智能风险控制

2020 年 3 月 17 日

Springer

目录

Chapter 1
智能风险控制

1.1 基于人脸的生物识别

人脸识别，是指对输入的图像和视频，检测其中存在的人脸，依据人脸的面部特征，完成身份识别的过程，属于生物特征识别技术。整个流程包含人脸检测、人脸对齐、人脸特征提取、人脸匹配几个阶段，如图 1-1 所示。目前人脸识别已经广泛应用于安防、金融、军事等领域。

人脸识别具有以下优点：

自然性：所谓自然性，是指人脸识别技术所利用的生物特征，与人类进行人脸识别时所利用的生物特征是一致的，与之相比，虹膜识别、指纹识别等技术，则不具备自然性。

非接触性：在人脸识别技术中，用户不会与识别设备发生任何接触，对于用户来说体验较好。而指纹识别则需要用户进行按压设备。

使用便捷：用户使用人脸识别技术时非常方便，基本上无需做特殊的配合。

人脸识别也具有一些缺点，比如易受光照条件的影响，易受人脸遮挡物的影响，跨年龄识别难度较高等。但总的来说，人脸识别是目前一种可靠的，实用的，便捷的身份核验技术。

图 1-1 人脸识别流程图

1.1.1 人脸检测及对齐

人脸检测是人脸识别的第一步，属于目标检测的子方向。其目的是找出图像中的人脸以及对应的位置。可能还会包含一些人脸的额外信息，比如人脸的关键点，姿态角度等。

典型的人脸检测是基于以下的流程：由于人脸可能出现的图像的任何位置，因此需要通过滑动窗口（sliding windows）来获取可能包含人脸的子图像。获取到的子图像，需要通过一个二分类的分类器，来判断图像中是否包含人脸，如果还需要确定人脸的精确位置，还需要加上一个回归人脸框的操作。同一个人脸可能会检测出多个人脸框，因此需要使用非极大值抑制(Non-Maximum suppression, NMS)来进行合并去重。接下来本文介绍一些具有代表性的人脸检测方法。

Viola-jones[**?**]使用Haar-like小波特征，并通过级联的AdaBoost分类器构造检测器。该方法具有检测效率高，并且能够保持较好的精度的特点，是第一个具有实用意义的人脸检测算法。MTCNN[**?**]将人脸分类、人脸框回归以及人脸关键点定位在同一个任务内完成了，是一个多任务(multi-task)的检测方法，这种思路在后续的很多方法里也得到了使用。

anchor的思想在目标检测方法Faster-rcnn[**?**]中首先被提出，在人脸检测中也经常使用到。anchor提出的目的是为了解决目标在图像中可能以不同的形状存在，比如不同的长宽比，所以加入人工的先验信息，预先定义不同比例的anchor来进行候选目标框的获取。Face r-cnn[**?**], Pyramidbox[**?**], Retinaface[**?**]这些方法，都用到了anchor的思想。另一个在人脸检测中经常使用的思想是特征金字塔网络(feature pyramid network)，为了自适应不同尺度人脸的检测，一般有两种做法，一种做法是图像金字塔，这种方法需要对输入图像做不同尺度的缩放，缺点是耗时较高；另一种更好的做法则是特征金字塔，其思想是在不同分辨率的特征图(feature map) 上检测对应尺度的目标，同时将不同分辨率的特征图与更高层的特征图进行特征融合，保证每一层的特征图都具有足够的表达能力。Pyramidbox,Retinaface都用到了特征金字塔，SSH[**?**]虽然没有直接用到特征金字塔，但其也是对网络3个不同尺度的特征图进行分别预测，来解决多尺度的人脸检测问题。

