

一种基于语音样本特征空间轨迹的说话人识别方法

技术领域

本发明涉及生物特征识别领域，具体涉及一种基于语音样本特征空间轨迹的说话人识别方法。

背景技术

随着人工智能技术的发展，音频感知已经成为音频处理技术研究的热点，其中音频分类或音频识别是音频感知的核心问题，在工程应用中，音频分类表现为说话人识别、音频事件识别、音频事件检测等。说话人识别技术是身份验证技术——生物特征识别技术的一种。生物特征识别技术是利用生物特征自动识别个体身份的技术，包括指纹识别、虹膜识别、基因识别、人脸识别等。与其他身份验证技术相比，说话人识别更加方便、自然，且具有比较低的用户侵犯性。说话人识别利用语音信号进行身份识别，具有人机交互自然、语音信号易于提取、可实现远程识别等优势。

现有的说话人识别系统包括两个阶段：训练阶段和识别阶段。在训练阶段，系统使用收集的说话人语音为说话人建立模型；在识别阶段，系统将输入语音与说话人模型进行匹配来作出判决。说话人识别系统需要从语音信号提取能反映说话人个性的特征，并建立准确的模型区分该说话人与其他说话人之间的差异。目前常用的音频分类技术主要有两大类，一类是生成统计模型，如混合高斯模型 GMM 和隐马尔可夫模型 HMM，另一类是基于神经网络的方法，如 DNN、RNN 或 LSTM 等。不论是哪一种技术，都需要大量的标注训练样本，为了达到较好的识别性能，神经网络方法对样本规模要求更高。基于 GMM 或 HMM 的方法对不同音频类之间的区分性信息并没有加以特别的考虑，也没有考虑不同类样本数据的共享，比如：MIT Reynold 教授等的论文《Speaker Verification Using Adapted Gaussian Mixture Models》(*Digital Signal Processing* 10 (2000), 19–41.) 提到的方法，具有较高的计算复杂度；在大样本支持下，神经网络方法表现出了很好的性能，比如谷歌公司的论文《End-to-End Text-Dependent Speaker Verification》(2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2016, Pages: 5115–5119) 中使用神经网络对语音提取特征并训练，但是神经网络的训练需要大量的有标注语音，而大量样本的获取成本是非常高昂的，而且神经网络方法缺乏解释性，相当一个黑匣子。

现有说话人识别技术往往计算复杂度较高，需要大量标注过的说话人语音数据来训练模型，而采集大量的有标注语音数据需要巨大的工作量。因此需要找到一种能够

更加方便有效的说话人识别方法和系统。

发明内容

本发明的目的是针对现有技术的不足，提供了一种基于语音样本特征空间轨迹的说话人识别方法，其中语音特征空间是不依赖说话人、文本和语言的，因此语音特征空间的构建可以采用任何合格的语音数据，实现了语音数据的共享；而说话人语音轨迹，即使一个样本也能构建，因此不需要大量的有标注语音数据，克服了现有技术需要采集大量有标注语音数据的缺陷。

本发明的目的可以通过如下技术方案实现：

一种基于语音样本特征空间轨迹的说话人识别方法，其中一个语音样本能够视为语音特征空间的一次运动，具有活动空间和空间中的轨迹特性，所述方法包括以下步骤：

步骤 1)、构建语音特征空间 Ω ：利用聚类方法将无标注语音样本在特征空间进行聚类，由聚类所得到的子类数据生成该子类数据的某种表达作为语音特征空间的表达 $\Omega = \{g_k, k=1, 2, \dots, K\}$ ；

步骤 2)、构建说话人知识：利用有说话人属性标注的纯净语音样本，获得其在语音特征空间 Ω 上的分布信息以及运动轨迹信息；

步骤 3)、说话人识别：对于待识别语音样本，首先获得该样本的语音特征空间分布表达以及轨迹，然后利用说话人语音特征空间分布信息计算样本分布与先验分布的差异以及沿轨迹的累计局部分布差异，作为说话人识别的依据并进行判断。

