采用 DTW 算法和语音增强的嵌入式声纹识别系统

周跃海[1],童 峰[1*],洪青阳[2]
(1. 厦门大学 海洋与环境学院,水声通信与海洋信息技术教育部重点实验室,
2. 厦门大学信息科学与技术学院,福建 厦门 361005)

摘要:动态时间规整(dynamic time warping,DTW)是一种相对简单成熟的算法,广泛用于语音识别系统中.针对环境噪声对声纹识别系统性能的影响,用信噪比关联谱减及自适应门限端点检测进行抗噪声处理,在此基础上采用 DTW 算法设计了基于嵌入式 ARM9 平台的声纹识别实现方案,并给出了带噪环境下的声纹识别实验结果.

关键词:声纹识别;动态时间规整;谱减;嵌入式系统
中图分类号:TN 912 **文献标志码**:A **文章编号**:0438-0479(2012)02-0174-05

声纹识别(voice print recognition)也称说话人识别,就是根据说话人的声音特征,识别出某段语音是谁说的.声纹是人的个性特征,很难找到两个声纹完全一样的人,因此,声纹识别广泛应用于安防、公安、军队、银行、证券、个人身份认定等领域.

在孤立词识别中,最有效最简单的方法是采用动态时间规整(dynamic time warping,DTW)算法,该方法最显著的优点是复杂度低、识别率比较高,因而在语音识别、说话人识别等领域被广泛研究.万春[1]通过对语音识别数学模型 DTW 算法的研究和改进,实现了语音识别系统;Lippmann[2]运用 DTW 算法在数字信号微处理器(DSP)上实现了一个功耗低、精度高、快速识别的声纹识别系统;Christophe Levy 等[3]则运用 DTW 算法和隐马尔科夫模型(HMM)在蜂窝手机上实现了声纹识别系统.

在声纹识别系统的实际应用环境中,环境噪声所引起的畸变严重影响着声纹识别的性能.因此有必要对语音进行消噪来提高信噪比,提高声纹识别的性能.谱减方法由于实现简单方便、约束条件少、物理意义直接,在语音信号的抗噪声处理中得到了广泛的研究.

针对嵌入式声纹识别系统在门禁等安防领域的应用并考虑到实际应用中背景噪声的影响,本文结合改进谱减算法和自适应门限端点检测算法进行语音信号增强,实现了采用 DTW 算法的嵌入式声纹识别系统,并对系统进行实验测试分析.

1 算法概述

1.1 特征向量提取

梅尔频标倒谱系数(mel frequency cepstrum coefficient,MFCC)是建立在傅里叶和倒谱分析基础上的.对短时音频帧上的采样点进行傅里叶变换,得到这个短时音频帧在每个频率上的能量.将整个频率分成 n 个就构成了 MFCC(也叫 Mel 系数).如果对提取出来的 Mel 系数再计算其对应的倒谱系数,就是 Mel 倒谱系数.它广泛地应用于各种说话人识别和处理领域中.通过式(1)就可以提取出特征向量.

$$\boldsymbol{C}_n=\sum_{k=1}^{M}\ln x'(k)\cos\left[\pi(k-0.5)n/M\right],n=1,2,3,\cdots,L. \tag{1}$$

1.2 DTW 算法

DTW 算法是把时间规整和间距测量计算结合起来的一种非线性规整技术,它将语音特征向量的时间序列与参考模板库中的模板进行相似度比较,将相似度最高的作为识别结果,从而有效地解决了小词汇量识别中说话速度不均匀的问题.通过将待识别的语音信号的时间轴进行不均匀的扭曲和弯曲,使其特征和模板特征对齐,并在两者之间不断地进行两个矢量最小的匹配路径计算,DTW 算法可以获得两个矢量匹

收稿日期:2011-06-14
基金项目:福建省高校产学合作重大项目(2010H6022);福建省杰出青年科学基金项目(2010D003);厦门市科技计划项目(3502Z20113008)
*通信作者:ftong@xmu.edu.cn

配距离最小的规整函数. 这是一个将时间规整和距离测度有机结合在一起的非线性规整技术,保证了待识别特征和模板特征之间最大的声学相似特性和最小的时差失真[4].