做完人脸检测后，一般需要进行人脸对齐。通过对人脸进行关键点定位，以及预先定义好的关键点模板，进行仿射变换，通过旋转、平移、

缩放等操作，进行人脸对齐，对齐后的人脸能够更好的进行人脸特征提取。目前常见的关键点个数，有5个关键点、68个关键点、90个关键点以及106个关键点等。

1.1.2 人脸特征提取

特征提取是人脸识别的关键步骤，它将人脸图像映射都某个特征空间中，使得映射后的特征能够很好地区分不同人之间的差异点。经过特征提取得到人脸的特征表示之后，可以进行特征匹配。如果是对两个特征进行比对，我们一般称为人脸比对或者人脸验证(verification)，如果是将一个特征与一组特征进行匹配，我们一般称为人脸检索或者人脸识别(identification)，如图 1-2 所示。

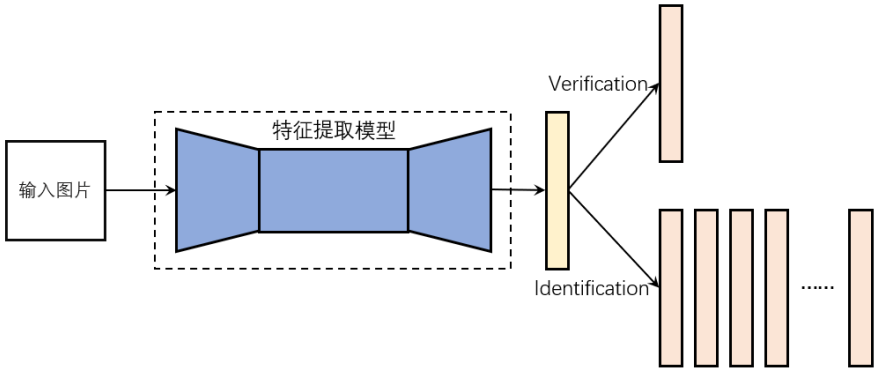


图 1-2 人脸特征抽取

传统的特征提取算法，通过一些降维方法，得到一系列降维后的特征，用来表示人脸。比如使用PCA进行降维的EigenFace，基于LDA进行降维的FisherFace等，都是早期人脸识别中非常经典的算法。但这些方法存在一些缺点，对光照、表情、姿态敏感，泛化能力不足，因此在实际使用中的准确度不高。

随着深度学习的广泛应用，越来越多具有实用价值的方法被提出，人脸识别的研究得到了极大的发展。基于深度学习的特征提取方法可以分为两大类：

度量学习(metric learning)

通过一个度量函数，来衡量相同人或者不同人的特征表示之间的距离，从而学习到每个人每个照片的特征表示，基本思路是同一个人的特征表示之间的距离尽可能小，不同人的特征表示的距离尽可能大。这一个方向的典型方法包含2014年的Contrastive loss[**?**]和2015年google提出的Triplet loss[**?**]。

DeepID2是基于Contrastive loss的模型，它在训练的时候，同时训练classification和verification两个信号，其中的verification信号，就是用特征表示之间的Contrastive loss来构造的。Contrastive loss是基于pairwise的思想，模型训练时，需要输入两张图片，如果两个图片是同一个人，则verification的标签为1，如果不是同一个人，则标签为0。

google于2015年提出的Facenet中，则用到了Triplet loss。其思想是以三元组的形式来训练模型，每次输入需要三张图片，其中两张图片是同一个人，以及一张其他人的图片，要求同一个人的照片之间的距离要小于不同人之间的距离，且要超过一个margin。

基于margin的分类方法(margin based classification)

第二类思想是基于分类的思想来进行特征提取，根据训练集中的数据，同一个人的照片属于同一类，训练集一共包含多少个id，则总共需要分多少类。由于用的分类的思想，所以自然而然可以使用分类的损失函数。而在此基础上，又提出了一系列的方法，用以最小化类内间距或者最大化类间间距。比较有代表性的方法有Center loss[**?**]，SphereFace[**?**]，CosFace[**?**]以及ArcFace[**?**]等。

由于是多分类任务，所以最基本的损失函数形式是softmax loss。但是直接用softmax loss训练出来的特征，往往效果不理想，某些类别的类内间距甚至比类间间距大，导致人脸识别的时候出现错误。Center loss引入了类内中心，为每个类别提供一个类内中心，最小化训练集中每个样本与其类内中心的距离，从而达到减少类内间距的效果。

