



(12) 发明专利

(10) 授权公告号 CN 103198833 B

(45) 授权公告日 2015. 10. 21

(21) 申请号 201310075089. 3

(22) 申请日 2013. 03. 08

(73) 专利权人 北京理工大学

地址 100081 北京市海淀区中关村南大街 5 号

(72) 发明人 罗森林 谢尔曼 潘丽敏

(51) Int. Cl.

G10L 17/02(2013. 01)

G10L 17/14(2013. 01)

(56) 对比文件

CN 101770774 A, 2010. 07. 07,

US 2003009333 A1, 2003. 01. 09,

US 6539352 B1, 2003. 03. 25,

范小春 邱政权. 基于 HAAR 小波的分级说话人辨识. 《计算机工程与应用》. 2010,

范小春 邱政权. 说话人识别中的 HOCOR 和改

进的 MCE. 《科学技术与工程》. 2008,

谢尔曼 罗森林 潘丽敏. 基于 Haar 特征的 Turbo-Boost 表情识别算法. 《计算机辅助设计与图形学学报》. 2011,

审查员 王玥

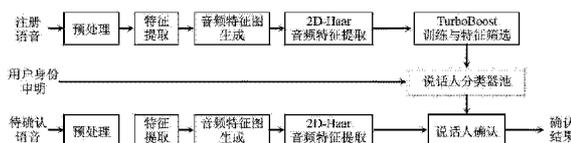
权利要求书4页 说明书12页 附图2页

(54) 发明名称

一种高精度说话人确认方法

(57) 摘要

本发明涉及一种基于文本无关说话人确认方法。本发明提出了 Turbo-Boost 分类算法与 2D-Haar 音频特征的相结合的说话人确认方法, 首先使用基础音频特征构成音频特征图; 进而利用音频特征图提取 2D-Haar 音频特征, 再使用 Turbo-Boost 算法, 通过两轮迭代运算分别完成对 2D-Haar 音频特征的筛选和说话人分类器的训练; 最终使用训练好的说话人分类器实现说话人确认。与现有技术相比, 本发明可以在同样的运算消耗下获得更高的准确率, 特别适合对于运算速度和说话人确认精度有着严格要求的说话人确认场合, 例如电话自动接听系统、计算机身份认证系统、高密级门禁系统等。



1. 一种高精度说话人确认方法,其特征在于,所述方法包括以下步骤:

步骤 1, 获取待确认说话人的语音信号, 形成基础语音库 S;

步骤 2, 对基础语音库 S 中的语音进行音频特征积分图计算, 形成基础特征库 R, 所述音频特征积分图计算的步骤具体包括:

步骤 2.1, 对于第 k 个待确认说话人, 对其音频文件 s_k 进行分帧处理, 帧长 f_s 、帧移 Δf_s 由用户设定, 并提取各帧的基础音频特征, 将各帧的基础音频特征组合, 形成一个包含 c 帧、每帧 p 维特征量的基础特征文件 v_k ;

v_k 中每一帧的特征向量的内容为: {[基础特征 1 (p_1 维)], [基础特征 2 (p_2 维)], ..., [基础特征 n (p_n 维)]},

对于一个时长为 t 的音频文件 s_k :

$$c = \left\lfloor \frac{t}{f_s - \Delta f_s} \right\rfloor, \quad p = \sum_{i=1}^n p_i,$$

步骤 2.2, 对于第 k 个待确认说话人的基础特征文件 v_k , 采用滑窗的方式, 以 a 为窗长、s 为步进, 将所有的 c 帧音频特征向量转换成音频特征图序列文件 G_k ;

$$G_k = \{g_1, g_2, g_3, \dots, g_u\}, \text{ 其中, } u = \left\lfloor \frac{c}{a-s} \right\rfloor;$$

步骤 2.3, 在步骤 2 的基础上, 计算对于第 k 个待确认说话人的特征图序列文件 G_k 中每幅音频特征图 g_u 的特征积分图 r_u , 形成该说话人的特征积分图序列文件 $R_k = \{r_1, r_2, r_3, \dots, r_u\}$, 将基础语音库 S 中所有 k 个待确认说话人的特征积分图序列文件集中起来, 形成基础特征库 $R = \{R_1, R_2, \dots, R_k\}$,

所述的特征积分图与原始的音频特征图尺寸相同, 其上任意一点 (x, y) 的值被定义为原始的音频特征图上对应点 (x', y') 及其左上方所有的特征值之和, 定义式如下:

$$ii(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'),$$

式中 $ii(x, y)$ 表示积分图上点 (x, y) 的取值, $i(x', y')$ 表示原始的音频特征图上点 (x', y') 的特征值;

步骤 3, 在基础特征库 R 的基础上, 生成每个待确认说话人的训练特征文件集 B;

步骤 4, 在步骤 3 的基础上, 提取 2D-Haar 音频特征, 并进行说话人注册, 也就是依次遍历特征文件集 B 中的 k 个文件夹, 并使用其中的训练特征文件为每个待确认说话人训练出单独的“1 对余”分类器, 最终得到由 k 个说话人分类器构成的分类器池, 所述“2D-Haar 音频特征”的计算方法为:

每维 2D-Haar 音频特征的取值都是在原始的音频特征图上, 任意尺寸、位置的方形区域中, 使用某一特定矩形区域的特征值之和减去另一个特定矩形区域的特征值之和, 可通过积分图快速计算获得, 其总维数 H 由采用的 2D-Haar 音频特征类型以及积分图的尺寸决定;

将每幅积分图相应的 H 维 2D-Haar 音频特征向量记作一行, 使特征文件夹 B_k 中所有 m 幅音频特征积分图的全部 H 维 2D-Haar 音频特征向量构成一个 m 行、H 列的特征矩阵 X;

步骤 5, 对用户提供的、申明是说话人 k 发声录制的语音文件, 提取其 2D-Haar 音频特征, 输入步骤 4 训练得到的说话人 k 的分类器, 以确认该文件中的语音是否确实由用户所申

明的说话人讲出。

2. 根据权利要求 1 所述的方法, 其特征在于, 所述获取待确认说话人的语音信号并不要求说话人按照特征模板中预置文本内容进行发音。

3. 根据权利要求 1 所述的方法, 其特征在于, 所述由 k 个说话人分类器构成的分类器池, 需通过 k 轮训练得到, 每轮训练都要包括两轮迭代过程: 第 1 轮进行 F 轮迭代, 从 H 维 2D-Haar 音频特征值集合中选择 F 维主要特征以完成特征筛选, 得到新的 F 维特征子空间; 第 2 轮进行 T 轮迭代, 在新的 F 维特征子空间中训练得到 T 个弱分类器, $T > F$, 将其组成强分类器, 具体方法为:

步骤 1, 初始化每幅积分图对应的权重, 记作 $D_1(i, l_i) = 1/(mk)$, $i = 1 \cdots m, l_i \in Y$, 表示第 i 个积分图所对应的说话人标签, $Y = \{1, 2, \cdots, k\}$ 是目标说话人标签集, k 为目标说话人数目, m 为音频特征积分图的数量;

步骤 2, 依次将记作特征矩阵 X 各列数据的、所有积分图的 H 组同维特征作为一个弱分类器的输入, 进行 H 轮运算, 按照下式计算 $r_{f,j}$ 的值:

$$r_{f,j} = \sum_{i \in (i,l)} D_f(i, l_i) K_i[l_i] h_j(x_i, l_i), \quad j = 1 \dots H$$

其中, $h_j(x_i, l_i)$ 表示以第 i 个积分图中提取的第 j 维特征值作为输入的弱分类器, $D_f(i, l_i)$ 表示第 f 轮迭代中第 i 个训练积分图的权重值, $K_i[l_i] = \begin{cases} +1 & l_i \in [1, \dots, k] \\ -1 & l_i \notin [1, \dots, k] \end{cases}$;

从上述 H 个弱分类器中选择一个 $h_j(x, l_i)$, 使得 $r_f = \max(r_{f,j})$, 将该分类器对应的特征 $f_j(x)$ 作为选中的特征维加入到新的特征空间; 其中, $f_j(x)$ 表示 H 维 2D-Haar 音频特征向量的第 j 维, $h_j(x, l)$ 表示采用第 j 维特征值作为输入的弱分类器;

