

一种声纹识别方法、装置、终端设备及存储介质

本申请以 2018 年 5 月 29 日提交的申请号为 201810528518.0，名称为“一种声纹识别方法、装置、终端设备及存储介质”的中国发明专利申请为基础，并要求其优先权。

技术领域

本申请涉及计算机通讯技术领域，尤其涉及一种声纹识别方法、装置、终端设备及存储介质。

背景技术

呼叫中心由交互式语音应答系统和人工话务系统构成。人工话务系统由签入系统，话务平台，接口机组成，客户代表也就是坐席员为了进行客户服务，需在签入系统进行签入操作，成功签入话务平台后，根据话务平台分配的人工服务请求和客户建立通话，也就是坐席员外呼，来进行客户服务，通常会根据业务需求将客户分为不同业务的不同等级，每个坐席员按照其职责来安排接待不同业务不同等级的客户。

虽然每个坐席员在外呼前已被告知相应的外呼权限，但现实生活中经常存在一些业务不熟练的坐席员对高等级客户进行外呼的情况。

现有技术一般是通过事后听取录音并对录音进行分析，从而获取不符合权限的外呼信息并进行相应处理，做不到及时的预警，导致客服语音呼叫的监控不及时，监控效率低。

发明内容

本申请实施例提供一种声纹识别方法、装置、终端设备及存储介质，以解决客服语音呼叫的监控不及时，监控效率低的问题。

第一方面，本申请实施例提供一种声纹识别方法，包括：

获取数据库中每个坐席员的呼叫权限，并根据预设的权限分组方式，将每个所述坐席员分配到所述呼叫权限对应的分组中；

针对每个所述分组，对该分组中的每个所述坐席员进行声纹特征采集，得到该分组的基础声纹特征集；

对每个所述基础声纹特征集中的声纹特征进行训练，得到每个所述基础声纹特征集的综合声纹特征，并基于所述分组、所述呼叫权限和所述综合声纹特征之间的对应关系，构建动态声纹识别模型；

若检测到待识别坐席员的呼叫操作，则获取被呼叫客户的客户级别和所述待识别坐席员的语音数据，并对所述语音数据进行声纹特征提取，得到目标声纹特征；

基于所述动态声纹识别模型，对所述目标声纹特征进行识别，确定所述待识别坐席员的呼叫权限；

若所述待识别坐席员的呼叫权限不满足所述客户级别的要求，则执行预设的预警处理措施。

第二方面，本申请实施例提供一种声纹识别装置，包括：

坐席权限分组模块，用于获取数据库中每个坐席员的呼叫权限，并根据预设的权限分组方式，将每个所述坐席员分配到所述呼叫权限对应的分组中；

声纹特征采集模块，用于针对每个所述分组，对该分组中的每个所述坐席员进行声纹特征采集，得到该分组的基础声纹特征集；

识别模型构建模块，用于对每个所述基础声纹特征集中的声纹特征进行训练，得到每

个所述基础声纹特征集的综合声纹特征，并基于所述分组、所述呼叫权限和所述综合声纹特征之间的对应关系，构建动态声纹识别模型；

目标特征获取模块，用于若检测到待识别坐席员的呼叫操作，则获取被呼叫客户的客户级别和所述待识别坐席员的语音数据，并对所述语音数据进行声纹特征提取，得到目标声纹特征；

呼叫权限确定模块，用于基于所述动态声纹识别模型，对所述目标声纹特征进行识别，确定所述待识别坐席员的呼叫权限；

实时语音预警模块，用于若所述待识别坐席员的呼叫权限不满足所述客户级别的要求，则执行预设的预警处理措施。

第三方面，本申请实施例提供一种终端设备，包括存储器、处理器以及存储在所述存储器中并可在所述处理器上运行的计算机可读指令，所述处理器执行所述计算机可读指令时实现所述声纹识别方法的步骤。

第四方面，本申请实施例提供一个或多个存储有计算机可读指令的非易失性可读存储介质，所述计算机可读指令被一个或多个处理器执行时，使得所述一个或多个处理器执行所述声纹识别方法的步骤。

本申请的一个或多个实施例的细节在下面的附图和描述中提出，本申请的其他特征和优点将从说明书、附图以及权利要求变得明显。

附图说明

为了更清楚地说明本申请实施例的技术方案，下面将对本申请实施例的描述中所需要使用到的附图作简单地介绍，显而易见地，下面描述中的附图仅仅是本申请的一些实施例，对于本领域普通技术人员来讲，在不付出创造性劳动性的前提下，还可以根据这些附图获得其他的附图。