基于SphereFace的训练方式，是在此基础上做了改进，对权重进行了归一化，且增加了角度裕量，在cos函数上对角度乘上因子m，加大分类

难度。CosFace和ArcFace更进一步，对特征表示也做了归一化，并分别引入了不同的margin形式，取得了更好的效果。

以上方法都是通过不同方式去减少同一个人的类内间距以及增大不同人之间的类间间距。但除了损失函数以及网络的设计之外，更为重要的是训练的数据的分布，比较好的训练数据是同一个人包含多张不同的照片，这些照片覆盖此人不同年龄段，不同姿态角度，不同遮挡程度，不同妆容情况等，这样的数据能够学习到鲁棒性更强，通用性更好的模型。

1.1.3 应用案例

人脸识别目前在安防、金融等领域都得到了广泛应用，下面介绍一些常见的应用案例。

门禁闸机

这是人脸识别的典型应用场景，属于人脸检索（1：N）的应用。门禁闸机在初始化的时候，会要求录入一个人脸库，该人脸库经过特征提取后，作为识别的底库。当有人通过闸机的时候，会拍摄来人的照片，通过特征提取转化为特征表示之后，与底库中的特征集合进行对比，找出该人员是否存在于底库中。

金融核身业务

目前几乎所有的金融核身业务都支持人脸核身功能，属于人脸比对（1:1）的应用。当用户的办理某些业务的时候，会被要求进行人脸核身，系统会通过摄像头采集用户的照片，与用户留底的另一张照片进行比对，以确定用户是否为本人。这种方式大大减少了金融业务中进行业务审核的人员数量及审核时间，节省了用户时间，提升了用户体验。

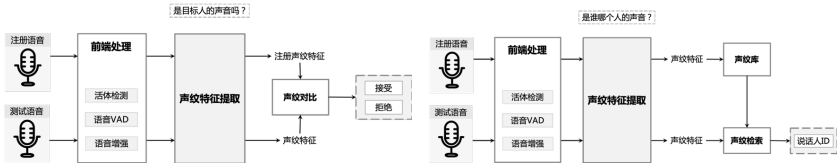
1.2 基于声纹的生物识别

声纹识别（Voice Print Recognition），也称作说话人识别（Speaker Recognition），是一种生物识别技术，能够根据说话人的声音特征提供精准、高效、便捷的身份识别服务。从感官直觉上来说，声纹虽然不像人脸、指纹的个体差异那样直观可见，但由于人在讲话时使用的发声器官–舌、牙齿、喉头、肺、鼻腔在尺寸和形态方面每个人的差异很大，因此反映到任何两个人的声纹图谱都存在有差异。最直观的感受是当我们打电话给认识的人的时候，通过很短一句话甚至一声“喂？”，就能准确地分辨出接电话的是谁。这种语音中承载的说话人身份信息的唯一性使得声纹也可以像人脸、指纹那样作为生物信息识别技术的生力军，可广泛应用于金融安全，公共安防，智能家居等领域。

声纹识别通常分为两大类，说话人确认和说话人辨别，也就是常说的声纹1：1识别和声纹1：N识别。声纹1：1识别是指确认某段语音是否是指定的某个人所说的，而声纹1：N识别是判断某段语音是若干人中的哪一个所说的。不同的任务和应用会使用不同的声纹识别技术，如诈骗电话需要缩小人员范围时可能需要声纹1：N技术进行辨别，而银行金融交易时则需要声纹1：1识别技术进行确认。从另一方面，声纹识别也分为文本相关（Text-Dependent）和文本无关（Text-Independent）两类。与文本相关的声纹识别要求用户按规定内容进行朗读，这样能更加准确的建立模型，在识别的时候也要求用户按规定内容朗读。文本无关的声纹识别并不要求用户根据指定文本进行朗读，这样也能建立模型，验证的时候同样不需要用户根据指定文本进行朗读。一般来说，文本相关的声纹识别效果会更好，安全性更高，但是用户体验较差和使用场景就相对较窄，通常用于安全要求比较高的场景，如金融核实身份。文本无关的声纹识别会和文本依赖比较弱，因此能进行跨语种使用，就算我们没有别的语种的语料，也是能应用到那语种上面去。文本无关的使用场景很宽广，在金融客服上就能建立声纹库，后面的用户在金融上快速验证和关联。