进一步地，步骤 1) 构建语音特征空间 Ω 的过程中，能够使用任何纯净语音样本，对说话人、语种因素没有任何约束。

进一步地，步骤 1) 中采用 K-means 或其它聚类方法将语音样本在特征空间进行聚类，所述语音特征空间表达 $\Omega = \{g_k, k=1, 2, \dots, K\}$ 能够是类数据的分布函数（比如高斯分布函数）、聚类中心矢量（质心）或者生成模型（比如隐马尔可夫模型或神经网络）这些具有定位能力的标识，称之为特征空间标识子，语音特征空间所使用的类标识子规模 K 决定语音特征空间表达粒度， K 越大，语音特征空间表达越精细。另一方面，空间表达的准确性与数据规模有关，数据越丰富，空间表达越完整；同时，构建语音特征空间的数据越有针对性，对于特定问题而言，空间表达会越精确。

进一步地，步骤 2) 中，利用有说话人属性标注的纯净语音样本对语音特征空间进行标注，在采用高斯分布 $g_k(m_k, U_k)$ 作为空间标识子时，说话人特征空间分布信息按以下方式获得：

一、计算语音样本每个特征 f_t 与空间标识子 $g_k(m_k, U_k)$ 的位置关联度, 定义为:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

式中, 空间标识子用多维高斯分布来表示, m_k 表示第 k 个高斯分布的均值向量, U_k 表示第 k 个多维高斯分布的方差矩阵;

二、计算说话人样本集与空间标识子 $g_k(m_k, U_k)$ 的位置关联度的期望值:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

式中, $p_{n,t}^k$ 表示第 n 个样本的第 t 帧特征与空间标识子 $g_k(m_k, U_k)$ 的关联度;

三、计算说话人特征空间分布为:

$$p_k = \frac{B^k}{\sum_k B^k}.$$

进一步地, 步骤 2) 中, 说话人语音样本在语音特征空间 Ω 上的运动轨迹时序信息表示为语音样本特征的邻域序列 $\Psi_1 \Psi_2 \cdots \Psi_T$, 而语音样本特征 f_t 的 δ 邻域为 $\Psi_t = \{g_k | d_{tk} < \delta\}$, 其中 d_{tk} 为语音样本特征与语音样本分布的马哈拉诺比斯距离 (Mahalanobis distance), 即 $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$ 。

进一步地, 所述语音样本特征 f_t 的 δ 邻域 $\Psi_t = \{g_k | d_{tk} < \delta\}$ 的判决门限参考正态分布的特性, 选取 $2 < \delta < 3$ 。

进一步地, 步骤 3) 中, 语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 的说话人识别过程包括以下步骤:

一、计算语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 在语音特征空间 Ω 的分布 $P = (p_1, p_2, \cdots, p_K)$,

$$\text{其中 } p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k};$$

二、确定语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 在语音特征空间 Ω 中的运动轨迹 $\Psi_1 \Psi_2 \cdots \Psi_T$,

$$\Psi_t = \{g_k | d_{tk} < \delta\};$$

三、计算样本分布 $P=(p_1, p_2, \dots, p_K)$ 与说话人 s 的先验特征空间分布 $P_s=(p_1^s, p_2^s, \dots, p_K^s)$ 的距离 $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$ ，然后筛选包含真实解的可能的解集 $S_p: \max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$;

四、计算运动轨迹 $\Psi_1 \Psi_2 \dots \Psi_T$ 的距离度量 $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ ，从 S_p 中选出可能的解 $s^* = \arg \min_{s \in S_p} (D_s)$ ，完成说话人识别。

具体地，步骤3)中，仅用语音样本的空间分布信息 $P=(p_1, p_2, \dots, p_K)$ 或者运动轨迹距离 $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ 均能够获得很好的说话人识别性能。

本发明与现有技术相比，具有如下优点和有益效果：

1、本发明提供一种基于语音样本特征空间轨迹的说话人识别方法，其中语音特征空间的建立是对大量的语音特征进行聚类，不需要有标注的数据，建立语音特征空间的数据样本来源于不同说话人即可，对于说话内容、说话人年龄、语种没有确切的要求，克服了神经网络方法中，需要大量有标注语音的问题，语音空间建立的数据采集方便实现。

2、本发明提供一种基于语音样本特征空间轨迹的说话人识别方法，基于说话人语音特征在语音特征空间中的定位与轨迹信息，不同于信号源生成类模型方法，如隐马尔可夫模型（HMM）等，定位是相对的，而生成模型是绝对的；与深度神经网络的方法相比，具有可解释性，每一个知识数据都具有一定的物理语义，比如样本特征在空间 Ω 上的关联度分布信息 $P=(p_1, p_2, \dots, p_K)$ 即表达了该样本的活动空间范围（非零元素所对应的标识子集合所代表的空间），又表达了在该空间的分布。