步骤 3, 计算由步骤 2 选择出的弱分类器 $h_j(x, l)$ 的权重 α_f :

$$\alpha_f = \frac{1}{2} \ln \left(\frac{1+r_f}{1-r_f} \right);$$

步骤 4, 计算下一轮迭代中各个积分图的权重 D_{f+1} :

$$D_{f+1} = \frac{D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i))}{Z_f}, \quad i=1 \dots m.$$

其中, $h_f(x_i, l_i)$ 表示第 f 轮迭代中以第 i 个积分图提取的第 j 维特征值作为输入的弱分类器, Z_f 是归一化因子

$$Z_f = \sum_{i,l} D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i)) \quad i=1 \dots m.$$

步骤 5, 将步骤 4 得到的新权重代入步骤 2, 按照步骤 2 至步骤 4 的方法, 选中一个新的特征维, 同时得到一个新的弱分类器添加到强分类器中;

步骤 6, 按照步骤 2 至步骤 5 的方法迭代 F 次, 从特征矩阵 X 中提取 F 列, 形成一个 m 行、 F 列的主要特征矩阵 X' , 并重新初始化每幅积分图对应的权重, 记作

$$D'_1(i, l_i) = 1/(mk), \quad i = 1 \cdots m, l_i \in Y;$$

步骤 7, 依次将记作主要特征矩阵 X' 各列数据的、所有图像的 F 组同维特征作为一个弱分类器的输入, 进行 F 轮运算, 按照下式计算 $r_{t,j}$ 的值:

$$r_{t,j} = \sum_{i,l} D'_t(i,l) K_i[l_i] h_j(x_i, l_i),$$

从F个弱分类器中选择一个 $h_j(x, l)$, 使得 $r_t = \max(r_{t,j})$; 将该弱分类器记作 $h_t(x, l)$, 添加到强分类器中; 其中 $D'_t(i, l)$ 表示第t轮迭代中第i个训练图像的权重值;

步骤8, 计算通过步骤7选择出的弱分类器 $h_j(x, l)$ 的权重 α_t :

$$\alpha_t = \frac{1}{2} \ln \left(\frac{1+r_t}{1-r_t} \right),$$

步骤9, 计算下一轮迭代中各个图像的权重 D'_{t+1} :

$$D'_{t+1} = \frac{D'_t(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i))}{Z_t}, i=1 \dots m.$$

其中, Z_t 是归一化因子

$$Z_t = \sum_{i,l} D'_t(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i)), i=1 \dots m.$$

步骤10, 将步骤9得到的新权重代入步骤7, 按照步骤7至步骤9的方法, 得到一个新的弱分类器添加到强分类器中;

按照上述步骤7至步骤10的方法进行T轮迭代, 得到由T个弱分类器组成的强分类器, 即第k个说话人的确认分类器, 表示为:

$$W_k(x) = \arg \max_l S_l, \quad S_l = (\sum_{t=1}^T \alpha_t h_t(x, l)) \quad (1).$$

4. 根据权利要求3所述的方法, 其特征在于, 所述迭代运算中所使用的弱分类器, 需满足以下条件: 1. 弱分类器的输入是单维特征值, 即特征向量中的某一特定维, 或特征矩阵X中的某一列; 2. 针对待确认的说话人标签 1_i , 弱分类器的输出是1或-1。

5. 根据权利要求1所述的方法, 其特征在于, 所述步骤5的具体步骤为:

步骤1, 对确认语音文件进行音频特征积分图提取, 得到待确认音频特征积分图序列 $G' = \{g'_1, g'_2, g'_3, \dots, g'_{u'}\}$, u' 表示待确认音频特征积分图序列中, 音频特征积分图的数量, 对于一个包含 c' 帧的待确认语音文件, 音频特征积分图序列包含的音频积分特征图的数量 u' 为:

$u' = \left\lfloor \frac{c'}{a-s} \right\rfloor$, a 表示生成音频特征图过程中设定的窗长, s 表示同一过程中, 滑窗移动的步进;

步骤2, 在步骤1的基础上, 为特征图序列中的每幅特征图提取2D-Haar音频特征, 构成2D-Haar音频特征矩阵 X' ;

步骤3, 从说话人分类器池中找到申明说话人k的说话人分类器 W_k , 再把步骤2得到的2D-Haar音频特征矩阵 X' 输入 W_k , 得到分类结果序列 R ;

步骤4, 对步骤3得到的分类结果序列进行结果综合, 得到最终的说话人确认结果。

6. 根据权利要求5所述的方法, 其特征在于, 所述分类结果序列 R 由 u' 个元素组成, 其中每个元素的具体计算方法为:

步骤1, 按照权利要求5步骤10中的(1)式, 读取说话人分类器中某个弱分类器

$h_t(x, 1)$ 及其相应 2D-Haar 音频特征 $f_j(x)$;

步骤 2, 对于每种待选标签 $l_i \in \{k, \text{other}\}$, 分别计算该弱分类器的输出 $h_t(f_j(x), 1)$, 并将该输出值以分类器中的权重 α_t 累加到待选标签 l_i 对应的加权值 S_{l_i} 中 ;

步骤 3, 按照步骤 1- 步骤 2 的方法进行 T 轮循环之后, 每种待选标签 l_i 将得到一个加权值 S_{l_i} , 选出取值最大的一个加权值 S_{l_i} , 同时记录与其相对应的待选标签 l_i 作为该音频特征图的分类结果, 记作 $(l_i, S_{l_i, u})$, 其中 l_k 为说话人标签, $S_{l_i, u}$ 为相应的强分类器加权值 ;

步骤 4, 将待确认音频的所有分类结果组合起来, 构成分类结果序列 $\mathbf{R} = \{(l_i, S_{l_i, u}) : (l_1, S_{l_1, 1}), (l_1, S_{l_1, 2}), (l_2, S_{l_2, 3}), \dots, (l_i, S_{l_i, u})\}$, $l_i \in \{k, \text{other}\}$ 。

7. 根据权利要求 5 所述的方法, 其特征在于, 所述“结果综合”的计算方法为 :

步骤 1, 统计结果序列中所有的强分类器判别权重 $S_{l_i, u}$ 按说话人标签 l_i 加权, 即分别

求出 $S_k = \sum_1^u S_{k, u}$ 和 $S_{\text{other}} = \sum_1^u S_{\text{other}, u}$;

步骤 2, 计算置信因子 $\eta = S_k / (S_{\text{other}} + S_k)$, 按照下式给出最终的说话人确认结果 V :

$$V = \begin{cases} \text{True}, & \text{if } \eta > \omega \\ \text{False}, & \text{if } \eta \leq \omega \end{cases}$$

式中 ω 为判别阈值, 可由用户指定。

一种高精度说话人确认方法

技术领域

[0001] 本发明涉及一种高精度的文本无关说话人确认方法,属于生物识别技术领域;从技术实现的角度来讲,亦属于计算机科学与语音处理技术领域。

背景技术

[0002] 说话人确认 (Speaker Verification) 技术是利用每个说话人的语音信号特点,从一段语音中提取说话人信息,进而确认某段语音是否是指定的某个人所说的,系统只给出“接受”或“拒绝”两种选择,是“一对一”的模式识别问题。

[0003] 说话人确认技术与说话人辨认技术同属说话人识别 (Speaker Recognition, SR) 的范畴,而与说话人辨认技术不同,说话人确认技术对于准确率、识别时间的要求更为严格,近年来,电话自动接听系统、计算机身份认证系统、高密级门禁系统等应用平台对这项技术的应用需求越来越强。

[0004] 按照说话内容的类型不同,说话人确认可以分为文本有关 (Text-dependent) 和文本无关 (Text-independent) 两大类。与文本有关的说话人确认系统要求用户按照规定的内容发音,每个人的识别模型逐个被精确地建立,而识别时也必须按规定的內容发音;文本无关的识别系统则不规定说话人的发音内容,模型建立相对困难,可应用范围较宽。有些情况下,人们无法(或者不希望)强迫说话人朗读一段特定的文字,在这些应用场景中,文本无关的说话人确认方法就显得格外重要。