声纹识别的流程如图 1-3所示。不管是1：1还是1：N大致都是分为三部分：前置处理，特征抽取和声纹匹配。前置处理通常是VAD检测、反欺诈活体检测、声音增强等。VAD检测处理是把有声音部分和静音部分区分出来，把有声音部分送到后续处理中去。反欺诈活体检测是主要应

对声纹识别是否被攻击，提高安全性。这些前置处理不在这里详细展开细说，匹配算法我们可以用最普通的计算向量距离如欧式距离或cos值，也可以用深度学习网络进行计算相似度。接下来我们重点介绍声纹的特征抽取。



(a) 声纹1：1识别 (b) 声纹1：N识别

图 1-3 声纹识别流程

1.2.1 特征抽取

声纹识别的特征抽取大致经历了三代算法，GMM模型到ivector最后到深度学习网络。我们重点介绍前面两代，深度学习网络的方法和人脸特征抽取方法类似，也是分为两大类：度量学习(Metric Learning)和基于margin的分类方法(Margin Based Classification)，将会在人脸识别中重点介绍。

1.2.1.1 GMM-UBM模型

GMM模型 [**?**]即高斯混合模型，是由多个高斯函数进行加权求和进行拟合复杂的函数，如

$$p(x) = \sum_{k=1}^{K} p(k)p(x|k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k) \tag{1.1}$$

其中，π_k表示第k个高斯函数的权重，μ_k和Σ_k表示第k个高斯的均值和方差。

GMM模型的参数求解，一般使用EM算法进行求解。通常情况下，GMM模型可以平滑地逼近任意形状的函数，具备对实际数据极强的表

征力。而声纹识别实际上就是从不同语音中抽取出相同的表征特征来。GMM模型同时还具备比较好的泛化能力。因此GMM模型在声纹识别初期获得比较好的效果。随着k的增大，所需要的训练数据也就更加大了，否则获取不到泛化能力较好的模型。

在实际使用过程中，每个人语音数据有限，很难获取到比较通用的声纹识别模型。为了解决这个问题，DA Reynolds提出了通用背景模型(Universal Background Model) [**?**]，简称UBM。先使用大量和说话人无关的语音数据训练一个GMM模型，然后再使用少量的说话人数据，通过自适应算法(如最大后验概率MAP、最大似然线性回归MLLR等）获取到说话人的个性特征的模型叫做UBM模型。这个思想有点像现在深度学习的finetune思想。这个模型就是GMM-UBM模型。该模型参数可以减半并有更快收敛的特点。

随着实际应用，GMM-UBM的存在问题：参数仍然很大和受信道的干扰比较大。学术界提出了GMM-SVM 模型 [**?**]、JFA模型 [**?**]等等去优化解决。

1.2.1.2 ivector模型

基于GMM-UBM的模型，基本是基于特征声纹空间与特征信道空间的独立假设，但是在现实使用中，数据之间都是具有相关性的。之前的假设更多是方便了公式推导同时也限制了模型的泛化能力。N.Dehak认为既然声纹信息与信道信息不能做到完全独立，那就用一个一段低维度的定长向量同时描述声纹信息和信道信息，从而提出了ivector模型 [**?**]。

对于每一段语音都有高斯均值向量M表示如下：

$$M = m + \omega T \tag{1.2}$$

其中，m表示通用背景模型(UBM)的高斯均值向量，该值和声纹信息、信道信息无关，ω是全局差异空间因子，即为ivector向量，它的先验服从标准正态分布$N(0,1)$，T表示全局的差异空间矩阵。接下来只需要估计ω和T值即可。