图1是本申请实施例1提供的声纹识别方法的实现流程图；

图2是本申请实施例1提供的声纹识别方法中步骤S2的实现流程图；

图3是本申请实施例1提供的声纹识别方法中语音信号交叠分帧的示例图；

图4是本申请实施例1提供的声纹识别方法中步骤S25的实现流程图；

图5是本申请实施例1提供的声纹识别方法中步骤S3的实现流程图；

图6是本申请实施例1提供的声纹识别方法中步骤S5的实现流程图；

图7是本申请实施例2中提供的声纹识别装置的示意图；

图8是本申请实施例4中提供的终端设备的示意图。

具体实施方式

下面将结合本申请实施例中的附图，对本申请实施例中的技术方案进行清楚、完整地描述，显然，所描述的实施例是本申请一部分实施例，而不是全部的实施例。基于本申请中的实施例，本领域普通技术人员在没有作出创造性劳动前提下所获得的所有其他实施例，都属于本申请保护的范围。

实施例1

请参阅图1，图1示出本实施例提供的声纹识别方法的实现流程。该声纹识别方法应用在呼叫中心的坐席员外呼场景中，该呼叫中心包括服务端和客户端，其中，服务端和客户端之间通过网络进行连接，坐席员通过客户端进行外呼呼叫，客户端具体可以但不限于各种个人计算机、笔记本电脑、智能手机、平板电脑和便携式可穿戴设备，服务端具体可以用独立的服务器或者多个服务器组成的服务器集群实现。本申请实施例提供的声纹识别方法应用于服务端，详述如下：

S1：获取数据库中每个坐席员的呼叫权限，并根据预设的权限分组方式，将每个坐席

员分配到呼叫权限对应的分组中。

具体地，在数据库中存储有每个坐席员外呼的呼叫权限，按照呼叫权限的不同，将每个坐席员分配到其呼叫权限对应的分组之中。

例如，在一具体实施方式中，数据库存储的呼叫权限有 6 个等级分别为 R_1 、 R_2 、 R_3 、 R_4 、 R_5 和 R_6 ，坐席员一共有 3000 人，将这些坐席员按照其对应权限分配到 6 个不同权限等级的分组中，得到权限按高到底的如下分组情况：分组一共 10 人，对应的权限为 R_1 ；分组二共 80 人，对应的权限为 R_2 ；分组三共 200 人，对应的权限为 R_3 ；分组四共 600 人，对应的权限为 R_4 ；分组五共 1800 人，对应的权限为 R_5 ；分组六共 310 人，对应的权限为 R_6 。

值得说明的是，高等级的权限包含低等级的权限，例如，属于分组一的坐席员具有所有呼叫权限，属于分组六的坐席员具有的权限最少，即 $R_6 \subset R_5 \subset R_4 \subset R_3 \subset R_2 \subset R_1$ 。

S2：针对每个分组，对该分组中的每个坐席员进行声纹特征采集，得到该分组的基础声纹特征集。

具体地，分别对同一分组的坐席员进行声纹特征采集，将同一分组的所有坐席员的声纹特征作为该分组的基础声纹特征集。

其中，声纹特征包括但不限于：声学特征、词法特征、韵律特征、语种方言口音信息和通道信息等。

优选地，本申请采用的声纹特征为声学特征。

S3：对每个基础声纹特征集中的声纹特征进行训练，得到每个基础声纹特征集的综合声纹特征，并基于分组、呼叫权限和综合声纹特征之间的对应关系，构建动态声纹识别模型。

具体地，对每个基础声纹特征集中的声纹特征进行训练，得到每个基础声纹特征集的综合声纹特征，即该分组的综合声纹特征，并根据步骤 S1 和步骤 S2 中提供的分组和呼叫权限的对应关系、分组与声纹特征的对应关系，来建立动态声纹识别模型。

其中，对每个基础声纹特征集中的声纹特征进行训练，得到每个基础声纹特征集的综合声纹特征，所采用的训练方式包括但不限于：模板匹配方法、最近邻方法、神经网络方法、隐式马尔可夫模型（Hidden Markov Model, HMM）、矢量量化（Vector Quantization, VQ）方法、多项式分类器（Polynomial Classifiers）方法等。

优选地，本申请实施例采用的训练方式为基于神经网络方法的遗传算法（Genetic Algorithm）的改进算法，通过将每个分组作为一个群体，分组中的每个坐席员的声纹特征作为一条染色体，通过使用遗传算法来进行初始化编码、个体评价、选择、交叉和变异，进而得到该分组的最优解或近似解，即该分组的综合声纹特征。

S4：若检测到待识别坐席员的呼叫操作，则获取被呼叫客户的客户级别和待识别坐席员的语音数据，并对语音数据进行声纹特征提取，得到目标声纹特征。

具体地，在检测到有待识别坐席员的呼叫操作时，根据被呼叫的客户的客户标识信息，再根据客户数据库里面查找该标识信息，获取被呼叫的客户的客户级别。

其中，客户级别为预先设置并存储在客户数据库中，客户级别与权限级别存在映射关系。

以步骤 S1 中存储的 6 个权限级别为例，客户数据库中的客户级别分为 1 至 10 级，依照权限分组的 6 个分组对应的权限分别为：分组一的坐席员可呼叫客户级别为 1 级至 10 级，分组二的坐席员可呼叫客户级别为 3 级至 10 级，分组三的坐席员可呼叫客户级别为 4 级至 10 级，分组四的坐席员可呼叫客户级别为 5 级至 10 级，分组五的坐席员可呼叫客户级别为 7 级至 10 级，分组六的坐席员可呼叫客户级别为 9 级至 10 级。