[0005] 文本无关的说话人确认的基本技术可分为语音采集,特征提取,分类方法三个层次,其中关键问题在于特征提取与分类方法。

[0006] 特征提取方面,目前的主流方法多采用基于底层声学原理的梅尔倒谱系数 (MFCC) 或线性预测倒谱系数 (Linear Predictive Coding Cepstrum, LPCC) 作为特征参数。

[0007] 分类方法方面,主流方法有动态时间规整 (DTW)、矢量量化 (VQ)、隐马尔可夫模型 (HMM)、高斯混合模型 (GMM)、人工神经网络 (ANN)、支撑向量机 (SVM) 等。目前广泛受到研究的是高斯混合模型 (GMM) 方法以及支撑向量机 (SVM) 方法。上述方法中,GMM-UBM 模型已经得到广泛应用,在更早的系统中,矢量量化也是一项获得了广泛研究的重要的技术。

[0008] 基于上述方法,文本无关的说话人确认技术已经在一些场合得到实际应用。然而,当待确认的人数不断增加时,上述方法的准确率会明显下降,当人数增加到一定规模时,将难以满足实际应用的需求,这是文本无关说话人确认技术需要解决的一个重要问题。

发明内容

[0009] 本发明的目标是:提出一种大规模说话人确认方法,能在获得高准确率的同时兼顾高运算速度的要求。具体实施方法上,本发明从特征提取和分类方法两个层次分别提出新的方法,提高特征的区分度,提升说话人分类器的速度与准确率。

[0010] 本发明的设计原理为:在特征提取层次,提出 2D-Haar 音频特征提取方法,引入一定的时序关系信息,并将音频特征空间扩展至数十万维,为确认算法提供更加庞大的特征

空间；在说话人分类器层次，提出 Turbo-Boost 算法，在庞大的 2D-Haar 特征空间中筛选具有代表性的特征组合，用于构建目标说话人的确认分类器。在相同的时间内，本发明可以将既有的识别准确率进一步提升，以满足说话人确认应用中快速、准确的技术要求。

[0011] 本发明的技术方案是通过如下步骤实现的：

[0012] 步骤 1，获取待确认说话人（即目标说话人）的语音信号，形成基础语音库 S。

[0013] 具体方法为：把麦克风与计算机连接，获取目标说话人的语音信号，并以音频文件的形式存储在计算机内，每个目标说话人对应一个音频文件，形成基础语音库 $S = \{s_1, s_2, s_3, \dots, s_k\}$ ，其中 k 为目标说话人的总数。

[0014] 步骤 2，对基础语音库 S 中的语音进行音频特征积分图计算，形成基础特征库 R。具体过程如下：

[0015] 步骤 2.1，对于第 k 个目标说话人，对其音频文件 s_k 进行分帧处理（帧长 f_s 、帧移 Δf_s 由用户设定），并提取各帧的基础音频特征（如 MFCC、LPCC、子带能量等），将各帧的基础音频特征组合，形成一个包含 c 帧、每帧 p 维特征量的基础特征文件 v_k 。

[0016] v_k 中每一帧的特征向量的内容为：{ [基础特征 1 (p_1 维)]， [基础特征 2 (p_2 维)]， \dots ， [基础特征 n (p_n 维)] }。

[0017] 以上描述中，对于一个时长为 t 的音频文件 s_k ：

$$[0018] \quad c = \left\lfloor \frac{t}{f_s - \Delta f_s} \right\rfloor, \quad p = \sum_{1}^n p_n.$$

[0019] 步骤 2.2，对于第 k 个目标说话人的基础特征文件 v_k ，采用滑窗的方式，以 a 为窗长、s 为步进，将所有的 c 帧音频特征向量转换成音频特征图序列文件 G_k （参见图 2）。

$$[0020] \quad G_k = \{g_1, g_2, g_3, \dots, g_u\}, \quad \text{其中}, \quad u = \left\lfloor \frac{c}{a-s} \right\rfloor.$$

[0021] 步骤 2.3，在步骤 2.2 的基础上，计算对于第 k 个目标说话人的特征图序列文件 G_k 中每幅特征图 g_u 的特征积分图 r_u ，形成该说话人的特征积分图序列文件 $R_k = \{r_1, r_2, r_3, \dots, r_u\}$ ，将基础语音库 S 中所有 k 个目标说话人的特征积分图序列文件集中起来，形成基础特征库 $R = \{R_1, R_2, \dots, R_k\}$ 。

[0022] 易知，基础特征库中所有说话人的特征积分图总数 m 的计算公式为：

[0023]

$$m = \sum_{1}^k u_k = \left\lfloor \left[\frac{\sum_{k} t_k}{k} \right] / (a-s) \right\rfloor.$$

[0024] 所述的特征积分图与原始特征图尺寸相同，其上任意一点 (x, y) 的值被定义为原图对应点 (x', y') 及其左上方所有的特征值之和。定义式如下：

$$[0025] \quad \ddot{i}(x, y) = \sum_{x' \leq x, y' \leq y} \dot{i}(x', y'),$$

[0026] 式中 $\ddot{i}(x, y)$ 表示积分图上点 (x, y) 的取值， $\dot{i}(x', y')$ 表示原始特征图上点 (x', y') 的特征值。

[0027] 步骤 3，在基础特征库 R 的基础上，生成每个目标说话人的训练特征文件集 B。具体过程如下：

[0028] 步骤 3.1, 对基础特征库 R 中的特征文件进行标注, 具体方法为:

[0029] 使用连续的整数编号作为说话人标签, 代表不同的目标说话人, 以便计算机处理。最终的标记形式为 $R' = \{(R_1, 1), (R_2, 2), \dots (R_k, k)\}$, 其中, $Y = \{1, 2, \dots, k\}$ 是目标说话人标签集, k 为目标说话人数目;

[0030] 步骤 3.2, 在步骤 3.1 的基础上, 为每个目标说话人建立用于说话人注册的特征文件集 B, 具体方法为:

[0031] 在标记好说话人标签的特征库 R' 中, 进行 k 轮整理, 在每轮整理工作中, 首先将第 k 个目标说话人的音频特征文件 r_k 作为正样本, 保留其说话人标签 k; 然后将其余的说话人音频特征文件作为负样本, 并将它们的说话人标签更改为“other”; 最后将上述 k 个音频特征文件存储到单独的文件夹中, 并将该特征文件夹命名为 B_k , 即:

[0032] $B_1 = \{(R_1, 1), (R_2, \text{other}), \dots (R_k, \text{other})\}$,

[0033] $B_2 = \{(R_1, \text{other}), (R_2, 2), \dots (R_k, \text{other})\}$,

[0034]

[0035] $B_k = \{(R_1, \text{other}), (R_2, \text{other}), \dots (R_k, k)\}$

[0036] k 轮整理工作之后, 最终形成由 k 个特征文件夹构成的特征文件集 $B = \{B_1, B_2, \dots, B_k\}$ 。

[0037] 步骤 4, 在步骤 3 的基础上, 提取 2D-Haar 音频特征, 并进行说话人注册, 也就是依次遍历特征文件集 B 中的 k 个文件夹, 并使用其中的训练特征文件为每个目标说话人训练出单独的“1 对余”分类器, 最终得到由 k 个说话人分类器构成的分类器池。

[0038] 对于第 k 个目标说话人, 其对应的分类器 W_k 的训练过程如下:

[0039] 步骤 4.1, 对步骤 3.2 所形成的特征文件夹 B_k 中的所有特征积分图序列文件 R_k 的每幅积分图进行 2D-Haar 音频特征提取。具体方法为:

[0040] 根据各个积分图计算相对应的 H 维 2D-Haar 音频特征值 (其中 H 由采用的 2D-Haar 音频特征类型以及积分图的尺寸决定), 得到用于说话人分类器训练的数据集合 $S = \{(x_1, l_1), \dots, (x_m, l_1)\}$ 。其中, x_i 表示第 i 个积分图所对应的全部 H 维 2D-Haar 音频特征向量, $l_i \in Y$, ($Y = \{1, 2, \dots, k\}$) 表示第 i 个积分图所对应的说话人标签。