对于ω和T的参数估计，我们基于假设：每一段语音都来自不同的说话人。首先计算训练数据中每个说话人所对应的Baum-Welch统计量，随后随机产生T的初始值。后续采取EM算法估算得到相关的参数。

1.2.2 应用案例

1.2.2.1 电话客服核身

在电话客服中应用声纹核身，可以节省核身时间，降低运营成本；并减少核身问题，提升客户体验；更重要的是，即使犯罪分子掌握了用户的所有信息，也能通过声音判断是否为本人，是否存在欺诈风险。例如，部分银行的服务热线目前已接入声纹核身技术，大致流程如图 1-4。

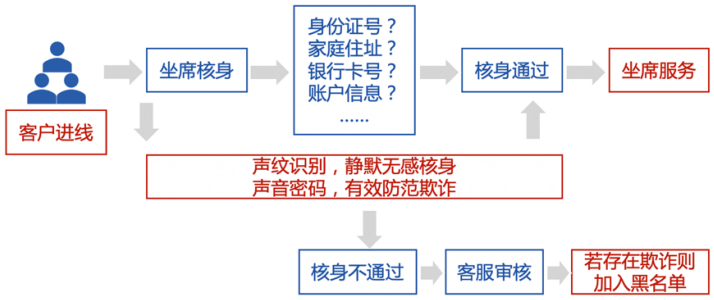


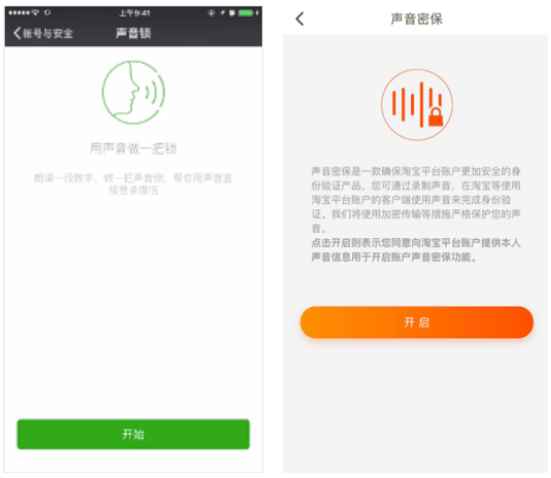
图 1-4 电话客服核身基本流程

1.2.2.2 社保核查

声纹验证能够解决参保人员面临的远程和现场身份核查及生存验证的问题，避免了指纹验证和人脸识别等需要现场办理、不易采集、伪造等问题，有效杜绝冒领养老金的可能性，节约社保资金和人力成本。例如，印尼新一代养老金认证系统已通过声纹验证技术，使其250万离退休人员在领取养老金时可通过电话或手机app进行远程身份认证，不仅节省了大量人力投入，还显著降低了传统骗保率。

1.2.2.3 声纹锁

如图 1-5所示声纹验证可作为登录、密保、修改账户信息的一种验证方式，也可以应用于门禁/闸机等。



(a) 声纹锁（一） (b) 声纹锁（二）

图 1-5 声纹锁样例

1.2.2.4 声纹支付

央行发布声纹识别安全应用技术标准，认定声纹技术可适用于手机银行、第三方支付，通过与语音交互硬件结合，能够解决无屏或远场进行身份验证的痛点。

(a) 声纹支付 (b) 声纹支付界面

图 1-6 声纹支付样例

1.2.2.5 声纹唤醒

通过声纹识别“主人”身份，只允许主人唤醒设备，或定制个性化服务，可适用于手机助手、智能音箱、智能家居、车载助手、服务机器人等智能设备。如图 1-7所示，在车联网应用中用户可以提前注册声纹信息并添加个性化配置，车机将通过声纹识别确认当前的驾驶人身份，可快速切换至对应的用户配置，令行车体验更加轻松。