1.3　信噪比关联谱减算法

谱减算法是噪声处理较为有效的方法. 其基本思想是从带噪语音的功率谱中减去噪声的功率谱,从而得到较为纯净的语音信号.

设 $s(t)$ 为纯净的语音信号,$n(t)$ 为噪声信号,$y(t)$ 为带噪语音信号,则有

$$y(t)=s(t)+n(t), \tag{2}$$

对式(3)作傅里叶变换,$Y(\omega)$,$S(\omega)$,$N(\omega)$ 分别为 $y(t)$,$s(t)$,$n(t)$ 的傅里叶变换,则

$$Y(\omega)=S(\omega)+N(\omega), \tag{3}$$

假设 $s(t)$,$n(t)$ 相互独立,则

$$E|Y(\omega)|^2=E|S(\omega)|^2+E|N(\omega)|^2, \tag{4}$$

假设语音信号短时平稳,对语音信号进行加窗处理,因此可得

$$|Y_i(\omega)|^2=|S_i(\omega)|^2+|N_i(\omega)|^2, \tag{5}$$

式中 i 表示加窗分帧后的第 i 帧.

根据式(6)就可以得到纯净语音信号的频谱

$$|S_i(\omega)|=[|Y_i(\omega)|^2-|N(\omega)|^2]^{1/2}, \tag{6}$$

对 $S_i(\omega)$ 进行傅里叶逆变换,就可以得到纯净的语音信号.

传统的谱减算法直接用带噪信号的频谱减去噪声的频谱,在平稳噪声的语音上可以取得不错的效果. 但是,由于语音的能量往往集中在某些频段内,尤其是共振峰对应频带处的幅度一般远大于噪声,而语音中的噪声往往是随机不平稳的,用加了不平稳噪声的语音信号的频谱减去一个固定的噪声频谱往往无法达到较好的语音增强效果,因此不用相同标准对语音信号进行谱减处理. 由于传统谱减方法的这一局限性,众多研究者对其进行改进,如:丁伟等[5]采用对幅度高的信号帧施加一个系数固定的加权谱减来改善对非平稳噪声的降噪效果;李晔等[6]、沈晓东等[7]则根据对多帧非平稳背景噪声的递推估计结果进行谱减.

在本文安防领域的嵌入式声纹识别系统中,由于采集的语音段短(本文算法中为 3 s),可近似认为在此段时间内背景噪声为平稳特性,此时我们根据语音信号的非平稳特性对谱减算法进行信噪比关联加权调整,即:当语音信号强、信噪比高时,由于遮蔽效应,背景噪声对人耳听觉影响小,此时采用较小的加权系数进行谱减;反之,语音信号弱、信噪比低时,由于背景噪声的相对影响大,采用较大的加权系数进行谱减. 同时,为了进一步改善不同信噪比下的不同加权谱减的效果,本文算法设定了加权系数与高、中、低信噪比的对应关系.

考虑嵌入式声纹门禁等安防领域的应用,使用者以刷卡或按键触发声纹识别程序后,从语音采集开始到使用者开始说话之间会有一段无语音的时间延迟,可将在该段"静音段"时间内录到的信号为背景噪声来近似代表较短的采集时间内的背景噪声特征,因此在算法具体实现上:

1) 取第 1 帧背景噪声为谱减算法中的基准噪声;

2) 获取基准噪声后通过信噪比来描述语音信号幅度的非平稳,将第 i 帧语音信号的能量除以第 1 帧噪声的能量得到的比值作为第 i 帧语音的信噪比(SNR_i);

3) 根据不同信噪比下背景噪声的不同影响,引进了一个与高、中、低信噪比关联的加权参数 b,对式(7)进行修正

$S_i(\omega)=(Y_i(\omega)-b\times N_i(\omega))^{1/2}$,如果 $S_i(\omega)<0$,则 $S_i(\omega)=0$.