[0041] 所述的 H 维 2D-Haar 音频特征值, 每维 2D-Haar 音频特征的取值是原始音频特征图上, 任意尺寸、位置的方形区域中, 使用某一特定矩形区域的特征值之和减去另一个特定矩形区域的特征值之和, 可通过积分图快速计算获得。

[0042] 将每幅积分图相应的 H 维 2D-Haar 音频特征向量记作一行, 使特征文件夹 B_k 中所有 m 幅积分图的全部 H 维 2D-Haar 音频特征向量构成一个 m 行、H 列的特征矩阵 X。

[0043] 步骤 4.2, 使用 Turbo-Boost 方法对步骤 4.1 得到的 2D-Haar 音频特征矩阵 X 进行特征筛选和分类器训练, 得到说话人分类器。所述的 Turbo-Boost 方法包括两轮迭代过程: 第 1 轮进行 F 轮迭代, 从 H 维 2D-Haar 音频特征值集合中选择 F 维主要特征以完成特征筛选, 得到新的 F 维特征子空间; 第 2 轮进行 T 轮迭代, 在新的 F 维特征子空间中训练得到 T 个弱分类器 ($T > F$), 将其组成强分类器。

[0044] 上述迭代运算中所使用的弱分类器, 需满足以下条件: 1. 弱分类器的输入是单维特征值 (即特征向量中的某一特定维, 或特征矩阵 X 中的某一行); 2. 针对待确认的说话人标签 l_i , 弱分类器的输出是 1 或 -1。

[0045] Turbo-Boost 的具体训练过程为：

[0046] 步骤 4.2.1, 初始化每幅积分图对应的权重, 记作 $D_1(i, l_i) = 1/(mk)$, $i=1 \dots m$, $l_i \in Y$ 。

[0047] 步骤 4.2.2, 依次将特征矩阵 X 的各列数据 (即所有积分图的 H 组同维特征) 作为一个弱分类器的输入, 进行 H 轮运算, 按照下式计算 $r_{f,j}$ 的值：

$$[0048] \quad r_{f,j} = \sum_{j,(i,l)} D_f(i, l_i) K_i[l_i] h_j(x_i, l_i), \quad j=1 \dots H$$

[0049] 其中, $h_j(x_i, l_i)$ 表示以第 i 个积分图中提取的第 j 维特征值作为输入的弱分类器,

$D_f(i, l_i)$ 表示第 f 轮迭代中第 i 个训练积分图的权重值, $K_i[l_i] = \begin{cases} +1 & l_i \in [1, \dots, k] \\ -1 & l_i \notin [1, \dots, k] \end{cases}$ 。

[0050] 从上述 H 个弱分类器中选择一个 $h_j(x, l_i)$, 使得 $r_f = \max(r_{f,j})$, 将该分类器对应的特征 $f_j(x)$ 作为选中的特征维加入到新的特征空间。其中, $f_j(x)$ 表示 H 维 2D-Haar 音频特征向量的第 j 维 (即特征矩阵 X 的第 j 列), $h_j(x, l)$ 表示采用第 j 维特征值作为输入的弱分类器；

[0051] 步骤 4.2.3, 计算由步骤 4.2.2 选择出的弱分类器 $h_j(x, l)$ 的权重 α_f ：

$$[0052] \quad \alpha_f = \frac{1}{2} \ln \left(\frac{1+r_f}{1-r_f} \right);$$

[0053] 步骤 4.2.4, 计算下一轮迭代中各个积分图的权重 D_{f+1} ：

$$[0054] \quad D_{f+1} = \frac{D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i))}{Z_f}, \quad i=1 \dots m.$$

[0055] 其中, $h_f(x_i, l_i)$ 表示第 f 轮迭代中以第 i 个积分图提取的第 j 维特征值作为输入的弱分类器, Z_f 是归一化因子

$$[0056] \quad Z_f = \sum_{i,l} D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i)) \quad i=1 \dots m.$$

[0057] 步骤 4.2.5, 将步骤 4.2.4 得到的新权重代入步骤 4.2.2, 按照步骤 4.2.2 至步骤 4.2.4 的方法, 选中一个新的特征维；

[0058] 步骤 4.2.6, 按照步骤 4.2.2 至步骤 4.2.5 的方法迭代 F 次, 从特征矩阵 X 中提取 F 列, 形成一个 m 行、 F 列的主要特征矩阵 X' , 并重新初始化每幅积分图对应的权重, 记作

$$[0059] \quad D'_1(i, l_i) = 1/(mk), \quad i=1 \dots m, \quad l_i \in Y.$$

[0060] 步骤 4.2.7, 依次将主要特征矩阵 X' 的各列数据 (即所有图像的 F 组同维特征) 作为一个弱分类器的输入, 进行 F 轮运算, 按照下式计算 $r_{t,j}$ 的值：

$$[0061] \quad r_{t,j} = \sum_{j,(i,l)} D'_t(i, l_i) K_i[l_i] h_j(x_i, l_i),$$

[0062] 从 F 个弱分类器中选择一个 $h_j(x, l)$, 使得 $r_t = \max(r_{t,j})$; 将该弱分类器记作 $h_t(x, l)$, 添加到强分类器中。其中 $D'_t(i, l)$ 表示第 t 轮迭代中第 i 个训练图像的权重值。

[0063] 步骤 4.2.8, 计算通过步骤 4.2.7 选择出的弱分类器 $h_j(x, l)$ 的权重 α_t ：

$$[0064] \quad \alpha_t = \frac{1}{2} \ln \left(\frac{1+r_t}{1-r_t} \right),$$

[0065] 步骤 4.2.9, 计算下一轮迭代中各个图像的权重 D'_{t+1} ;

$$[0066] \quad D'_{t+1} = \frac{D_t^+(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i))}{Z_t}, \quad i=1 \dots m.$$

[0067] 其中, Z_t 是归一化因子

$$[0068] \quad Z_t = \sum_{i,l} D_t^+(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i)), \quad i=1 \dots m.$$

[0069] 步骤 4.2.10, 将步骤 4.2.9 得到的新权重代入步骤 4.2.7, 按照步骤 4.2.7 至步骤 4.2.9 的方法, 得到一个新的弱分类器添加到强分类器中;

[0070] 按照上述步骤 4.2.7 至步骤 4.2.10 的方法进行 T 轮迭代, 得到由 T 个弱分类器组成的强分类器, 即第 k 个说话人的确认分类器, 表示为:

$$[0071] \quad W_k(x) = \underset{l}{\operatorname{argmax}} S_l, \quad S_l = \left(\sum_{t=1}^T \alpha_t h_t(x, l) \right) \quad (1)$$

[0072] 步骤 4.2.11, 待 k 轮训练结束后, 将所有的 k 个说话人分类器集合起来, 构成说话人分类器池 $W = \{W_1(x), W_2(x), \dots, W_k(x)\}$ 。

[0073] 步骤 5, 对用户提供的、申明是说话人 k 发声录制的语音文件, 提取其 2D-Haar 音频特征, 输入步骤 4 训练得到的说话人 k 的分类器, 以确认该文件中的语音是否确实由用户所申明的说话人讲出。具体步骤为:

[0074] 步骤 5.1, 对确认语音文件进行音频特征积分图提取, 得到待确认音频特征积分图序列 $G' = \{g'_{1}, g'_{2}, g'_{3}, \dots, g'_{u'}\}$ 。具体方法与步骤 2 所述方法相同。其中, 音频特征图序列转换过程中(对应于步骤 2.2), 窗长 a、步进 s 的取值与步骤 2 中的相同; 类似的, 对于一个包含 c' 帧的待确认语音文件, 特征图序列包含的特征图数量 u' 为: $u' = \left\lfloor \frac{c'}{a-s} \right\rfloor$ 。