3、本发明提供一种基于语音样本特征空间轨迹的说话人识别方法，本质是语音特征在空间中定位，对于不同的说话人的语音特征，在建立的语音特征空间上进行定位，使用关联度来表示不同说话人的语音特征定位信息，通过较少的计算量就表达出了不同说话人之间的区分性，相比于需要用生成模型对每个说话人进行建模的 GMM 或 HMM 的方法，具有更低的计算复杂度。

4、本发明提供一种基于语音样本特征空间轨迹的说话人识别方法，其中语音特征空间标识子集合是用来定位说话人语音特征的参照系统，是相对关系，与待识别样本没有严格的关系要求，因此特征空间具有共享性，建立的语音特征空间可以迁移到其他的说话人数据集上进行识别，比如：一个语种的说话人语音特征空间可作为另一语种的说话人识别的语音特征空间，实现了数据的共享。

附图说明

图 1 为本发明实施例 1 中说话人识别方法的概略流程图。

图 2 为本发明实施例 1 中语音特征空间建立的步骤流程图。

图 3 为本发明实施例 1 中说话人语音特征空间分布信息与轨迹信息生成的步骤流程图。

图 4 为本发明实施例 1 中对待识别语音样本进行识别的步骤流程图。

具体实施方式

下面结合实施例及附图对本发明作进一步详细的描述，但本发明的实施方式不限于此。

实施例 1:

本实施例提供了一种基于语音样本特征空间轨迹的说话人识别方法，概略流程图如图 1 所示，包括以下三个步骤：

1) 建立语音特征空间 Ω ，可以使用任何纯净语音样本，对说话人、语种等因素没有任何约束，然后利用聚类方法将语音样本在特征空间进行聚类，聚类所得到的子类数据表达为语音特征空间的表达 $\{g_k, k=1, 2, \dots, K\}$ ；

2) 构建说话人知识，包括说话人在语音特征空间中的分布信息和运动轨迹信息两部分；

3) 对于待识别语音样本，利用说话人语音特征空间分布信息以及语音样本的运动轨迹信息进行识别。

参考图 2，是本实施例中语音特征空间建立的步骤流程图。使用 aishell 中文语料库中的说话人语音数据作为无标注的语音样本集，aishell 中一共包含 400 个说话人，选择每个人的 60 个 wav 文件用来训练语音特征空间，提取无标注语音样本集 $X = \{x_1, x_2, \dots, x_N\}$ 的 12 维 MFCC 特征，从而获得特征集 $F_x = \{f_i^x, i=1, 2, \dots, t_N\}$ ，其中

f_i^x 是短时帧特征， t_N 为所有样本的帧数之和；

然后使用 $F_x = \{f_i^x, i=1, 2, \dots, t_N\}$ 特征序列训练一个混合度为 K 的 GMM，舍弃 GMM 的权重信息，保留每一个高斯分量作为语音特征空间的标识子集合。其中，K 为音频特征空间标识子的数量，标识数量 K 选择 4096，以便对音频特征空间给予精度较高的描述；

语音特征空间标识子表示为 $\Omega = \{g_k, k=1, 2, \dots, K\}$ ，其中， $g = N(m, U)$ 为多维高斯分布函数；

参考图 3, 是本实施例中说话人特征空间分布信息生成的步骤流程图。对于 aishell 的中每个人, 使用 20 个 wav 文件用来对语音特征空间进行标注。目标说话人语音样本集 $Y = \{(y_1, s_1), (y_2, s_2), \dots, (y_M, s_M)\}$, $s_i \in S = \{S_l, l=1, 2, \dots, L\}$ (说话人集), 说话人 S_l 的样本为 $Y_l = \{y_m | s_m = S_l, m=1, 2, \dots, M\}$, 提取其音频特征序列为 $F_l = \{f_1, f_2, \dots, f_{l_i}\}$; 计算语音样本所有特征 f_t 与空间标识子 $g_k(m_k, U_k)$ 的位置关联度:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

计算说话人样本集与空间标识子 $g_k(m_k, U_k)$ 的位置关联度的期望值:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