进一步地, 获取待识别坐席员的语音数据, 并对该语音数据进行声纹特征提取, 得到该坐席员的目标声纹特征, 提取方法请参照下文中步骤 S22 至步骤 S25 的描述, 为避免重复, 此处不再赘述。

S5: 基于动态声纹识别模型, 对目标声纹特征进行识别, 确定待识别坐席员的呼叫权限。

具体地, 通过将目标声纹特征输入到动态声纹识别模型中进行识别, 来判断该目标声纹特征对应的分组, 进而找到该分组对应的呼叫权限, 即为待识别坐席的呼叫权限。

S6: 若待识别坐席员的呼叫权限不满足被呼叫客户的客户级别的要求, 则执行预设的预警处理措施。

具体地, 根据待识别坐席员的呼叫权限对应的客户级别 R 与被呼叫客户的客户级别 r 之间的关系, 来确定当前的呼叫是否符合规范, 若 $r \notin R$, 则认为待识别坐席员的呼叫权限不满足被呼叫客户的客户级别的要求, 将执行预设的处理措施, 并发送本次呼叫的信息给监控人员, 记录此次违规的呼叫记录。

以步骤 S4 中的 6 个权限分组对应的 10 个客户级别为例, 获取到的待识别坐席员所属分组为分组五, 对应的呼叫权限为 R_5 , 即可外呼的客户级别为 7 级至 10 级, 若当前的被呼叫的客户客户级别为 7 级至 10 级中的任一等级, 则当前外呼是满足要求的, 若当前的被呼叫客户的客户级别为 1 级至 6 级中的任一等级, 则认为待识别坐席员的呼叫权限不满足被呼叫客户的客户级别的要求, 将采取相应预警措施来阻止该呼叫继续进行。

其中, 预设的处理措施包括但不限于: 向待识别坐席员发送提醒信息、向监控中心推送警告信息、强制呼叫转移和强制挂断等, 具体依据实际情况而定, 在此不作限制。

在图 1 对应的实施例, 一方面, 通过获取数据库中每个坐席员的呼叫权限, 并根据预设的权限分组方式, 将每个坐席员分配到呼叫权限对应的分组中, 针对每个分组, 对该分组中的每个坐席员进行声纹特征采集, 得到该分组的基础声纹特征集, 并对每个基础声纹特征集中的声纹特征进行训练, 得到每个基础声纹特征集的综合声纹特征, 并基于分组、呼叫权限和综合声纹特征之间的对应关系, 构建动态声纹识别模型, 这种通过分组的方式, 提取每个分组的综合声纹特征, 能够在检测到呼叫操作时, 快速判断当前语音数据所属分组, 提升识别效率, 进而实现及时监控, 提升监控效率; 另一方面, 当检测到待识别坐席员的呼叫操作时, 获取被呼叫客户的客户级别和待识别坐席员的语音数据, 并对该语音数据进行声纹特征提取, 得到目标声纹特征, 同时基于动态声纹识别模型, 对该目标声纹特征进行识别, 确定待识别坐席员的呼叫权限, 若待识别坐席员的呼叫权限不满足该客户级别的要求, 则执行预设的预警处理措施, 通过采用声纹识别的方式, 判断坐席员的呼叫权限, 进而确定坐席员当前的呼叫是否合法, 并在非法时能够及时采取预警措施, 实现实时监控, 有效提高监控的实时和坐席呼叫的呼叫质量。

接下来, 在图 1 对应的实施例的基础之上, 下面通过一个具体的实施例来对步骤 S2 中所提及的针对每个分组, 对该分组中的每个坐席员进行声纹特征采集, 得到该分组的基础声纹特征集的具体实现方法进行详细说明。

请参阅图 2, 图 2 示出了本申请实施例提供的步骤 S2 的具体实现流程, 详述如下:

S21: 针对每个分组中的每个坐席员, 获取该坐席员对预设的多个不同语料进行朗读的录音样本, 并对该录音样本进行声纹解析, 得到该坐席员的初始声纹样本。

具体地, 针对每个坐席员, 提供预设的不同内容的语料, 在坐席员朗读该语料时进行录音, 得到录音样本, 对该录音文本进行语音信号提取, 进而获取该坐席员的初始声纹样本。

例如, 在一具体实施方式中, 提供给某一坐席员的预料有三个, 分别是语料一、语料二和语料三, 该坐席员对这三个语料每个重复读 5 遍, 进而获取到了 15 个录音文本, 对这 15 个录音文本进行语音信号提取, 得到 15 段语音信号, 将这 15 段语音信号作为该