[0075] 步骤 5.2, 在步骤 5.1 的基础上, 根据步骤 4.1 所述的 2D-Haar 音频特征提取方法, 以及步骤 4.2.5 的特征筛选结果, 为特征图序列中的每幅特征图提取 F 维 2D-Haar 音频特征, 构成 2D-Haar 音频特征矩阵 X' 。

[0076] 步骤 5.3, 从说话人分类器池中找到申明说话人 k 的说话人分类器 W_k , 再把步骤 5.2 得到的 2D-Haar 音频特征矩阵 X' 输入 W_k 得到分类结果序列 R。

[0077] 所述分类结果序列 R 由 u' 个元素组成, 其中每个元素的具体计算方法为:

[0078] 步骤 5.3.1, 按照步骤 4.2.10 中的(1)式, 读取说话人分类器中某个弱分类器 $h_t(x, l)$ 及其相应 2D-Haar 音频特征 $f_j(x)$;

[0079] 步骤 5.3.2, 对于每种待选标签 $l_i \in \{k, \text{other}\}$, 分别计算该弱分类器的输出 $h_t(f_j(x), l)$, 并将该输出值以分类器中的权重 α_t 累加到待选标签 l_i 对应的加权值 S_{li} 中;

[0080] 步骤 5.3.3, 按照步骤 5.3.1-步骤 5.3.2 的方法进行 T 轮循环之后, 每种待选标签 l_i 将得到一个加权值 S_{li} 。选出取值最大的一个加权值 S_{li} , 同时记录与其相对应的待选标签 l_i 作为该音频特征图的分类结果, 记作 $(l_i, S_{li, u'})$, 其中 l_k 为说话人标签, $S_{li, u'}$ 为相应的强分类器加权和。

[0081] 步骤 5.3.4, 将待确认音频的所有分类结果组合起来, 构成分类结果序列 $R = \{ (l_i, S_{1i,u}) : (l_1, S_{11,1}), (l_1, S_{11,2}), (l_2, S_{12,3}), \dots (l_i, S_{1i,u}) \}$, $l_i \in \{k, \text{other}\}$ 。

[0082] 步骤 5.4, 对步骤 5.3 得到的分类结果序列进行结果综合, 得到最终的说话人确认结果。

[0083] 具体方法为:

[0084] 步骤 5.4.1, 统计结果序列中所有的强分类器判别权重 $S_{1i,u}$ 按说话人标签 l_i 加权, 即分别求出 $S_k = \sum_{\mathbf{I}} S_{k,u}$ 和 $S_{\text{other}} = \sum_{\mathbf{I}} S_{\text{other},u}$

[0085] 步骤 5.4.2, 计算置信因子 $\eta = S_k / (S_{\text{other}} + S_k)$, 按照下式给出最终的说话人确认结果 V :

$$[0086] \quad V = \begin{cases} \text{True} & \text{if } \eta > \omega \\ \text{False} & \text{if } \eta \leq \omega \end{cases}$$

[0087] 式中 ω 为判别阈值, 可由用户指定。

[0088] 有益效果

[0089] 相比于基于底层声学原理的梅尔倒谱系数 (MFCC) 或线性预测倒谱系数 (LPCC) 等特征参数提取方法, 本发明提出的 2D-Haar 音频特征提取方法引入 $1e$ 一定的时序关系信息, 并将音频特征空间扩展至数十万维, 为确认算法提供更加庞大的特征空间。

[0090] 与 GMM、SVM 等说话人分类方法相比, 本发明采用使用 Turbo-Boost 算法, 结合单特征输入的 Decision Stump 弱分类器进行特征筛选, 大大减少了说话人确认阶段的计算负担, 在相同运算开销下, 具有更高的准确率, 可以满足说话人确认“快速、准确”的实用要求, 具有较高的实用价值。

附图说明

[0091] 图 1 为本发明的原理框图;

[0092] 图 2 为本发明提出的音频特征图和特征图序列提取原理意图;

[0093] 图 3 为本发明的说话人注册过程原理图;

[0094] 图 4 为本发明的说话人确认过程原理图;

[0095] 图 5 为具体实施方式中说话人训练及确认过程中所使用的 5 类 2D-Haar 音频特征;

[0096] 图 6 为具体实施方式中, 使用 TIMIT 语音库进行测试时, 本发明与 GMM-UBM 算法、AdaBoost.MH 算法的性能比对。

具体实施方式

[0097] 为了更好的说明本发明的目的和优点, 下面结合附图和实施例对本发明方法的实施方式做进一步详细说明。

[0098] 以下所有测试均在同一台计算机上完成, 具体配置为: Intel 双核 CPU (主频 1.8G), 1G 内存, WindowsXP SP3 操作系统。

[0099] 第一环节

[0100] 本环节将使用 TIMIT 音频库的语音文件, 详细说明当目标说话人规模为 200 人时,

本发明的说话人注册 / 训练、说话人确认的具体过程。

[0101] TIMIT 语音库是由麻省理工大学、斯坦福研究院、德州仪器联合制作的标准库, 包含了 630 个说话人 (438 个男性和 192 个女性) 的语料, 每个人 10 条语音。

[0102] 从所有说话人中随机选取 200 人的全部语音数据, 再从每个人的 10 条语音中选取 1 条持续时间大于 5 秒的文件作为说话人注册 / 训练语音文件; 另外随机选取 1 个人的任意一条语音作为确认语音文件。

[0103] 具体实施步骤如下:

[0104] 步骤 1, 获取待确认说话人 (即目标说话人) 的语音信号, 形成基础语音库 S。

[0105] 由于 TIMIT 语音库已经是存储完整的音频文件, 因此直接将 200 条目标说话人的语音文件形成基础语音库 $S = \{s_1, s_2, s_3, \dots, s_k\}$, 其中 $k=200$ 为目标说话人的总数。

[0106] 步骤 2, 对基础语音库 S 中的语音进行音频特征积分图计算, 形成基础特征库 R。具体过程如下:

[0107] 步骤 2.1, 对于第 k 个目标说话人, 对其音频文件 s_k 进行分帧处理, 并提取各帧的基础音频特征 (本实施例中, 使用 MFCC, LPCC, PLPC), 将各帧的基础音频特征组合, 形成一个包含 c 帧、每帧 p 维特征量的基础特征文件 v_k 。

[0108] 本实施例中, v_k 中每一帧的特征向量的内容为: $\{[MFCC (12 维)], [LPCC (12 维)], [PLPC (8 维)]\}$, 分帧操作的帧长设定为 $f_s=30ms$, 帧移设定为 $\Delta f_s=20ms$ 。

[0109]

$$c = \left\lfloor \frac{t_k}{f_s - \Delta f_s} \right\rfloor, \quad p = \sum_1^n p_n = 12 + 12 + 8 = 32.$$

[0110] 步骤 2.2, 对于第 k 个目标说话人的基础特征文件 v_k , 采用滑窗的方式, 以 a 为窗长、s 为步进, 将所有的 c 帧音频特征向量转换成音频特征图序列文件 G_k (参见图 2)。本实施例中, $a=32, s=16$ 。

[0111] $G_k = \{g_1, g_2, g_3, \dots, g_{u_k}\}$, 其中, $u_k = \left\lfloor \frac{c}{a-s} \right\rfloor$ 。

[0112] 步骤 2.3, 在步骤 2.2 的基础上, 计算对于第 k 个目标说话人的特征图序列文件 G_k 中每幅特征图 g_u 的特征积分图 r_u , 形成该说话人的特征积分图序列文件 $R_k = \{r_1, r_2, r_3, \dots, r_{u_k}\}$, 将基础语音库 S 中所有 200 个目标说话人的特征积分图序列文件集中起来, 形成基础特征库 $R = \{R_1, R_2, \dots, R_k\}$ 。

[0113] 易知, 基础特征库中所有说话人的特征积分图总数 m 的计算公式为:

[0114]

$$m = \sum_1^k u_k = \left\lfloor \left\lfloor \frac{\sum_k t_k}{f_s - \Delta f_s} \right\rfloor / (a-s) \right\rfloor.$$