其中 $p_{n,t}^k$ 为第 n 个样本的第 t 帧特征与标识子 $g_k(m_k, U_k)$ 的位置关联度;

计算说话人特征空间分布为:

$$p_k = \frac{B^k}{\sum_k B^k}$$

对目标说话人集中每个说话人的注册语音进行处理, 得到每个说话人的语音特征分布信息。

语音样本的运动轨迹时序信息表示为语音样本特征的邻域序列 $\Psi_1 \Psi_2 \dots \Psi_T$, 而样本特征 f_t 的 δ 邻域为 $\Psi_t = \{g_k | d_{tk} < \delta\}$, 其中 d_{tk} 为特征与分布的 Mahalanobis distance,

即 $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$;

参考图 4, 是本实施例中对语音样本进行识别的步骤流程图。待识别语音样本的特征为 $f = \{f_1, f_2, \dots, f_T\}$, 识别过程如下:

计算语音 $f = \{f_1, f_2, \dots, f_T\}$ 在特征空间 Ω 的分布 $P = (p_1, p_2, \dots, p_K)$;

其中, 特征 f_t 与空间标识子 $g_k(m_k, U_k)$ 的位置关联度为:

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

待识别说话人样本与空间标识子 $g_k(m_k, U_k)$ 的位置关联度为:

$$p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k}$$

确定语音特征 $f = \{f_1, f_2, \dots, f_T\}$ 在特征空间 Ω 中的轨迹 $\Psi_1 \Psi_2 \dots \Psi_T$ ，其中 $\Psi_t = \{g_k | d_{tk} \leq \delta\}$ ；

计算样本分布 $P = (p_1, p_2, \dots, p_K)$ 与说话人 s 的先验特征空间分布 $P_s = (p_1^s, p_2^s, \dots, p_K^s)$ 的距离 $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$ ，其中 α 取 2，然后筛选包含真实解的可能的解集 S_p ： $\max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$ ，选择距离最小的 10 个说话人作为候选的识别结果；

计算轨迹 $\Psi_1 \Psi_2 \dots \Psi_T$ 的距离度量 $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ ，其中 α 取 2，从候选的 10 个说话人中选择轨迹距离最小的说话人作为识别结果，即 $s^* = \arg \min_{s \in S_p} (D_s)$ 。

实施例 2:

本实施例提供了一种基于语音样本特征空间轨迹的说话人识别方法，包括以下步骤：

步骤 1，使用英文语料库 timit 的语音数据来建立语音特征空间标识子集合；

步骤 2，使用 aishell 语料库中的语音数据对目标说话人集进行注册，同实施例 1；

步骤 3，对待识别语音样本进行识别，同实施例 1。

得到的识别效果与实施例 1 中相比有少量差距，可以证明另外一个语种的说话人语音特征空间可作为另一语种的说话人识别的语音特征空间，实现了数据的共享。

以上所述，仅为本发明专利较佳的实施例，但本发明专利的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明专利所公开的范围内，根据本发明专利的技术方案及其发明专利构思加以等同替换或改变，都属于本发明专利的保护范围。

权 利 要 求 书

1、一种基于语音样本特征空间轨迹的说话人识别方法，其中一个语音样本能够视为语音特征空间的一次运动，具有活动空间和空间中的轨迹特性，其特征在于，所述方法包括以下步骤：

步骤 1)、构建语音特征空间 Ω ：利用聚类方法将无标注语音样本在特征空间进行聚类，由聚类所得到的子类数据生成该子类数据的某种表达作为语音特征空间的表达 $\Omega = \{g_k, k=1, 2, \dots, K\}$ ；

步骤 2)、构建说话人知识：利用有说话人属性标注的纯净语音样本，获得其在语音特征空间 Ω 上的分布信息以及运动轨迹信息；

步骤 3)、说话人识别：对于待识别语音样本，首先获得该样本的语音特征空间分布表达以及轨迹，然后利用说话人语音特征空间分布信息计算样本分布与先验分布的差异以及沿轨迹的累计局部分布差异，作为说话人识别的依据并进行判断。

2、根据权利要求 1 所述的一种基于语音样本特征空间轨迹的说话人识别方法，其特征在于：步骤 1) 构建语音特征空间 Ω 的过程中，能够使用任何纯净语音样本，对说话人、语种因素没有任何约束。

