

矢量量化的音频信号类型确认算法的研究

陈宝远, 姜海鹏, 史玲玲, 王正, 王晓亮

(哈尔滨理工大学 测控技术与通信工程学院, 黑龙江 哈尔滨 150040)

摘要:语音识别和控制环境中,非语音信号(噪音)很容易与语音信号相混淆,一部分噪音被机器误认为语音,导致系统性能恶化.针对这一问题,以说话人识别理论为基础,提出了一种采用矢量量化来对音频信号类型(噪音或语音)进行确认的方法,以便于后期对非语音信号的噪音进行鉴别并且予以消除,提高语音信号处理系统的工作效率.在此基础上,在MATLAB平台上设计了一个基于此方法的语音与非语音识别的仿真系统.测试结果表明,系统实现了语音与非语音识别的基本功能,达到较好的工作效果.

关键词:说话人识别; MFCC; 矢量量化; lbg 算法

中图分类号: TP912.3 **文献标志码:** A **文章编号:** 1007-2683(2010)01-0108-04

Algorithm Research of Confirm the Type of Audio Signal Based on Vector Quantization

CHEN Bao-yuan, JIANG Hai-peng, SHI Ling-ling, WANG Zheng, WANG Xiao-liang

(School of Measure-control Technology and Communicat Engineering, Harbin University of Science and Technology, Harbin 150040, China)

Abstract: Among the environment of speech recognition and control, the non-speech signal (noise) was very easy to confuse with speech signal, apart of noise could be mistaken for speech result in the deteriorate of system. To solve this problem, this article based on the theory of speaker recognition, a method with vector quantization is presented to confirm the type of audio signal, so that to distinguish the non-speech signal and eliminate it later, advance the system's work efficiency. This article has designed a simulation system based on matlab. The results show that this system realized the basic function of distinguish the non-speech signal and speech signal, reached ideal effect.

Key words: speaker recognition; MFCC; vector quantization; lbg algorithm

0 引言

在音频矩阵的控制过程中,经常会出现各种类型的噪音如(咳嗽声,物体碰撞声,关门声),机器容易错误的将这些信号作为有效语音信号^[1],导致说话人的语音被很强的背景噪音或突发噪音打断,这

类噪声严重影响了音频矩阵控制的准确性.由于噪声来源众多,并且因应用场合而异,它们的特性也各不相同,因此,即便是在实验室仿真条件下,也难找到一种通用的、能适用于各种噪声环境的方法.

为解决这一问题,本文考虑到人耳的听觉特性,采用反应人类听觉特性的 Mel 倒谱系数 MFCC 作为音频信号特征,提出了一种用矢量量化(VQ)来进

收稿日期: 2009-02-25

作者简介: 陈宝远(1970—),男,副教授,研究生导师, E-mail: chenbaoyuan@126.com;

姜海鹏(1985—),男,硕士研究生.

行声音信号类型确认的方法^[1]. 该方法首先利用从不同场景中提取的音频信号特征,通过矢量量化的方法来训练语音与非语音的模板,然后在类型确认时,取与待测音频信号匹配距离最小的模板所对应的场景作为确认的结果. 本文采取方案如下:① 对若干段已知类型的声音文件(语音/非语音)进行训练分别生成语音、噪音两套码本,然后对要识别的语音文件的语音特征矢量用不同的码本进行矢量量化(VQ),最后比较这些量化的畸变值来判断语音的类型;② 将若干已知语音和噪音作为样本,抽取语音的特征参数,经过矢量量化,为每一段声音分别建立一个 VQ 码本,将未知声音按各个码本分别量化,认定量化误差最小的一个码本,即为最近邻 VQ 码本,则该未知声音与此最近邻码本同一类型声音信号.

2 语音/噪音识别系统的组成

语音/噪音识别与说话人识别系统相同,分为两个阶段,即训练(学习)和识别阶段^[1]. 在训练阶段,系统根据已知的若干训练文本建立相应的模型或参量参考集. 在本系统中通过 `function code = train(traindir, n)` 函数实现. 在识别阶段,由待识别音频信号中提取特征参量并与训练过程中建立的参考参量集或模板进行失真测度加以比较. 我们用 VQ 方法计算平均失真测度(本系统在计算距离 d 时,采用欧氏距离测度),选取距离最小的模板来判断待识别的声音所属类型. 在本系统中通过 `function test(testdir, n, code)` 函数来实现此过程^[1]. 图 1 是语音/噪音识别系统的框图.