其中,$b=\begin{cases}15, & a_1<\mathrm{SNR}_i\leqslant a_2,\\ 10, & a_2<\mathrm{SNR}_i\leqslant a_3,\\ 6, & \mathrm{SNR}_i>a_3.\end{cases}$

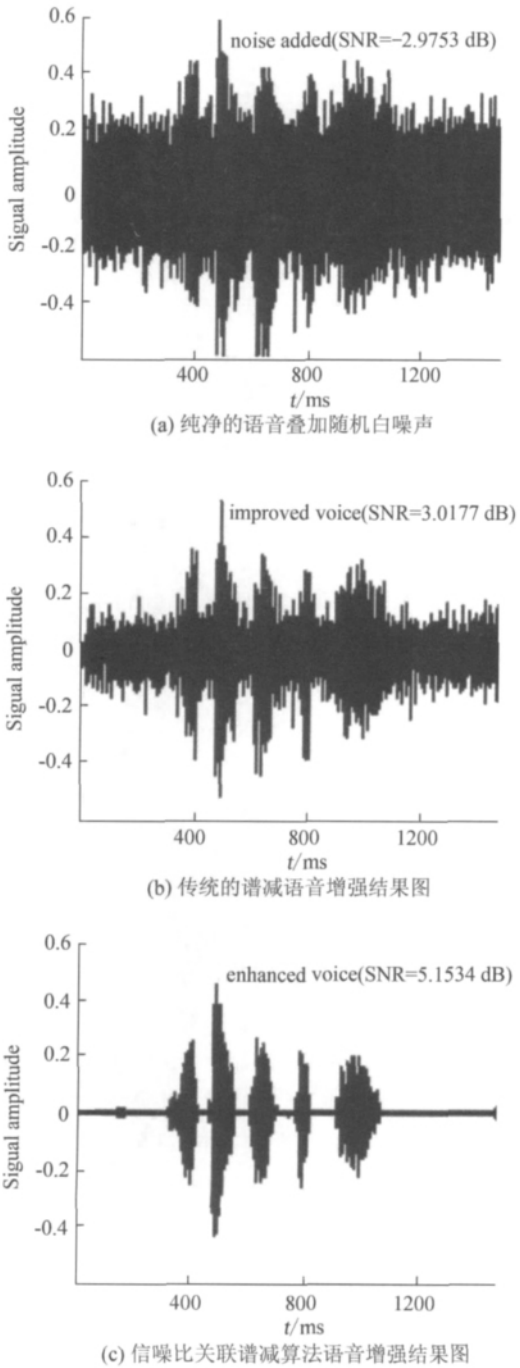
a_1,a_2,a_3 为信噪比的门限值,单位为 dB.

图 1 给出对一段背景噪声为白噪声的语音信号进行传统谱减和信噪比关联谱减结果的对比图. 在本文中,取 $a_1=0$ dB,$a_2=2$ dB,$a_3=4$ dB. 从图 1 中可以看出谱减算法能够根据语音信号的信噪比,比较好地对语音进行增强,达到消除噪声的效果. 从图 1(b)、(c)可以看出信噪比关联谱减算法的语音增强效果明显优于传统的谱减算法.

1.4　自适应门限端点检测

在声纹识别中,一个关键问题就是如何利用端点检测将语音信号精确地检测出来,为获得准确的识别提供前提. 针对背景噪声对传统端点检测算法的影响,在传统的短时平均能量和短时平均过零率的基础上,考虑到本文应用采集的语音段短,可近似认为此段时间内背景噪声为平稳特性,提出一种自适应门限的端点检测算法,利用语音信号前后的"静音段"获取的背景噪声特性作为门限改善在噪声背景下的语音端点检测效果.

