7].

3) 将每个滤波器的输出取对数,得到相应频带的对数功率谱,并进行离散余弦变换(Discrete Cosine Transform,DCT),将滤波器输出变换到倒谱域,得到L个MFCC,它的变换式可表示为

$$C_n = \sum_{k=1}^{M} \log X(k)\cos[\pi(k-0.5)n/M], \quad n = 1,2,\cdots,L, \tag{4}$$

其中,$X(k)$为第k个滤波器的输出,L为MFCC系数的个数[6]. Mel滤波器组如图3所示,图中纵坐标单位为归一化单位.

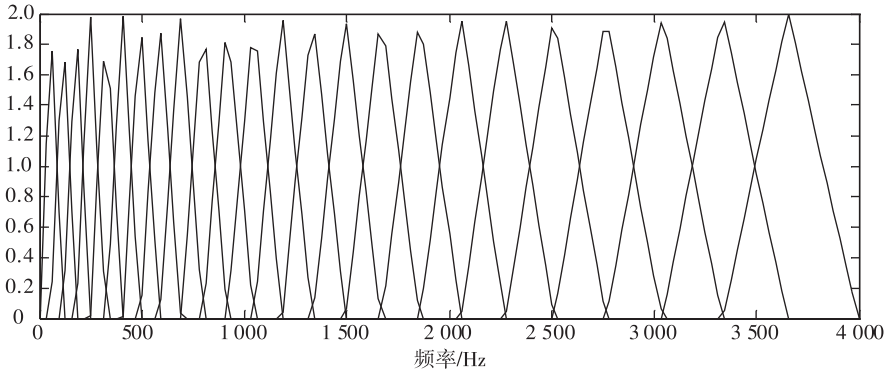

图3 Mel滤波器组

Fig. 3 Mel filter bank

4) 由于 MFCC 主要反映语音静态特征,为得到语音信号的动态特征,对静态特征进行一阶和二阶差分[8].

2 模板匹配与训练方法的改进

2.1 模板匹配

本文采用 DTW 算法:假设参考模板的特征矢量序列为 $X=\{x_1,x_2,\cdots,x_I\}$,输入语音特征矢量序列为 $Y=\{y_1,y_2,\cdots,y_J\}$,$I\neq J$. DTW 算法就是要寻找一个最佳的时间规整函数 $j=w(i)$,使待测语音的时间轴 j 非线性地映射到参考模板的时间轴 i 上,使总的累积失真最小,如图 4 所示.

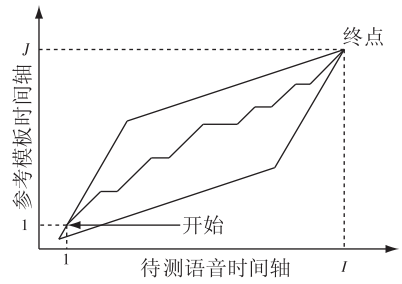

图 4 动态时间规整过程

Fig. 4 Dynamic programming

使时间规整函数满足:

$$D=\min_{\omega(i)}\sum_{i=1}^{I}d[T(i),R(\omega(i))]. \qquad (5)$$

式(5)中:$d[T(i),R(w(i))]$是第 i 帧测试矢量$T(i)$和第 j 帧模板矢量 $R(j)$ 之间的距离测度;D 则是处于最优时间规整情况下两矢量的距离[9].

如图 4 所示,规整函数 $\omega(i)$ 被限制在一个平行四边形内,它的一条边的斜率是 2,另一条边的斜率为 1/2. 规整函数的起点为(1,1),终点为(I,J). $\omega(i)$的斜率为 0、1 或 2. 这是一种简单的路径限制. 需要寻找一个规整函数,在平行四边形内由点(1,1)到点(I,J)具有最小代价函数. 由于已经对路径进行了限制,所以计算量可相应的减少[10]. 总代价函数的计算式为

$$D[c(k)]=d[c(k)]+\min D[c(k-1)]. \qquad (6)$$

式(6)中:$D[c(k)]$ 为匹配点 $c(k)$ 本身的代价;$\min D[c(k-1)]$ 是在 $c(k)$ 以前所有允许值(由路径限制而定)中最小的一个. 因此,总代价函数是该点本身的代价与到该点的最佳路径的代价之和[11].

动态规划算法从过程的最后阶段开始,即按逆序进行. 进行时间规整时,对每一个 i 值都要考虑沿纵轴方向能够达到的当前值的所有可能的点,根据路径限制可以减少这些可能的点,而得到几种可能的先前点,对每一个新的可能点按上述方法继续寻找最佳先前点,得到此点的代价. 随着过程的进行,路径要分叉,同时分叉的可能性也不断增大. 不断重复这一过程,得到从(I,J)点到(1,1)点的最佳路径[12].

2.2 训练

在整个算法过程中,模板建立的好坏将直接影响匹配结果. 在传统的训练方法中,给每个词选择一个样本作为这个词的参考模板. 由于只使用单个参考模板,这种训练方法的鲁棒性不强,因为即使是同一个人在不同时刻发出的同一个语音,也不可能完全一样[10]. 语音信号的产生取决于很多因素,因此,如果建立的模板不够理想,我们就需要不断地替换它,直到适合为止. 为了解决这个问题,同时又不增加计算量,本文采用一种改进的训练算法.