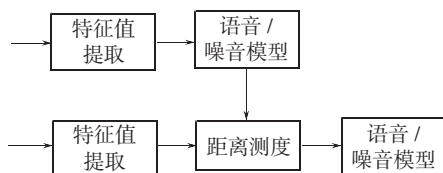


图 1 语音/噪音识别系统框图

3 声音特征 MFCC 的提取

在说话人识别系统中特征提取是最重要的一环,特征提取就是从说话人的音频信号中提取出表示说话人个性的基本特性^[1]. 代表性的特征参数有倒谱和基音参数. 试验证明,用倒谱特征可以得到比较好的识别性能,而且稳定的基音特征较难提取. 一

般来说,人能从声音的音色、频高、能量的大小等各种信息中知觉说话人的个性特征. 所以,如果利用复数特征的有效组合,可以得到比较稳定的识别性能. 实际上,人的听觉系统是一个特殊的非线性系统,它响应不同频率信号的灵敏度是不同的,基本上是一个对数的关系. 近年来,一种能够比较充分利用人耳这种特殊的感知特性的参数得到了广泛应用,这就是 Mel 尺度倒谱系数,称为 MFCC.

Mel 倒谱系数(MFCC)主要着眼于人耳的听觉特性,利用听觉前段滤波器组模型,提取说话人的个性特征参数. 该特征谱由 Davies 和 Mermelstein 提出,已经被证明为在语音相关的识别任务中应用最成功的特征描述之一. Mel 倒谱系数(MFCC)是在 Mel 标度频率域提取出来的倒谱参数^[1]. Mel 标度描述了人耳对频率感知的非线性特性,它与频率的关系可用下式近似表示为

$$\text{Mel}(f) = 2595 \log(1 + f/700) \quad (1)$$

式中 f 为频率, Hz.

MFCC 参数的提取过程如下:

1) 首先对输入的音频信号进行分帧和加窗处理,然后作离散傅里叶变换,获得频谱分布信息. 设音频信号的 DFT 为

$$X_a(k) = \sum_{n=1}^{N-1} x(n) e^{-j2\pi nk/N}, 0 \leq k \leq n-1 \quad (2)$$

式中: $x(n)$ 为输入的音频信号; N 表示傅里叶变换的点数.

2) 再求频谱幅度的平方,得到能量谱.

3) 将能量谱通过一组 Mel 尺度的三角形滤波器组.

我们定义一个有 M 个滤波器的滤波器组(滤波器的个数和临界带的个数相近),采用的滤波器为三角滤波器,中心频率为

$$f(m), m = 1, 2, \dots, M, \text{本系统取 } M = 24.$$

4) 计算每个滤波器组输出的对数能量.

$$S(m) = \ln \left(\sum_{k=1}^{N-1} |X_a(k)|^2 H_m(k) \right), 0 \leq m \leq M-1 \quad (3)$$

其中 $H_m(k)$ 为三角滤波器的频率响应.

5) 经过离散预选变换(DCT)得到 MFCC 系数.

$$C(n) = \sum_{m=0}^{M-1} S(m) \cos(\pi n(m - 0.5/M)) \quad 0 \leq n \leq N-1 \quad (4)$$

MFCC 系数个数通常不用 0 阶倒谱系数,因为它反应的是频谱能量,故在一般识别系统中,将 C_0

称为能量系数,并不作为倒谱系数,本系统选取 16 阶倒谱系数.在语音信号处理工具箱 Voice Box 中,调用了函数 Melbankm. m,用来计算 M 个滤波器的系数. Melbankm. m 的语法为:

$$h = \text{melbankm}(m, n, f_s) \quad (5)$$

其中, m 为滤波器的个数, n 为语音帧的长度, f_s 为采样频率. 本文设计了一个滤波器组,包含 20 个滤波器,音频信号的帧长度为 256,信号的采样频率为 8 000 Hz. 图 2 为响应的 mel 尺度滤波器组.