坐席员的初始声纹样本。

其中，语料是指在统计自然语言处理中实际上不可能观测到大规模的语言实例。

优选地，每个录音样本的录音时长为 30 秒钟。

S22：对初始声纹样本进行预加重处理，生成具有平坦频谱的加重处理声纹样本。

具体地，由于声门激励和口鼻辐射会对语音信号的平均功率谱产生影响，导致高频在超过 800Hz 时会按 6dB/倍频跌落，所以在计算语音信号频谱时，频率越高相应的成分越小，为此要在预处理中进行预加重（Pre-emphasis）处理，预加重的目的是提高高频部分，使信号的频谱变得平坦，保持在低频到高频的整个频带中，能用同样的信噪比求频谱，以便于频谱分析或者声道参数分析。预加重可在语音信号数字化时在反混叠滤波器之前进行，这样不仅可以进行预加重，而且可以压缩信号的动态范围，有效地提高信噪比。预加重可使用一阶的数字滤波器来实现，例如：有限脉冲响应（Finite Impulse Response, FIR）滤波器。

值得说明的是，利用设备获取的语音信号都是模拟信号，在对这些模拟信号进行预加重处理之前，需要经过采样和量化将模拟信息转化为数字信号，根据语音的频谱范围 200-3400Hz，采样率可设置为 8KHz，量化精度为 16bit。

应理解，此处采样率和量化精度的数值范围，为本申请优选范围，但可以根据实际应用的需要进行设置，此处不做限制。

语音信号在经过预加重后，频谱的高频部分得到提升，信号也变得平坦，生成具有平坦频谱的加重处理声纹样本，有利于后续的声纹特征提取。

S23：采用分帧和加窗的方式，对加重处理声纹样本进行分帧处理，得到初始语音帧。

具体地，语音信号具有短时平稳的性质，语音信号在经过预加重处理后，需要对其进行分帧和加窗处理，来保持信号的短时平稳性，通常情况下，每秒钟包含的帧数在 33~100 帧之间。为了保持帧与帧之间的连续性，使得相邻两帧都能平滑过渡，采用交叠分帧的方式，如图 3 所示，图 3 示出了交叠分帧的样例，图 3 中第 k 帧和第 k+1 帧之间的交叠部分即为帧移。

优选地，帧移与帧长的比值的取值范围为 (0, 0.5)。

例如，在一具体实施方式中，预加重后的语音信号为 $s'(n)$ ，帧长为 N 个采样点，帧移为 M 个采样点。当第 l 帧对应的采样点为第 n 个时，原始语音信号 $x_l(n)$ 与各参数之间的对应关系为：

$$x_l(n) = x[(l-1)M + n]$$

其中， $n = 0, 1, \dots, N-1$, $N = 256$ 。

进一步地，声纹样本经过分帧之后，使用相应的窗函数 $w(n)$ 与预加重后的语音信号 $s'(n)$ 相乘，即得到加窗后的语音信号 S_w ，将该语音信号作为初始语音帧信号。

其中，窗函数包括但不限于：矩形窗（Rectangular）、汉明窗（Hamming）和汉宁窗（Hanning）等。

矩形窗表达式为：

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

汉明窗表达式为：

$$w(n) = \begin{cases} 0.54 - 0.46 \cos(2 * \pi * n / (N-1)) & (0 \leq n \leq N-1) \\ 0 & (n < 0, n > N) \end{cases}$$

汉宁窗表达式为：

$$w(n) = \begin{cases} 0.5(1 - \cos(2 * \pi * n / (N - 1))) & (0 \leq n \leq N - 1) \\ 0 & (n < 0, n > N) \end{cases}$$

对经过预加重处理的声纹样本进行分帧和加窗处理，使得声纹样本保持帧与帧之间的连续性，并剔除掉一些异常的信号点，提高了声纹样本的鲁棒性。

S24：对初始语音帧信号进行静默音分离，得到目标语音帧。

具体地，在通话持续期间，语音信号可分为激活期和静默期两个状态，静默期不传送任何语音信号，上、下行链路的激活期和静默期相互独立。在步骤 S21 中，用户录音过程中在每次发音前后，均会有停顿的状态，这个状态会带来语音信号的停顿，即静默期，在进行声纹特征提取的时候，需要检测出静默期状态，进而将静默期与激活期进行分离，以得到持续的激活期，将保留下来的持续的激活期的语音信号作为目标语音帧。

其中，检测静默音状态的方式包括但不限于：语音端点检测、FFMPEG 探测音频静音算法和语音活动检测（Voice Activity Detection, VAD）算法等。