由于语音信号是一种非平稳信号,用移动的有限长的窗口进行分帧,并认为在一帧内它是平稳的. 一帧内的信号能量值和过零次数分别被称为短时平均能量和短时平均过零率. 窗长的选择一般包含1～7个基音



(a) 纯净的语音叠加随机白噪声

(b) 传统的谱减语音增强结果图

(c) 信噪比关联谱减算法语音增强结果图

图 1 谱减对比图

Fig. 1 Comparison of spectral subtraction results

周期,一般选取 10～30 ms 的时间作为窗长[8]. 设语音信号为 $x(n)$,短时平均能量定义为

$$Mn = \{\sum_{m=n-N+1}^{n}[x(m)\times x(m)\times w(n-m)]\}/N, \tag{8}$$

短时平均过零率定义为

$$Zn = \sum_{m=n-N+1}^{n} |\operatorname{sgn}[(x(m)-T_0]-\operatorname{sgn}[x(m-1-T_0)|\, w(n-m), \tag{9}$$

其中 sgn ()为符号函数,窗函数为

$$w(n)=\begin{cases}1, & (0\leqslant n\leqslant N-1),\\ 0, & (n>N-n, n<0),\end{cases}$$

其中 N 为矩形窗的长.

在传统的语音端点检测中,短时平均能量和短时平均过零率是一个固定的门限值,由于环境的背景噪声不同和外界的干扰差异,使固定的门限在端点检测过程中不能真正地检测到语音的起始点. 自适应门限就是为了克服这个问题而提出来的. 一般而言,有效语音前后都包含一些体现背景噪声和干扰的语音段,把第一帧和最后一帧的短时平均能量和短时平均过零率加上某个经验值来作为检测起点和终点短时平均能量和短时平均过零率的门限. 实验表明该算法能够较好地检测带噪语音的端点.

图 2 给出了一段带噪语音信号端点检测前和端点检测后的对比图,图 2(a)中的语音信号中叠加有随机白噪声. 从图 2(b)中可以得出该算法在有噪声的情况下能够较好地检测出语音的端点.

2 嵌入式声纹识别系统实现方案

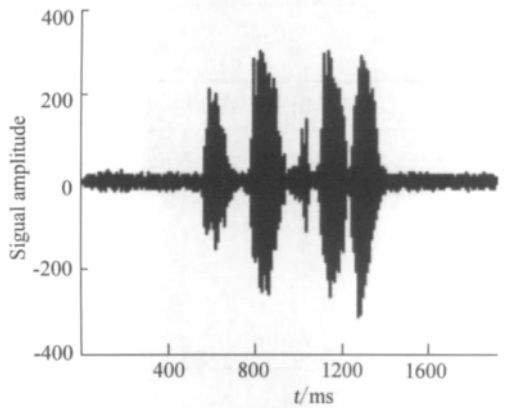
2.1 总体设计

本文系统的流程图如图 3 所示,其核心为 ARM9 S3C2440 微处理器. MIC 前置放大模块用于对语音进行放大,LCD 触摸显示屏用于显示结果和用户接口,外部的 Nand Flash 用于保存模板和用户数据,Nor Flash 用于保存程序,外部随机存储器(RAM)用于存储计算中间变量.

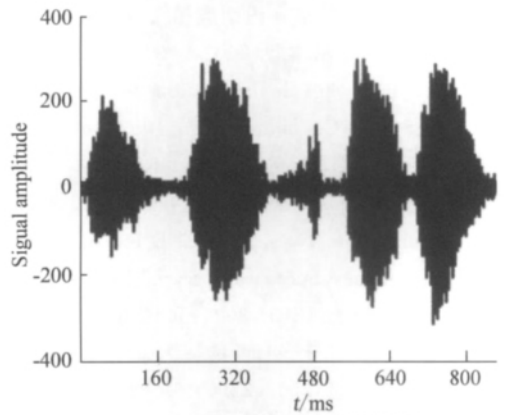
2.2 硬、软件设计

本系统采用 Mini2440 开发板,系统硬件相关配置为:CPU 主频 400 MHz,定时器 0 时钟为 0.5 MHz,音频接口 AD 采样频率为 8 kHz,音频的采样精度为 10 bit,录音时间为 3 s.