给每个词准备 N(N 一般不大于 4)个样本,计算这些样本的特征矢量序列沿 DTW 的平均路径长度. 找出路径长度与平均路径长度最接近的模板,把它定义为初始参考模板,对剩余的模板在 DTW 过程中进行匹配,使它们的路径长度与初始参考模板一致. 最后在每帧上对匹配好的的模板求平均,得到最终的参考模板. 最终参考模板的求取过程如下.

1) 将第 1 个模板和初始参考模板进行匹配,找出最佳的规整函数 $w(i)$,$1\leqslant i\leqslant I$.

2) 从最后一帧开始,进行后向搜索,直到第 1 帧,在 $w(i)$ 路径上寻找语音信号每一帧先前可能路径的斜率.

3) 有 3 种可能的斜率:

① 斜率为 1,保持不变;

② 斜率为 2,语音信号帧重复,即 $w(i-1)$与 $w(i)$表示相同帧;

③ 斜率为 0.5,对连续的两帧求平均,即求 $w(i)$与 $w(i-1)$的平均值.

4) 对剩余的模板重复步骤Ⅰ和Ⅱ,得到一组相等长度的模板.

5) 在每帧上对求得的模板取平均,得到最终参考模板.

因为使用了多个样本,这种训练方法建立的模板可有效地提高识别率和系统鲁棒性;同时,与矢量量化相比,此方法避免了量化误差,并且不需要大量样本,从而减少了计算量[13].

3 仿真实验及结果分析

本系统的语音采样频率为 8 kHz,采样精度为 16 bit. 语音信号通过一阶数字预加重滤波器来补偿语音中的高频部分,预加重系数为 0.95. 汉明窗窗长为 32 ms,帧移为 10 ms. 实验的指令词来自自制的语音库,包括汉语发音的 0 到 9,以及‘我是中国人’和‘南信大’,每个指令词有 10 个不同发音,总共 120 个样本.

1) 实验 1. 首先用第 1 组发音作为训练模板,其余样本为待测数据进行识别;接着以前 3 组发音作为训练模板,其余样本作为待测数据进行识别.

2) 实验 2. 对语音信号样本加噪,分别采用白噪声、粉红噪声、广播噪声. 任取一‘我是中国人’为例,分别加上各种噪声后的语音信号. 如图 5 所示.

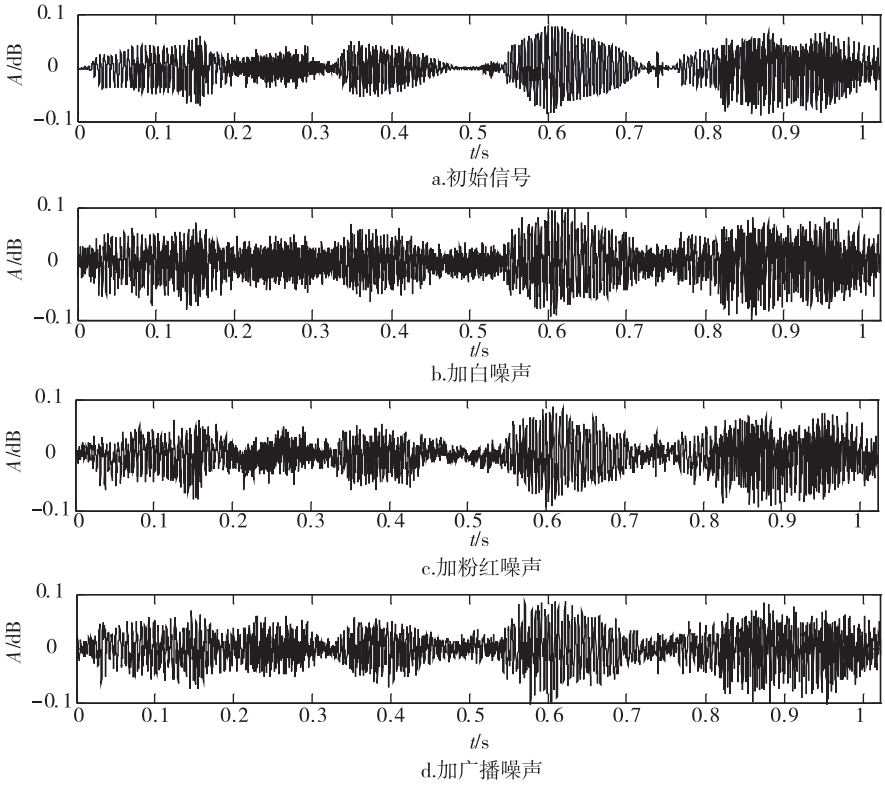

图 5 语音信号加各种噪声的前后对比

Fig. 5 Comparison of speech signals with different noise add

对加噪后的语音信号进行处理得到特征向量,之后按实验一的步骤进行识别. 实验结果如表 1 所示.