S25：基于目标语音帧，提取声纹特征。

具体地，在经过预加重处理、分帧和加窗和静默音分离之后，获取了稳定性强的声纹样本，使用该样本进行声纹特征的提取。

其中，声纹特征提取是提取并选择对说话人的声纹具有可分性强、稳定性高等特性的声学或语言特征。

优选地，本申请选择提取的声纹特征为声学特征中的线性倒谱特征。

S26：在每个分组中，将该分组中每个坐席员的声纹特征组成该分组的基础声纹特征集。

具体地，在步骤 S25 中获取到每个坐席的声纹特征之后，将每个分组的坐席的声纹特征组成一个声纹特征集，用于后续的动态声纹识别模型的构建。

在图 2 对应的实施例中，针对每个分组中的每个坐席员，获取该坐席员对预设的多个不同语料进行朗读的录音样本，并对该录音样本进行声纹解析，得到该坐席员的初始声纹样本，并对初始声纹样本进行预加重处理，生成具有平坦频谱的加重处理声纹样本，采用分帧和加窗的方式，对加重处理声纹样本进行分帧处理，得到具有鲁棒性的初始语音帧，然后对初始语音帧信号进行静默音分离，过滤掉有干扰的静默音，得到目标语音帧，使得声纹样本具有连续性和鲁棒性，进而从该声纹样本中提取坐席员的声纹特征，在每个分组中，将该分组中每个坐席员的声纹特征组成该分组的基础声纹特征集，提升了基础声纹特征集中声纹特征的准确度和鲁棒性。

在图 2 对应的实施例的基础之上，下面通过一个具体的实施例来对步骤 S25 中所提及的基于目标语音帧，提取声纹特征的具体实现方法进行详细说明：

请参阅图 4，图 4 示出了本申请实施例提供的步骤 S25 的具体实现流程，详述如下：

S251：使用莱文森-德宾算法对目标语音帧的预测误差计算公式（F1）中的线性预测系数 $a_1', a_2', \dots, a_p'$ 进行预测，并根据预测结果确定目标线性预测系数 $a_1', a_2', \dots, a_p'$ ：

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^p a_i s(n-i) \quad F1$$

其中， $e(n)$ 为预测误差， $\hat{s}(n)$ 为目标语音帧序列的预测值， n 为目标语音帧序列包含的总点数， $s(n-i)$ 为目标语音帧序列在前 $n-i$ 个点的取值总和， a_i 为目标语音帧序列中第 i 点的线性预测系数， p 为预设的线性预测的阶数， $i \in [1, p]$ ， n 为正整数。

具体地，由于语音相邻样点间具有相关性，可以利用线性预测，根据过去的样点值来预测现在或者未来的样点值，即利用过去若干个语音抽样或者他们的线性组合，来逼近该

条语音的现在的抽样值。预测的目的就是找出一组合适的系数,使得预测误差 $e^{(n)}$ 的均方值最小。实际预测过程一般是先把抽样序列按一定的数目组成帧,然后逐帧进行预测,每帧都找出该帧的P个最佳预测系数。预测的精确度不是以某一个样值的预测结果来衡量,而是要看帧内各样值预测的总效果。

在本申请实施例中,通过采用信号过去的P个样点值,来对信号现在或者未来的样点进行预测,得到预测值,假定语音信号的序列为 $s^{(n)}$,n为[1,n]的任意正整数, $s^{(n)}$ 为当前时刻n点的取值,其预测值为 $\hat{s}^{(n)}$,进而将预测值与真实值之差作为该点的预测误差。

进一步地,要使得预测精准,就需要降低预测误差,在本申请中采取的是对预测误差进行均方误差求导,得到尤尔-沃克方程(Yule-Walker Equation),并使用莱文森-德宾算法对尤尔-沃克方程进行求解,得到的一组线性预测系数即为使得预测误差值最小的线性预测系数,将该组线性预测系数作为目标预测系数。

S252: 根据目标线性预测系数,按照公式(F2)计算目标语音帧的线性预测倒谱系数,并将该线性预测倒谱系数作为声纹特征:

$$\hat{h}(n) = \begin{cases} a'_1 & n=1 \\ a'_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a'_k \hat{h}(n-k) & 1 < n \leq p \\ \sum_{k=1}^p \left(1 - \frac{k}{n}\right) a'_k \hat{h}(n-k) & n > p \end{cases} \quad \text{F2}$$

其中, $\hat{h}(n)$ 为线性预测倒谱系数。

具体地,依据n的取值范围来使用公式(F2)计算出对应的线性预测倒谱系数,并将得到的线性预测倒谱系数作为语音信号序列 $s^{(n)}$ 的声纹特征。

在图4对应的实施例中,通过对目标语音帧的预测误差的表达式进行求极值,并找出使得预测误差最小的一组线性预测系数作为目标线性预测系数,使用该目标线性预测系数通过公式(F3)计算得线性预测倒谱系数,将该线性倒谱系数作为声纹特征,使得采集的声纹特征精准度较高,有利于提高动态声纹识别模型的预测准确率。