图 4 为系统的软件流程图,通过触摸屏控制系统的流程. 如果是注册,则把注册成功的模板保存在 Nand Flash 中;如果是识别则调出模板,用 DTW 匹配方式进行用户识别. 该声纹识别系统是在 Mini2440 平台上实现的,其开发环境是 ADS 1.2,调试环境是 AXD,开发语言为 C 语言. 该系统设置成 5 个用户,将这些训练成功的模板保存在 Nand Flash 中第 100～104 块,将用户信息保存在第 105 块,通过操作 Nand Flash 的这些块,可以进行用户注册、识别、删除等,并且将数据保存在NandFlash中实现了断电保存数据



(a) 端点检测前的语音信号



(b) 端点检测后的语音信号

图 2　端点检测试验结果

Fig. 2　Endpoint detection results

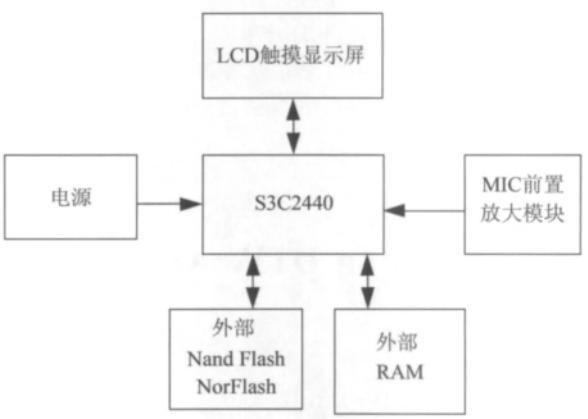


图 3　系统的总体结构

Fig. 3　System structure

的功能.

由于嵌入式处理器硬件资源有限，在程序设计中将一些常量和显示在屏幕上的图片数据定义为常量保存在 FLASH 中，节省 RAM 的开销. 另外，S3C2440 对浮点比对定点计算消耗更多的时钟周期，因此在程序中尽量避免对浮点的计算，但是为了保证数据的精度，在谱减算法中保留了对浮点的运算.

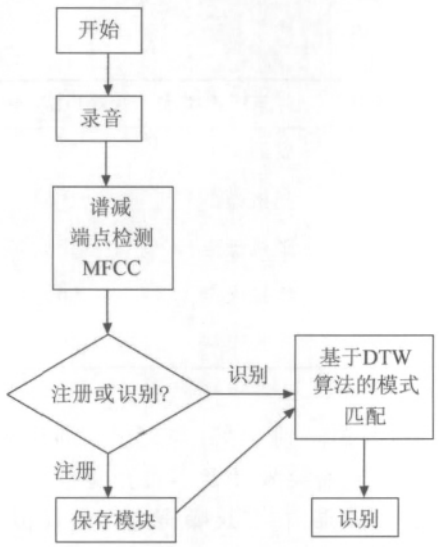


图 4　系统流程图

Fig. 4　System flowchart

3　实验结果和结论

3.1　实验设计

为了测试本文系统的性能，在办公室进行嵌入式声纹识别实验. 实验所在办公室内比较安静(背景噪声级为 58 dB(A)). 实验中分别采用随机噪声和实际环境噪声对系统的抗噪声性能进行测试. 其中：随机噪声是用Matlab软件产生一段高斯随机信号；实际环境噪声则是在室外实录的背景环境噪声，此段噪声包含车辆噪声、行人噪声等. 实验中分别将这两段噪声通过音箱播放出来模拟声纹识别实验的背景噪声，并用声级计分别测量噪声的强度.

实验中录音的时间为 3 s，说话人以正常的语速说话，说话长度为 4～5 个字. 在识别前，先登记 5 个模板，测试者分别说登记时的短句来测试系统的性能.