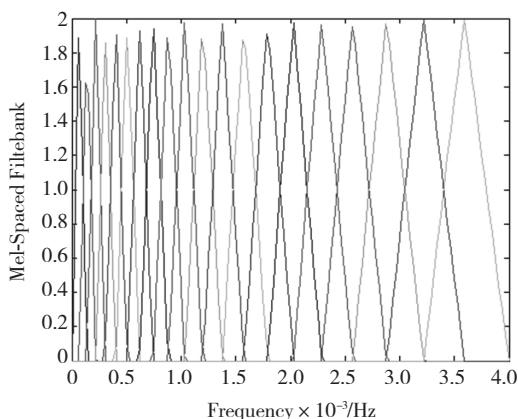


图2 mel 尺度滤波器组

4 基于 VQ 的语音/噪音识别算法

目前说话人识别的主要方法有基于隐马尔科夫模型(HMM)、基于矢量量化(VQ)以及基于高斯混合模型法(GMM). Matsui 和 Furui 从语音波动的识别性能对基于 VQ 方法和各态历经的离散的和连续的 HMM 方法进行了比较,结果表明,当用于训练的数据量较小时,基于 VQ 方法比连续的 HMM 方法识别性能更好.同时,基于 VQ 方法比较简单,实时性也较好,综合各方面因素,本文采用基于矢量量化的说话人识别方法^[7].

在说话人识别系统中,VQ 起着双重作用.在训练阶段,提取每一断声音的特征参数进行聚类,本文采用 lbg 聚类算法为每个说话者产生一个码本.在识别阶段,我们用 VQ 方法计算平均失真测度,从而判断说话人是谁.本文中方案 1 同样为每一段声音建立一个码本,方案 2 则提出了一种为若干段同类声音建立一个公共码本的方法.

4.1 用矢量量化聚类法生成码本

我们将待识别的声音看作是一个信源,用一个码本来表征.码本是从该说话人的训练序列中提取的 MFCC 特征矢量聚类而生成.只要训练的序列足

够长,可认为这个码本有效地包含了声音的个性特征^[8].

1) 提取出来的所有帧的特征矢量的型心(均值)作为第一个码字矢量 B_1 .

2) 取当前的码本 B_m 根据以下规则分裂,形成 $2m$ 个码字

$$\begin{cases} B_m^+ = B_m(1 + \varepsilon) \\ B_m^- = B_m(1 - \varepsilon) \end{cases} \quad (6)$$

其中 m 从 1 变化到当前的码本的码字数, ε 是分裂时的参数,本文 $\varepsilon = 0.01$.

3) 根据得到的码本把所有的训练序列(特征矢量)进行分类,然后按照下面两个公式计算训练矢量量化失真量的总和 $D^{[k]}$ 以及相对失真(n 为迭代次数,初始 $n = 0, D^{[-1]} = \infty, B$ 为当前的码书),若相对失真小于某一阈值 ε ,迭代结束,当前的码书就是设计好的 $2m$ 个码字的码书,转 5). 否则,转下一步.

量化失真量和:

$$D^{(n)} = \sum_{k=1}^K \text{mind}(X_k, B) \quad (7)$$

相对失真:

$$\left| \frac{D^{(n-1)} - D^n}{D^n} \right| \quad (8)$$

4) 重新计算各个区域的新型心,得到新的码书,转步骤 3.

5) 重复 2), 3), 4) 步骤,直到形成有 M 个码字的码书(M 是所要求的码字数).

4.2 测度判决

识别的时候,用同样的方法从待识别的音频中提取音频信号的特征参数矢量序列,用系统中建立的码本对其进行矢量量化,用下式定义的平均量化畸变程度判断该矢量序与哪一个码本的分布最为接近.

$$D_i = \frac{1}{M} \sum_{n=1}^M \min_{1 \leq l \leq L} [d(X_n, Y_l^i)] \quad (9)$$

式中: Y_l^i 表示第 i 个码本的第 l 个码子; M 是特征矢量序列的长度,也就是待识别语音所包含帧的总数, $d(X_n, Y_l^i)$ 采用欧式距离测度.最后选择平均量化误差最小的码本对所对应的说话人作为系统的识别结果.

5 实验分析及结果

本系统录音使用普通的话筒,通过计算机的声卡来实现.对音频信号的采样频率为 8 000 Hz,量化

位数为 16bit. 在提取特征参数时,每帧的采样点数为 256,帧移为 80 点. 本系统使用 16 维 MFCC 参数进行训练和识别^[9].