在图1对应的实施例的基础之上,下面通过一个具体的实施例来对步骤S3中所提及的对每个基础声纹特征集中的声纹特征进行训练,得到每个基础声纹特征集的综合声纹特征的具体实现方法进行详细说明。

请参阅图5,图5示出了本申请实施例提供的步骤S3的具体实现流程,详述如下:

S31: 针对每个基础声纹特征集,采用遗传算法对该基础声纹特征集中的声纹特征进行优化,得到该基础声纹特征集对应的码本集。

具体地,采用遗传算法对该基础声纹特征集中的声纹特征进行优化的具体过程请参照步骤S3中的描述,为避免重复,此处不在赘述。

S32: 针对第x个基础声纹特征集,获取该基础声纹特征集对应的码本集内的码本总数q,在该码本集中任意选取k个码本作为训练样本,将训练样本的码本矩阵按对应位取算术平均值得到类别码本 D_x ,将该类别码本 D_x 作为第x个分组的基础声纹特征集的综合声纹特征,其中 $k \in [1, q]$,x和q为正整数。

可以理解地,针对每个分组的基础声纹集,均得到该分组的综合声纹特征,在本申请实施例中采用按对应位取算术平均值的方式得到类别码本,在其他实施例中,也可以根据

实际情况,选择加权平均值、几何平均值,平方平均值和调和平均值等方式得到类别码本,此处不作具体限制。

值得说明的是,在生成类别码本后,使用下述步骤 S52 中的公式 (F3) 来计算该分组内每个声纹特征到该类别码本的距离,并将这些距离与预设距离进行比较,若大于预设距离的声纹特征个数超过预设个数阈值,则判定该类别码本不符合要求,随机选取 k 个码本进行重新训练,直到生成的类别码本到该分组内各声纹的欧式距离,大于预设距离的个数不超过预设的个数阈值为止,将此时的类别码本作为该分组的类别码本。

在图 5 对应的实施例中,针对每个基础声纹特征集,采用遗传算法对该基础声纹特征集中的声纹特征进行优化,得到该基础声纹特征集对应的码本集,在每个基础声纹特征集中,获取该基础声纹特征集对应的码本集内的码本总数,在该码本集中任意选取预设个数码本作为训练样本,将训练样本的码本矩阵按对应位取算术平均值得到类别码本,按次方法得到所有分组的类别码本,通过提取每个分组的综合声纹特征,能够在检测到呼叫操作时,快速判断当前语音数据所属分组,提升识别效率。

在图 5 对应的实施例的基础之上,下面通过一个具体的实施例来对步骤 S5 中所提及的基于动态声纹识别模型,对目标声纹特征进行识别,确定待识别坐席员的呼叫权限的具体实现方法进行详细说明。

请参阅图 6,图 6 示出了本申请实施例提供的步骤 S5 的具体实现流程,详述如下:

S51: 将目标声纹特征进行预处理,得到目标声纹特征的目标优化码本。

具体地,对步骤 S4 中获取到的目标声纹特征进行预处理,预处理过程可参照步骤 S22 至步骤 S25 的描述,进而采用遗传算法对该基础声纹特征集中的声纹特征进行优化,具体优化过程请参照步骤 S3 中的描述,为避免重复,此处不再赘述,得到目标声纹特征的目标优化码本。

S52: 按照公式 (F3) 计算目标优化码本到每个类别码本的欧式距离:

$$\rho_j(A, D_j) = \sqrt{\sum_{t=1}^L (a_t - b_t)^2} \quad F3$$

其中, $\rho_j(A, D_j)$ 为目标优化码本 A 到第 j 个类别码本 D_j 的欧式距离, a_t 为目标优化码本 A 中第 t 个编码序列特征, b_t 为类别码本 D_j 中第 t 个编码序列特征, $j \in [1, J]$, J 为基础声纹特征集的个数。

其中, 编码序列特征是指每个目标优化码本中用来表示声纹特征的系数,例如步骤 S252 中的线性倒谱系数。

S53: 从 J 个欧式距离中选取数值最小的欧氏距离,作为目标欧氏距离。

具体地,由步骤 S52 可知,目标优化码本到每个类别码本均得到一个欧式距离,容易理解地,最后获取的欧式距离总数与分组的总数一致,将这些欧式距离进行比较,获取数值最小的欧式距离,作为目标欧氏距离。

例如,在一具体实施方式中,共分了 10 个组,即 10 个基础声纹特征集,每个基础声纹特征集对应存在 1 个类别码本,分别计算目标优化码本到每个类别码本的欧式距离,得到 10 个欧式距离,从这些欧式距离中选出数值最小的欧氏距离作为目标欧式距离。