[0115] 本实施例中, 所有 200 个音频文件的总时长为 1202.30s, 因此:

[0116]

$$m = \left\lfloor \left\lfloor \frac{\sum_k t_k}{f_s - \Delta f_s} \right\rfloor / (a-s) \right\rfloor = \left\lfloor \left\lfloor \frac{1202.30}{0.03 - 0.02} \right\rfloor / (32 - 16) \right\rfloor = 7514$$

[0117] 所述的特征积分图与原始特征图尺寸相同,其上任意一点 (x, y) 的值被定义为原图对应点 (x', y') 及其左上方所有的特征值之和。定义式如下:

$$[0118] \quad \bar{i}(x, y) = \sum_{x' \leq x, y' \leq y} i(x', y'),$$

[0119] 式中 $\bar{i}(x, y)$ 表示积分图上点 (x, y) 的取值, $i(x', y')$ 表示原始特征图上点 (x', y') 的特征值。

[0120] 步骤 3, 在基础特征库 R 的基础上, 生成每个目标说话人的训练特征文件集 B。具体过程如下:

[0121] 步骤 3.1, 对基础特征库 R 中的特征文件进行标注, 具体方法为:

[0122] 使用连续的整数编号作为说话人标签, 代表不同的目标说话人, 以便计算机处理。最终的标记形式为 $R' = \{(R_1, 1), (R_2, 2), \dots, (R_{200}, 200)\}$, 其中, $Y = \{1, 2, \dots, 200\}$ 是目标说话人标签集;

[0123] 步骤 3.2, 在步骤 3.1 的基础上, 为每个目标说话人建立用于说话人注册的特征文件集 B, 具体方法为:

[0124] 在标记好说话人标签的特征库 R' 中, 进行 200 轮整理, 在每轮整理工作中, 首先将第 k 个目标说话人的音频特征文件 r_k 作为正样本, 保留其说话人标签 k; 然后将其余的说话人音频特征文件作为负样本, 并将它们的说话人标签更改为“other”; 最后将上述 200 个音频特征文件存储到单独的文件夹中, 并将该特征文件夹命名为 B_k , 即:

$$[0125] \quad B_1 = \{(R_1, 1), (R_2, \text{other}), \dots, (R_{200}, \text{other})\},$$

$$[0126] \quad B_2 = \{(R_1, \text{other}), (R_2, 2), \dots, (R_{200}, \text{other})\},$$

[0127]

$$[0128] \quad B_{200} = \{(R_1, \text{other}), (R_2, \text{other}), \dots, (R_{200}, 200)\}$$

[0129] 200 轮整理工作之后, 最终形成由 200 个特征文件夹构成的特征文件集 $B = \{B_1, B_2, \dots, B_{200}\}$ 。

[0130] 步骤 4, 在步骤 3 的基础上, 提取 2D-Haar 音频特征, 并进行说话人注册, 也就是依次遍历特征文件集 B 中的 200 个文件夹, 并使用其中的训练特征文件为每个目标说话人训练出单独的“1 对余”分类器。

[0131] 对于第 k 个目标说话人, 其对应的分类器 W_k 的训练过程如下:

[0132] 步骤 4.1, 对步骤 3.2 所形成的特征文件夹 B_k 中的所有特征积分图序列文件 R_k 的每幅积分图进行 2D-Haar 音频特征提取。

[0133] 根据各个积分图计算相对应的 H 维 2D-Haar 音频特征值, 得到用于说话人分类器训练的数据集合 $S = \{(x_1, l_1), \dots, (x_m, l_m)\}$ 。其中, x_i 表示第 i 个积分图所对应的全部 H 维 2D-Haar 音频特征向量, $l_i \in Y$, ($Y = \{1, 2, \dots, k\}$) 表示第 i 个积分图所对应的说话人标签。

[0134] 图 5 展示了本实施例使用的 5 类 2D-Haar 音频特征的计算模式, 每维 2D-Haar 音频特征的取值为: 原始音频特征图上, 任意尺寸、位置的方形区域上, 按照图 5 中某一类模式, 计算黑色区域的特征值之和减去白色区域的特征值之和。该特征具有如下三个特点:

[0135] 1. 运算速度快。配合积分图, 任何尺寸 2D-Haar 音频特征的提取只需执行固定次数的数据读取和加减运算。包含 2 个矩形的 2D-Haar 音频特征只需从积分图中读取 6 个点进行加 / 减运算, 3 个矩形的特征只需读取 8 个点, 4 个矩形的特征只需读取 9 个点。

[0136] 2. 区分性强。2D-Haar 音频特征空间的维数很高,以本实施例使用的 5 类模式为例,一幅 32×32 的积分图,5 类模式可以产生总维数超过了 51 万的 2D-Haar 音频特征,具体数量如表 1 所示。

[0137] 表 1 一幅 32×32 积分图 5 类 2D-Haar 音频特征的数量

	I 型	II 型	III 型	IV 型	V 型	总数
[0138]	135168	135168	87120	87120	65536	510112

[0139] 这一维数远远超过了音频 FFT 能量谱的原始信息,也远远超出了 SVM 非线性映射后特征空间的维度。此外,由于音频特征图是由一定数量的连续音频帧组成,因此 2D-Haar 音频特征也能反映一定的时序信息。

[0140] 在本实施例中,2D-Haar 音频特征提取的具体方法为:首先根据积分图和上述方法,计算所有的 510112 维 2D-Haar 音频特征值,得到 2D-Haar 音频特征值集合;进而将每幅积分图相应的 510112 维 2D-Haar 音频特征向量记作一行,使特征文件夹 B_k 中所有 m 幅积分图的全部 H 维 2D-Haar 音频特征向量构成一个 m 行、510112 列的特征矩阵 X ,如步骤 2.2 所示,在本实施例中, $m=7514$ 。

[0141] 步骤 4.2,使用 Turbo-Boost 方法对步骤 4.1 得到的 2D-Haar 音频特征矩阵 X 进行特征筛选和分类器训练,得到说话人分类器。所述的 Turbo-Boost 方法包括两轮迭代过程:第 1 轮进行 F 轮迭代,从 H 维 2D-Haar 音频特征值集合中选择 F 维主要特征以完成特征筛选,得到新的 F 维特征子空间;第 2 轮进行 T 轮迭代,在新的 F 维特征子空间中训练得到 T 个弱分类器 ($T > F$),将其组成强分类器。

[0142] 上述迭代运算中所使用的弱分类器,其定义式为:

$$[0143] \quad h_j(x,y) = \begin{cases} 1 & p_{j,y} x_j < p_{j,y} \theta_{j,y} \\ -1 & p_{j,y} x_j \geq p_{j,y} \theta_{j,y} \end{cases}, \quad (2)$$

[0144] 其中, x_j 表示弱分类器的输入, $\theta_{j,y}$ 表示训练后得到的阈值, $p_{j,y}$ 指示不等号的方向。

[0145] Turbo-Boost 的具体训练过程为(本实施例中,训练过程中所涉及的参数取值为: $H=510112$, $m=7514$, $F=200$, $T=400$, $Y=\{k, \text{other}\}$, $k=200$):

[0146] 步骤 4.2.1,初始化每幅积分图对应的权重,记作 $D_1(i, l_i) = 1/(mk)$, $i=1 \dots m$, $l_i \in Y$ 。

[0147] 步骤 4.2.2,依次将特征矩阵 X 的各列数据(即所有积分图的 H 组同维特征)作为一个弱分类器的输入,进行 H 轮运算,按照下式计算 $r_{f,j}$ 的值:

$$[0148] \quad r_{f,j} = \sum_{j(i,l)} D_f(i, l_i) K_i[l_i] h_j(x_i, l_i), \quad j=1 \dots H$$

[0149] 其中, $h_j(x_i, l_i)$ 表示以第 i 个积分图中提取的第 j 维特征值作为输入的弱分类器,

$D_f(i, l_i)$ 表示第 f 轮迭代中第 i 个训练积分图的权重值, $K_i[l_i] = \begin{cases} +1 & l_i \in [1, \dots, k] \\ -1 & l_i \notin [1, \dots, k] \end{cases}$ 。