在训练过程中,分别选用 7 人的说话录音和 7 种不同的噪音作为样本,其中内容各不相同. 为了确保实验结果的有效性,测试时选用的声音样本与训练样本不同.

按照方案(1)进行的实验中,码本 1~7 是语音的模板,码本 8~14 是噪音的模板. 如表 1 所示,输入待识别的信号为噪音时,经过 VQ 算法计算得到的量化畸变,选取最小的,即最接近模板 13,是属于噪音信号.

表 1 待识别信号为噪声的识别结果

模板 1	量化畸变	模板 2	量化畸变
1	5.637 1	8	7.088 6
2	6.959 8	9	8.942 6
3	11.179 8	10	7.053 7
4	13.679 7	11	6.211 2
5	7.149 7	12	6.555 8
6	9.098 5	13	5.118 4
7	5.883 6	14	5.809 2

在实验中,运用矢量量化算法(VQ)对采集到的,不同人的 23 个说话声和 39 个不同的噪声进行训练和识别,其结果如表 2 所示.

表 2 方案(1)识别结果

模板类型	噪声	语音
识别率/%	81.25	87.5

在按方案(2)进行的实验中,提取所有语音公共参数建立的码本 1 为语音的模板,提取所有噪音公共参数建立的码本 2 为噪音的模板. 类似的,如表 3 所示,输入待识别的信号为噪音时,即最接近模板 2,是属于噪音信号. 当输入的待识别信号为语音信号时,如表 4 可以看到,运用 VQ 算法进行比较,得到结果是跟模板 1 比较接近,即属于语音类的信号.

表 3 待识别信号为噪声的识别结果

模板	1	2
量化畸变	6.755 0	5.108 5

表 4 待识别信号为语音的识别结果

模板	1	2
量化畸变	8.691 4	9.653 7

表 5 方案(2)识别结果

模板类型	噪声	语音
识别率/%	87.5	89.8

由上述实验结果可知,采用方案(2)的计算量小,并且准确率高方案(1). 用不同的码字系统去量化未知声音,源自同种类型的声音,量化误差必然最小,这就是本系统的基本思想.

6 结 论

通过大量的实验数据表明,利用本系统建立的声音数字参数的 VQ 码本,解决了未知声音信号的类型确认问题,在准确性和实时性上得以保证,解决了传统的端点检测方法只能识别短促噪音无法识别持续噪音的问题^[10],并且扩展了说话人识别的应用范围. 但是实验用码本,码字规模较小,远没有把不同类型的声音的全部特征包括进来参与模板建立,对系统的准确性造成影响,故系统应该选取尽可能多的具有代表性的声音建立模板,提高 VQ 码本的质量. 基于本文介绍方法加以扩展,在取得大量素材的前提下,可以用于更为复杂的语种类型确认或动物类型确认的研究中.

参 考 文 献:

[1] DOUGLAS A Reynolds. An Overview of Automatic Speaker Recognition Technology [A]. Proc ICASSP [C]. Orlando, Florida, USA:IEEE,2002:4072-4075.

[2] 孙圣和,陆哲明. 矢量量化技术及应用 [M]. 北京:科学出版社,2002.

[3] 王仁华,自动说话人识别 [J]. 信号处理,1991,7(4):193-200.

[4] 何 强,何 英. MATLAB 扩展编程 [M]. 北京:清华大学出版社,2002.

[5] 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. 20(4):2-10.

[6] 郭春霞,裴雪红. 基于 MFCC 的说话人识别系统 [J]. 电子技术,2005,22(3):53-56.

[7] 胡 征. 矢量量化原理及应用 [M]. 西安:西安电子科技大学出版社. 1998.

[8] 姜洪臣,梁 伟,张树武,等. 音频场景分类的音频特征提取和分析 [J]. 声学技术,2005,24(6):91-104.

[9] 赵 立. 语音信号处理 [M]. 北京:机械工业出版社,2003.

[10] 陈宝远,梁伟明. 基于小波分析的语音端点检测算法研究与仿真 [J]. 哈尔滨理工大学学报,2009. 14(1):51-54.

(编辑:王 萍)