S54: 获取目标欧式距离对应的类别码本所在的分组,将该分组作为当前外呼组别。

具体地,由步骤 S42 可知,每个类别码本对应一个分组,获取目标欧式距离对应的分组,将该分组作为当前的外呼组别。

S55: 根据对应关系,确定当前外呼组别对应的当前呼叫权限,作为待识别坐席员的呼叫权限。

具体地,根据步骤 S3 中确定的外呼组别与呼叫权限之间的对应关系,获取当前外呼

组别对应的呼叫权限，将该权限作为待识别的坐席员的呼叫权限。

在图 6 对应的实施例中，通过将目标声纹特征进行预处理，得到目标声纹特征的目标优化码本，进而计算目标优化码本到每个类别码本的欧式距离，并从所有欧式距离中选取数值最小的欧氏距离，作为目标欧氏距离，再获取目标欧氏距离对应的类别码本所在的分组，将该分组作为当前外呼组别，根据分组与呼叫权限的对应关系，确定当前外呼组别对应的当前呼叫权限，作为待识别坐席员的呼叫权限。从而实现了实时快速的对当前外呼进行权限的认定，有利于实现实时监控，有效提高监控的实时性和坐席呼叫的呼叫质量。

应理解，上述实施例中各步骤的序号的大小并不意味着执行顺序的先后，各过程的执行顺序应以其功能和内在逻辑确定，而不对本申请实施例的实施过程构成任何限定。

实施例 2

对应于实施例 1 中的声纹识别方法，图 7 示出了与实施例 1 提供的声纹识别方法一一对应的声纹识别装置，为了便于说明，仅示出了与本申请实施例相关的部分。

如图 7 所示，该声纹识别装置包括：坐席权限分组模块 10、声纹特征采集模块 20、识别模型构建模块 30、目标特征获取模块 40、呼叫权限确定模块 50 和实时语音预警模块 60。各功能模块详细说明如下：

坐席权限分组模块 10，用于获取数据库中每个坐席员的呼叫权限，并根据预设的权限分组方式，将每个坐席员分配到呼叫权限对应的分组中；

声纹特征采集模块 20，用于针对每个分组，对该分组中的每个坐席员进行声纹特征采集，得到该分组的基础声纹特征集；

识别模型构建模块 30，用于对每个基础声纹特征集中的声纹特征进行训练，得到每个基础声纹特征集的综合声纹特征，并基于分组、呼叫权限和综合声纹特征之间的对应关系，构建动态声纹识别模型；

目标特征获取模块 40，用于若检测到待识别坐席员的呼叫操作，则获取被呼叫客户的客户级别和待识别坐席员的语音数据，并对该语音数据进行声纹特征提取，得到目标声纹特征；

呼叫权限确定模块 50，用于基于动态声纹识别模型，对目标声纹特征进行识别，确定待识别坐席员的呼叫权限；

实时语音预警模块 60，用于若待识别坐席员的呼叫权限不满足客户级别的要求，则执行预设的预警处理措施。

进一步地，声纹特征采集模块 20 包括：

初始声纹获取单元 21，用于针对每个分组中的每个坐席员，获取该坐席员对预设的多个不同语料进行朗读的录音样本，并对该录音样本进行声纹解析，得到该坐席员的初始声纹样本；

声纹加重处理单元 22，用于对初始声纹样本进行预加重处理，生成具有平坦频谱的加重处理声纹样本；

初始语音帧生成单元 23，用于采用分帧和加窗的方式，对加重处理声纹样本进行分帧处理，得到初始语音帧；

目标语音帧获取单元 24，用于对初始语音帧信号进行静默音分离，得到目标语音帧；

声纹特征提取单元 25，用于基于目标语音帧，提取声纹特征；

声纹特征整合单元 26，用于在每个分组中，将该分组中每个坐席员的声纹特征组成该分组的基础声纹特征集。

优选地，声纹特征提取单元 25 包括：

目标线性预测系数递推子单元 251，用于使用莱文森-德宾算法目标语音帧的预测误差计算公式 (F1) 中的线性预测系数进行预测，并根据预测结果确定目标线性预测系数 $a_1', a_2', \dots, a_p'$ ；

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{i=1}^p a_i s(n-i) \quad F1$$

其中, $e(n)$ 为预测误差, $\hat{s}(n)$ 为目标语音帧序列的预测值, n 为目标语音帧序列包含的总点数, $s(n-i)$ 为目标语音帧序列在前 $n-i$ 个点的取值总和, a_i 为目标语音帧序列中第 i 点的线性预测系数, p 为预设的线性预测的阶数, $i \in [1, p]$, n 为正整数;

线性预测倒谱系数计算子单元 252, 用于根据目标线性预测系数, 按照公式 (F2) 计算目标语音帧的线性预测倒谱系数, 并将线性预测倒谱系数作为声纹特征:

$$\hat{h}(n) = \begin{cases} a'_1 & n=1 \\ a'_n + \sum_{k=1}^{n-1} \left(1 - \frac{k}{n}\right) a'_k \hat{h}(n-k) & 1 < n \leq p \\ \sum_{k=1}^p \left(1 - \frac{k}{n}\right) a'_k \hat{h}(n-k) & n > p \end{cases} \quad F2$$