表 1 加噪信号识别率比较

Table 1 Recognition accuracy of speech signals with different noise add

信号方式	信噪比	传统训练方法/%	改进训练方法/%
原始信号		93	98
加白噪声信号	5.36	89	95
加白噪声信号	7.33	91	96
加粉红噪声信号	3.70	86	93
加粉红噪声信号	5.68	89	95
加广播噪声信号	4.55	86	92
加广播噪声信号	6.19	88	94

实验测试结果表明:本文提出的语音识别算法的特征数据传送稳定,在采用了改进训练方法后,无论对于实验室环境或是噪声环境,对孤立词的识别率都得到了提高.

4 结语

本文提出了一种基于平均路径长度的改进的 DTW 语音识别算法,应用改进的模板提取方法,在稳定完成语音识别过程的前提下,有效提高了识别率. 采用梅尔频率倒谱系数进行特征提取,使系统有着较高的鲁棒性. 然而,本语音识别系统还存在许多难点,包括同一发音信号的随机性变化,环境和噪声影响,端点检测的精确度等,都对最终的识别有较大影响. 因此,在以后的研究中,需要进一步对算法进

行优化,达到更好地抑制噪声的目的.

参考文献
References

[1] 韩纪庆,张磊,郑铁然. 语音信号处理[M]. 北京:清华大学出版社,2004:44-56
HAN Jiqing, ZHANG Lei, ZHENG Tieran. Speech signal processing[M]. Beijing: Tsinghua University Press, 2004:44-56

[2] 吴家安. 现代语音编码技术[M]. 北京:科学出版社,2008:26-30
WU Jiaan. Modern speech coding technology[M]. Beijing: Science Press, 2008:26-30

[3] 刘幺和,宋庭新. 语音识别与控制应用技术[M]. 北京:科学出版社,2008:28-31
LIU Yaohe, SONG Tingxin. Speech recognition and control technology[M]. Beijing: Science Press, 2008:28-31

[4] Rabiner L, Juang B H. Fundamentals of speech recognition[M]. New Jersey: PTR Prentice Hall, 1993:41-43

[5] Reynolds D A, Rose R C. Robust text-independent speaker identification using gaussian mixture speaker models [J]. IEEE Transactions on Speech and Audio Processing, 1995, 3(1): 72-83

[6] 王金明,张雄伟. 话者识别系统中语音特征参数的研究与仿真[J]. 系统仿真学报,2003,15(9):1276-1278
WANG Jinming, ZHANG Xiongwei. Study and simulation of the acoustic features in speaker recognition system [J]. Acta Simulata Systematica Sinica, 2003, 15(9): 1276-1278

[7] Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis[J]. IEEE Transactions on Neural Networks, 1999, 10(3):626-634

[8] Morris R W, Clements M A. Reconstruction of speech from whispers [J]. Medical Engineering & Physics, 2002, 24(7):515-520

[9] Zhang J, Zhang Y, Huang Z T. A recognition algorithm without ending-point detection of Chinese based on the DTW and HMM unified model[C]// IEEE International Conference on System, Man, and Cybernetics, 1998, 5: 4279-4283

[10] 张雄伟,陈亮,杨吉斌. 现代语音处理技术及应用[M]. 北京:机械工业出版社,2003:134-138
ZHANG Xiongwei, CHEN Liang, YANG Jibin. Modern speech processing technology and applications[M]. Beijing: China Machine Press, 2003:134-138

[11] Bojana G, Paliwal K K. Robust feature extraction using subband spectral centroid histograms[C]// Procceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001, 1: 85-88

[12] 许丽红. 孤立词特定人语音识别鲁棒性研究[D]. 上海:上海大学通信与信息工程学院,2002
XU Lihong. Study on robustness of isolated word and speaker-dependent speech recognition [D]. Shanghai: School of Communication and Information Engineering, Shanghai University, 2002

[13] Abdulla W H, Chow D, Sin G. Cross-words reference template for DTW-based speech recognition systems [J]. IEEE Conference on Convergent Technologies for Asia-Pacific Region TENCON, 2003, 4: 1576-1579

Research and simulation on speech recognition based on average length

ZHANG Yanping[1] ZHANG Yansheng[1]

1 School of Electronic & Information Engineering, Nanjing University of Information Science & Technology, Nanjing 210044

Abstract This paper designs a DTW-based speech recognition system. The method applied in this paper belongs to the small glossary's isolated words' speech recognition, which includes starting & ending point measuring, feature extraction and mode matching. The system takes pre-process of the speech signal, and then adopts the MFCC as a characteristic parameter drawing algorithm, and takes the DTW as the recognition algorithm, uses an improved template training technique to extracts the reference template with only a small quantity of samples by computing the average length. An experiment with speech signal recorded in lab-environment is given to simulate the proposed speech recognition method. The simulation results show that compared with using a single reference template, this method improves the recognition accuracy and the robustness. Furthermore, compared with the VQ, our method needs fewer samples and reduces the complexity of the algorithm.

Key words speech recognition; DTW; template training; average length