实验分别在安静环境下、随机噪声环境下和实录环境噪声环境下进行测试，测试结果如表 1 所示：

3.2　结果与讨论

表 1 中“正确识别”表示能够正确地识别出自己登记的模板，“错误识别”表示系统识别到了其他用户的模板，“被拒绝识别”表示没有识别到任何模板. 在较安静环境下(58 dB(A))，该系统可以取得较好的识别效果，识别率达 92%；在带噪声环境下，识别效果有所下降. 在 110 dB(A)的随机噪声条件下，使用改进型谱减算法后可以有效地抑制白噪声，提高了系统的识别率，

表 1 嵌入式声纹识别系统实验结果

Tab. 1 Experimental results

实验条件	背景噪声类型	背景声级/dB(A)	次数(10 人)	正确识别	错误识别	被拒绝识别	识别率/%
无语音增强	安静	58	100	92	2	6	92
无语音增强	随机噪声	100	100	71	5	24	71
有语音增强	随机噪声	100	100	84	3	13	84
无语音增强	实录噪声	90	100	60	14	26	60
有语音增强	实录噪声	90	100	74	6	20	74

识别率从 71%提高到 84%. 与理想高斯白噪声不同,实际环境噪声具有突发性及不可预见性,这种噪声严重影响着识别效果. 在实录噪声进行的模拟实际背景噪声测试中,本文采用的语音增强技术使系统的识别率从 60%提高到 74%. 因此,从表 1 实验结果可以看出,通过采用信噪比关联谱减及自适应门限端点检测进行的语音增强处理,本文设计的嵌入式声纹检测系统可以有效改善噪声背景下 DTW 声纹识别的性能.

4 结 论

本文设计了采用 DTW 算法和语音增强处理的嵌入式声纹识别系统,噪声背景下的实验结果表明,该系统有较高的识别率,说明基于信噪比关联谱减及自适应门限端点检测的语音增强处理可以有效提高嵌入式 DTW 声纹识别的性能. 当然,由于 MFCC 程序复杂和谱减算法涉及到对浮点的计算,嵌入式系统中大量资源都花费在这些浮点计算上,因此有必要对算法及程序设计进行进一步精简优化.

参考文献:

[1] 万春. 基于 DTW 的语音识别应用系统研究与实现[J]. 集美大学学报:自然科学版,2002,7(2):104-108.

[2] Lippmann R P. Speech recognition by machines and humans[J]. Speech Communication, 1997,22(1):1-15.

[3] Levy C, Linares G, Nocera P, et al. Recucing somputational and memory cost for cellular phone embedded speech recognition system[C]//2004 IEEE International Conference on Acoustics, Speech and Signal Processing. Montreal, Quebec, Canada: IEEE, 2004:309-312.

[4] 王社国,赵建光. 基于 ARM 的嵌入式语音识别系统研究[J]. 微计算机信息,2007,23(5):149-150.

[5] 丁伟,吴小培. 基于改进谱减方法的语音增强研究[J]. 计算机技术与发展,2008,18(9):98-100.

[6] 李晔,崔慧娟,唐昆. 基于谱减的语音增强算法的改进[J]. 清华大学学报:自然科学版,2006,46(10):1685-1686.

[7] 沈晓东,岳晓果. 一种基于谱减的语音增强算法的改进[J]. 软件导刊,2009,8(11):72-73.

[8] 陈迪,龚卫国,杨利平. 基于基音周期的语音 MFCC 参数提取[J]. 计算机应用,2007,27(5):1217-1219.

Embedded Voiceprint Recognition System Based on DTW Algorithm and Speech Enhancement

ZHOU Yue-hai, TONG Feng*, HONG Qing-yang

(1. Key Laboratory of Underwater Acoustic Communication and Marine Information Technology, Ministry of Education, College of Oceanography and Environmental of Science, Xiamen University, 2. School of Information Science and Technology, Xiamen University, Xiamen 361005, China)

Abstract: Dynamic time warping (DTW) is a relatively simple and effective algorithm, which is widely used in speech recognition systems. In order to deal with the performance degration caused by background noise, this paper utilized signal to noise ratio (SNR) spectral substration algorithm and adaptive threshold endpoint detection algorithm. Basing on the two algorithms, the speech recognition system is successfully implemented on ARM9 embedded microprocessor. Finally the experimental results are given to verify the effectiveness of the proposed system.

Key words: speech recognition; dynamic time warping; spectral subtraction; embedded systems