其中, $\hat{h}(n)$ 为线性预测倒谱系数。

进一步地, 识别模型构建模块 30 包括:

特征码本优化单元 31, 用于针对每个基础声纹特征集, 采用遗传算法对该基础声纹特征集中的声纹特征进行优化, 得到该基础声纹特征集对应的码本集;

类别码本训练单元 32, 用于针对第 x 个基础声纹特征集, 获取该基础声纹特征集对应的码本集内的码本总数 q , 在该码本集中任意选取 k 个码本作为训练样本, 将该训练样本的码本矩阵按对应位取算术平均值得到类别码本 D_x , 将该类别码本 D_x 作为第 x 个分组的基础声纹特征集的综合声纹特征, 其中 $k \in [1, q]$, x 和 q 为正整数。

进一步地, 呼叫权限确定模块 50 包括:

目标优化码本获取单元 51, 用于将目标声纹特征进行预处理, 得到目标声纹特征的目标优化码本;

欧氏距离计算单元 52, 用于按照公式 (F3) 计算目标优化码本到每个类别码本的欧氏距离:

$$\rho_j(A, D_j) = \sqrt{\sum_{t=1}^J (a_t - b_t)^2} \quad F3$$

其中, $\rho_j(A, D_j)$ 为目标优化码本 A 到第 j 个类别码本 D_j 的欧氏距离, a_t 为目标优化码本 A 中第 t 个编码序列特征, b_t 为类别码本 D_j 中第 t 个编码序列特征, $j \in [1, J]$, J 为基础声纹特征集的个数;

目标欧式距离选取单元 53, 用于从 n 个欧式距离中选取数值最小的欧氏距离, 作为目标欧氏距离;

当前外呼组别确定单元 54, 用于获取目标欧式距离对应的类别码本所在的分组, 将该分组作为当前外呼组别;

当前呼叫权限获取单元 55, 用于根据对应关系, 确定当前外呼组别对应的当前呼叫权限, 作为待识别坐席员的呼叫权限。

本实施例提供了一种声纹识别装置中各模块实现各自功能的过程, 具体可参考前述实

实施例 1 的描述，此处不再赘述。

实施例 3

本实施例提供一个或多个存储有计算机可读指令的非易失性可读存储介质，该非易失性可读存储介质上存储有计算机可读指令，该计算机可读指令被一个或多个处理器执行时实现实施例 1 中声纹识别方法，或者，该计算机可读指令被一个或多个处理器执行时实现实施例 2 中声纹识别装置中各模块/单元的功能。为避免重复，这里不再赘述。

可以理解地，所述非易失性可读存储介质可以包括：能够携带所述计算机可读指令代码的任何实体或装置、记录介质、U 盘、移动硬盘、磁碟、光盘、计算机存储器、只读存储器（Read-Only Memory, ROM）、随机存取存储器（Random Access Memory, RAM）、电载波信号和电信信号等。

实施例 4

图 8 是本申请一实施例提供的终端设备的示意图。如图 8 所示，该实施例的终端设备 80 包括：处理器 81、存储器 82 以及存储在存储器 82 中并可在处理器 81 上运行的计算机可读指令 83，例如声纹识别程序。处理器 81 执行计算机可读指令 83 时实现上述各个声纹识别方法实施例中的步骤，例如图 1 所示的步骤 S1 至步骤 S6。或者，处理器 81 执行计算机可读指令 83 时实现上述各装置实施例中各模块/单元的功能，例如图 7 所示模块 10 至模块 60 的功能。

示例性的，计算机可读指令 83 可以被分割成一个或多个模块/单元，一个或者多个模块/单元被存储在存储器 82 中，并由处理器 81 执行，以完成本申请。一个或多个模块/单元可以是能够完成特定功能的一系列计算机可读指令的指令段，该指令段用于描述计算机可读指令 83 在终端设备 80 中的执行过程。例如，计算机可读指令 83 可以被分割成坐席权限分组模块、声纹特征采集模块、识别模型构建模块、目标特征获取模块、呼叫权限确定模块和实时语音预警模块，各模块的具体功能如实施例 2 所示，为避免重复，此处不一一赘述。

所属领域的技术人员可以清楚地了解到，为了描述的方便和简洁，仅以上述各功能单元、模块的划分进行举例说明，实际应用中，可以根据需要而将上述功能分配由不同的功能单元、模块完成，即将所述装置的内部结构划分成不同的功能单元或模块，以完成以上描述的全部或者部分功能。

以上所述实施例仅用以说明本申请的技术方案，而非对其限制；尽管参照前述实施例对本申请进行了详细的说明，本领域的普通技术人员应当理解：其依然可以对前述各实施例所记载的技术方案进行修改，或者对其中部分技术特征进行等同替换；而这些修改或者替换，并不使相应技术方案的本质脱离本申请各实施例技术方案的精神和范围，均应包含在本申请的保护范围之内。