[0150] 从上述 H 个弱分类器中选择一个 $h_j(x, l_i)$, 使得 $r_f = \max(r_{f,j})$, 将该分类器对应的特征 $f_j(x)$ 作为选中的特征维加入到新的特征空间。其中, $f_j(x)$ 表示 H 维 2D-Haar 音频特征向量的第 j 维(即特征矩阵 X 的第 j 列), $h_j(x, l)$ 表示采用第 j 维特征值作为输入的弱

分类器；

[0151] 步骤 4.2.3, 计算由步骤 4.2.2 选择出的弱分类器 $h_j(x, l)$ 的权重 α_f ;

$$[0152] \quad \alpha_f = \frac{1}{2} \ln \left(\frac{1+r_f}{1-r_f} \right);$$

[0153] 步骤 4.2.4, 计算下一轮迭代中各个积分图的权重 D_{f+1} ;

$$[0154] \quad D_{f+1} = \frac{D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i))}{Z_f}, i=1 \dots m.$$

[0155] 其中, $h_f(x_i, l_i)$ 表示第 f 轮迭代中以第 i 个积分图提取的第 j 维特征值作为输入的弱分类器, Z_f 是归一化因子

$$[0156] \quad Z_f = \sum_{i,l} D_f(i, l_i) \exp(-\alpha_f K_i[l_i] h_f(x_i, l_i)) \quad i=1 \dots m.$$

[0157] 步骤 4.2.5, 将步骤 4.2.4 得到的新权重代入步骤 4.2.2, 按照步骤 4.2.2 至步骤 4.2.4 的方法, 选中一个新的特征维;

[0158] 步骤 4.2.6, 按照步骤 4.2.2 至步骤 4.2.5 的方法迭代 F 次, 从特征矩阵 X 中提取 F 列, 形成一个 m 行、 F 列的主要特征矩阵 X' , 并重新初始化每幅积分图对应的权重, 记作

$$[0159] \quad D'_1(i, l_i) = 1/(mk), i=1 \dots m, l_i \in Y.$$

[0160] 步骤 4.2.7, 依次将主要特征矩阵 X' 的各列数据 (即所有图像的 F 组同维特征) 作为一个弱分类器的输入, 进行 F 轮运算, 按照下式计算 $r_{t,j}$ 的值:

$$[0161] \quad r_{t,j} = \sum_{j \in (i,l)} D'_t(i, l_i) K_i[l_i] h_j(x_i, l_i),$$

[0162] 从 F 个弱分类器中选择一个 $h_j(x, l)$, 使得 $r_t = \max(r_{t,j})$; 将该弱分类器记作 $h_t(x, l)$, 添加到强分类器中。其中 $D'_t(i, l)$ 表示第 t 轮迭代中第 i 个训练图像的权重值。

[0163] 步骤 4.2.8, 计算通过步骤 4.2.7 选择出的弱分类器 $h_j(x, l)$ 的权重 α_t ;

$$[0164] \quad \alpha_t = \frac{1}{2} \ln \left(\frac{1+r_t}{1-r_t} \right),$$

[0165] 步骤 4.2.9, 计算下一轮迭代中各个图像的权重 D'_{t+1} ;

$$[0166] \quad D'_{t+1} = \frac{D'_t(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i))}{Z_t}, i=1 \dots m.$$

[0167] 其中, Z_t 是归一化因子

$$[0168] \quad Z_t = \sum_{i,l} D'_t(i, l_i) \exp(-\alpha_t K_i[l_i] h_t(x_i, l_i)), i=1 \dots m.$$

[0169] 步骤 4.2.10, 将步骤 4.2.9 得到的新权重代入步骤 4.2.7, 按照步骤 4.2.7 至步骤 4.2.9 的方法, 得到一个新的弱分类器添加到强分类器中;

[0170] 按照上述步骤 4.2.7 至步骤 4.2.10 的方法进行 T 轮迭代, 得到由 T 个弱分类器组成的强分类器, 即第 k 个说话人的确认分类器, 表示为:

$$[0171] \quad W_k(x) = \underset{l}{\operatorname{argmax}} S_l, \quad S_l = (\sum_{t=1}^T \alpha_t h_t(x, l)) \quad (1)$$

[0172] 步骤 4.2.11, 待 k 轮训练结束后, 将所有的 k 个说话人分类器集合起来, 构成说话人分类器池 $W = \{W_1(x), W_2(x), \dots, W_k(x)\}$ 。

[0173] 步骤 5, 对用户提供的、申明是说话人 k 发声录制的语音文件, 提取其 2D-Haar 音频特征, 输入步骤 4 训练得到的说话人 k 的分类器, 以确认该文件中的语音是否确实由用户所申明的说话人讲出。具体步骤为:

[0174] 步骤 5.1, 对确认语音文件进行音频特征积分图提取, 得到待确认音频特征积分图序列 $G' = \{g'_{1}, g'_{2}, g'_{3}, \dots, g'_{u'}\}$ 。具体方法与步骤 2 所述方法相同。其中, 音频特征图序列转换过程中(对应于步骤 2.2), 帧长设定为 $f_s = 30\text{ms}$, 帧移设定为 $\Delta f_s = 20\text{ms}$; 音频特征图序列转换过程中(对应于步骤 2.2), 窗长 $a = 32$ 、步进 $s = 16$; 本实施例中, s_k 的总时长为 6.54s, 因此

[0175]

$$c = \left\lfloor \frac{t}{f_s - \Delta f_s} \right\rfloor = 654, \quad p = \sum_1^n p_n = 12 + 12 + 8 = 32.$$

[0176] 类似的, 待确认语音的总帧数 c' 的取值也由待确认语音文件的长度确定, 特征图序列包含的特征图数量 u' 为: $u' = \left\lfloor \frac{c'}{a-s} \right\rfloor = 40$ 。

[0177] 步骤 5.2, 在步骤 5.1 的基础上, 根据步骤 4.1 所述的 2D-Haar 音频特征提取方法, 以及步骤 4.2.5 的特征筛选结果, 为特征图序列中的每幅特征图提取 F 维 2D-Haar 音频特征, 构成由 510112 列, 40 行的 2D-Haar 音频特征矩阵 X' 。

[0178] 步骤 5.3, 从说话人分类器池中找到申明说话人 k 的说话人分类器 W_k , 再把步骤 5.2 得到的 2D-Haar 音频特征矩阵 X' 输入 W_k , 得到分类结果序列 R。

[0179] 所述分类结果序列 R 由 40 个元素组成, 其中每个元素的具体计算方法为:

[0180] 步骤 5.3.1, 按照步骤 4.2.10 中的(1)式, 读取说话人分类器中某个弱分类器 $h_t(x, 1)$ 及其相应 2D-Haar 音频特征 $f_j(x)$;

[0181] 步骤 5.3.2, 对于每种待选标签 $l_i \in \{k, \text{other}\}$, 分别计算该弱分类器的输出 $h_t(f_j(x), 1)$, 并将该输出值以分类器中的权重 α_t 累加到待选标签 l_i 对应的加权值 S_{li} 中;

[0182] 步骤 5.3.3, 按照步骤 5.3.1-步骤 5.3.2 的方法进行 T 轮循环之后, 每种待选标签 l_i 将得到一个加权值 S_{li} 。选出取值最大的一个加权值 S_{li} , 同时记录与其相对应的待选标签 l_i 作为该音频特征图的分类结果, 记作 $(l_i, S_{li, u'})$, 其中 l_k 为说话人标签, $S_{li, u'}$ 为相应的强分类器加权和。

[0183] 步骤 5.3.4, 将待确认音频的所有分类结果组合起来, 构成分类结果序列 $R = \{(l_i, S_{li, u'}) : (l_1, S_{11,1}), (l_1, S_{11,2}), (l_2, S_{12,3}), \dots, (l_i, S_{li, u'})\}$, $l_i \in \{k, \text{other}\}$ 。

[0184] 步骤 5.4, 对步骤 5.3 得到的分类结果序列进行结果综合, 得到最终的说话人确认结果。

[0185] 具体方法为:

[0186] 步骤 5.4.1, 统计结果序列中所有的强分类器判别权重 $S_{li, u'}$ 按说话人标签 l_i 加

权,即分别求出 $S_k = \sum_1^{u'} S_{k,u'}$ 和 $S_{other} = \sum_1^{u'} S_{other,u'}$

[0187] 步骤 5.4.2, 计算置信因子 $\eta = S_k / (S_{other} + S_k)$, 按照下式给出最终的说话人确认结果 V :

$$[0188] \quad V = \begin{cases} True, & \text{if } \eta > \omega \\ False, & \text{if } \eta \leq \omega \end{cases},$$

[0189] 式中 ω 为判别阈值, 可由用户指定。

[0190] 在本实施例中, $\eta = 75\%$, $\omega = 60\%$, 由于 $\eta > \omega$, 所以输出“True”, 表示待确认语音的确为用户所声明的说话人所讲出。

[0191] 第二环节

[0192] 本环节将对本发明的性能进行测试, 测试平台、说话人注册 / 训练流程说话人确认流程与实施例 1 相同, 以下将不再赘述, 重点说明性能测试的方法与结果。

[0193] 实验数据通过以下步骤生成: (1) 从所有说话人中随机选取 200 人的全部语音数据, (2) 从每个人的语音中选取 1 句作为训练数据, 3 句作为目标测试数据, (3) 针对每个目标说话人, 随机选取 3 句他人语句作为冒认测试数据, 并对每个说话人分别进行 1 真、1 真 1 假、2 真 1 假、2 真 2 假、3 真 2 假、3 真 3 假的 6 组测试, 记录每组测试下 200 人的识别结果。

[0194] 为了进行比较, 采用 GMM-UBM 方法、AdaBoost 方法进行对比, 记录三种方法的错误接受率 (False Acceptance Rate, FAR) 和错误拒绝率 (False Rejection Rate, FRR), 绘制 DET 曲线, 并统计准确率和确认耗时。其中:

[0195]

$$FAR = \frac{\text{错误接受的样本数}}{\text{应被拒绝的样本数}},$$

[0196]

$$FRR = \frac{\text{错误拒绝的样本数}}{\text{应被接受的样本数}},$$

[0197] 准确率 = 1 - 等错率。

[0198] 当测试规模从 200 次增加到 1200 次时, 三种方法的表现如图 6 所示。可见, 当测试次数不断增加时, 对比方法的确认准确率下降比较明显, 而本文所提方法下降趋势较缓, 在 1200 次的测试规模下, 较对比方法的准确率分别高出 3.2% 和 2.6%。

[0199] 为了评价本文所提算法的时间效率, 统计不同 2D-Haar 特征维数 T 下每秒钟语音数据的平均识别耗时 t 。由表 2 可知, 本文所提方法具有较高的识别速度。

[0200] 表 2 不同 T 值下本文所提方法的平均识别耗时

T 值	100	200	300	400	500
平均耗时(ms)	27.5	36.6	38.5	47.6	58.7

[0202] 由上述实验可知, 2D-Haar 音频特征在引入了时序信息的同时, 有效地扩充了特征空间的维度, 为训练出性能更优的分类器提供了可能; 同时, 使用 Turbo-Boost 算法, 结合单特征输入的 Decision Stump 弱分类器进行特征筛选, 既提高了特征向量的代表性和区分度, 也减少了确认阶段的计算负担, 确认速度较高。

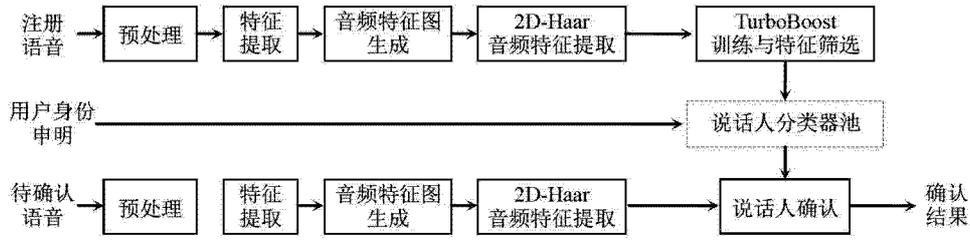


图 1

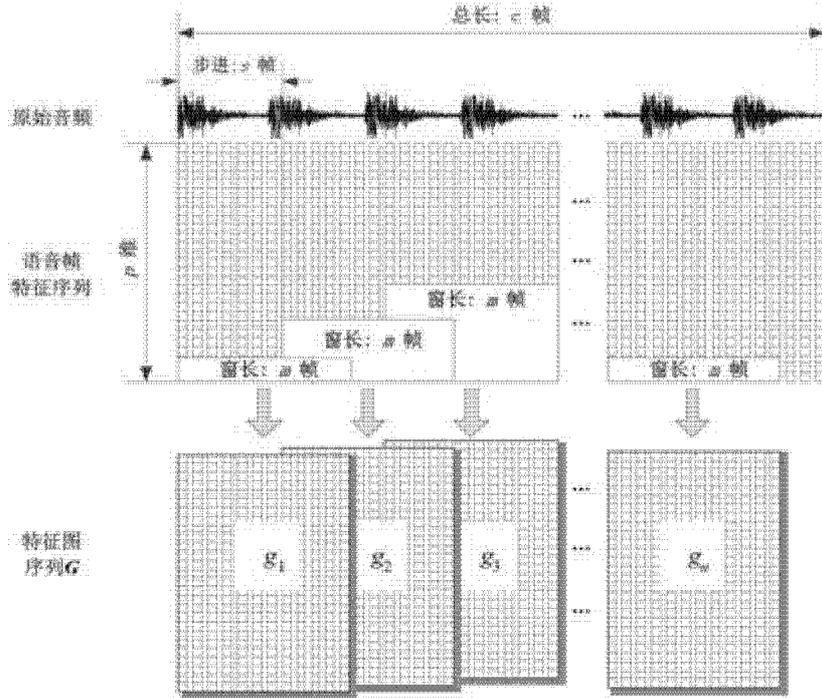


图 2

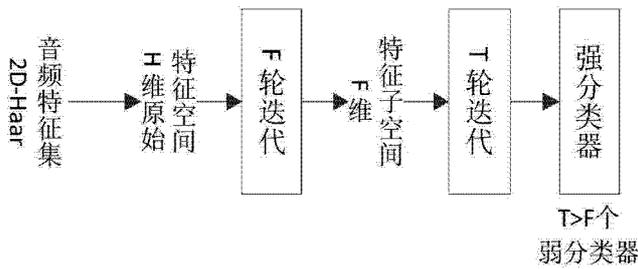


图 3

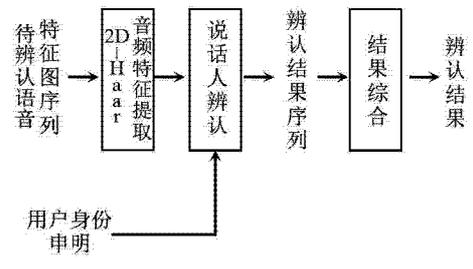


图 4

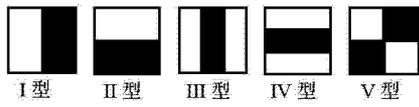


图 5

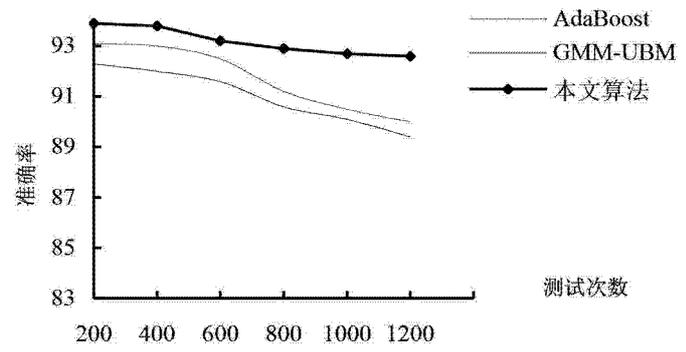


图 6