3、根据权利要求 1 所述的一种基于语音样本特征空间轨迹的说话人识别方法，其特征在于：所述语音特征空间表达 $\Omega = \{g_k, k=1, 2, \dots, K\}$ 能够是类数据的分布函数、聚类中心矢量或者生成模型这些具有定位能力的标识，称之为特征空间标识子，语音特征空间所使用的类标识子规模 K 决定语音特征空间表达粒度， K 越大，语音特征空间表达越精细。

4、根据权利要求 1 所述的一种基于语音样本特征空间轨迹的说话人识别方法，其特征在于：步骤 2) 中，利用有说话人属性标注的纯净语音样本对语音特征空间进行标注，在采用高斯分布 $g_k(m_k, U_k)$ 作为空间标识子时，说话人特征空间分布信息按以下方式获得：

一、计算语音样本每个特征 f_t 与空间标识子 $g_k(m_k, U_k)$ 的位置关联度，定义为：

$$p_t^k = \frac{1}{\left[(f_t - m_k)^T U_k^{-1} (f_t - m_k) \right]^\alpha}$$

式中，空间标识子用多维高斯分布来表示， m_k 表示第 k 个高斯分布的均值矢

量, U_k 表示第 k 个多维高斯分布的方差矩阵;

二、计算说话人样本集与空间标识子 $g_k(m_k, U_k)$ 的位置关联度的期望值:

$$B^k = \sum_n \sum_t p_{n,t}^k$$

式中, $p_{n,t}^k$ 表示第 n 个样本的第 t 帧特征与空间标识子 $g_k(m_k, U_k)$ 的关联度;

三、计算说话人特征空间分布为:

$$p_k = \frac{B^k}{\sum_k B^k}.$$

5、根据权利要求 4 所述的一种基于语音样本特征空间轨迹的说话人识别方法, 其特征在于: 步骤 2) 中, 说话人语音样本在语音特征空间 Ω 上的运动轨迹时序信息表示为语音样本特征的邻域序列 $\Psi_1 \Psi_2 \cdots \Psi_T$, 而语音样本特征 f_t 的 δ 邻域为

$\Psi_t = \{g_k | d_{tk} \leq \delta\}$, 其中 d_{tk} 为语音样本特征与语音样本分布的马哈拉诺比斯距离 (Mahalanobis distance), 即 $d_{tk} = \sqrt{(f_t - m_k)^T U_k^{-1} (f_t - m_k)}$ 。

6、根据权利要求 5 所述的一种基于语音样本特征空间轨迹的说话人识别方法, 其特征在于: 所述语音样本特征 f_t 的 δ 邻域 $\Psi_t = \{g_k | d_{tk} < \delta\}$ 的判决门限参考正态分布的特性, 选取 $2 < \delta < 3$ 。

7、根据权利要求 5 所述的一种基于语音样本特征空间轨迹的说话人识别方法, 其特征在于: 步骤 3) 中, 语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 的说话人识别过程包括以下步骤:

一、计算语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 在语音特征空间 Ω 的分布 $P = (p_1, p_2, \cdots, p_K)$,

$$\text{其中 } p_k = \frac{\sum_t p_t^k}{\sum_k \sum_t p_t^k};$$

二、确定语音样本 $f = \{f_1, f_2, \cdots, f_T\}$ 在语音特征空间 Ω 中的运动轨迹 $\Psi_1 \Psi_2 \cdots \Psi_T$,

$$\Psi_t = \{g_k | d_{tk} < \delta\};$$

三、计算样本分布 $P = (p_1, p_2, \cdots, p_K)$ 与说话人 s 的先验特征空间分布 $P_s = (p_1^s, p_2^s, \cdots, p_K^s)$ 的距离 $d_s = \|P - P_s\|^\beta = \sum_k (p_k - p_k^s)^\beta$, 然后筛选包含真实解的可能的解集 S_p : $\max_{l \in S_p} \{d_{sl}\} < \min_{l \notin S_p} \{d_{sl}\}$;

四、计算运动轨迹 $\Psi_1 \Psi_2 \cdots \Psi_T$ 的距离度量 $D_s = \sum_t \sum_{k: g_k \in \Psi_t} (p_k - p_k^s)^\beta$ ，从 S_p 中选出可能的解 $s^* = \arg \min_{s \in S_p} (D_s)$ ，完成说话人识别。