图 1-7 汽车声纹唤醒

1.2.2.6 公安

以破案、追逃为导向，利用声纹识别技术公安可进行“案查人”、“人查案”、“案查案”与“人查人”等多种排查方式：

- 案查人：如电信诈骗，主要线索只有语音的情况下，将该语音进行声纹库大库检索，快速锁定嫌疑人。
- 人查案：公安抓捕到可疑人员后，提取出该人的声纹特征，将其放入尚未侦破的语音案件中，排查该人是否为在逃人员。
- 案查案：公安人员可使用声纹识别技术将尚未侦破的语音案件以及语音线索归纳整理，从中排查是否有多起案件是同一人所为，帮助侦察人员获得更多线索，提高排查效率。
- 人查人：公安机关在抓捕到可疑人员后，提取出该人的声纹特征，为避免该人使用伪造身份，可将其声纹特征放入已知人员的声纹库，查询其真实身份。

声纹识别技术还能应用于重点人员监管、反电信诈骗、反恐、刑事案件侦破、身份查询与核验，助力公安有效遏制与打击犯罪，构建和强化安全的社会公众环境。例如：

- 反电信网络欺诈：在通信系统或安全监测系统中嵌入声纹识别技术，能够对黑名单人员语音对话实时预警，提示重点人员可疑行为；语音内容关键词识别动态预警，提示可疑案件与犯罪意图。
- 动态声纹布控：通过声纹识别和声纹大数据技术，进行对重点人员和关键卡口的布控监管，在第一时间完成举报人或嫌疑人身份鉴定，辅助刑事案件侦破和案情分析。

图 1-8 公安布控监管系统

1.2.2.7 金融黑名单识别

将信贷黑名单、风险等级高、不良中介、金融欺诈等用户声纹加入黑名单库，当其再次办理业务时，匹配到黑名单库的用户，直接给出风险预警。例如，车险业务能够针对报假案、修理厂、黑中介等不良用户建立黑名单声纹库，当不良用户再次报案时，业务员端能够及时给出预警。

1.3 反欺诈

1.3.1 人脸识别中的反欺诈

人脸活体检测（ Face Anti-Spoofing）技术是人脸识别系统中，用以确认待认证对象是否为真实生物活体的一项技术。一方面，人脸识别技术的商业化愈加成熟和广泛，极大改善和推进了社会金融活动的智能化和便捷性；另一方面，由于人脸照片、视频数据相对容易获取和复制，若无活体检测这一环节，那么使用被盗取的合法用户的照片、视频或者3D面及头套等即可入侵人脸识别系统，由此将带来极大危害[**?**]。在目前的人脸识别系统中，常见的活体检测技术包括动作活体、3D活体、红外活体、光线活体等。下面将逐一简单介绍。

动作活体是通过利用人脸关键点和人脸跟踪等技术，检测用户眨眼、张嘴、摇头等多种动作及其组合，可有效抵御照片、换脸、面具、遮挡以及屏幕翻拍等常见的攻击手段，但较难抵御视频回放攻击。

3D活体通过专用硬件（例如3D结构光、ToF等）获取人脸部3D结构信息，可以有效防御如手机、电脑等屏幕显示和打印照片等2D攻击手段，但是需要配合其他方法抵御近几年出现的高质量3D面具攻击[**?**]。

红外活体检测一般利用人体皮肤对近红外光的反射率较高，相比于其他材质有明显区别的特性，通过专用红外设备获取人脸部红外图像判断是否为活体。实际应用中通常使用主动红外摄像，即通过红外LED照射人脸，利用红外摄像头获取人脸部图像，分类判断是否为活体。红外活体检测技术对于常见攻击手段具有较好的防御效果，缺点在于需要特定红外设备[**? ?**]。

光线活体是近两年出现的一种活体检测技术。由于3D活体、红外活体需要特殊设备，在已有系统中部署较为困难。光线活体技术利用屏幕发出不同颜色和强度的光线照射在人脸上，由于人脸自身的三维结构以及皮肤等生理组织对于不同颜色光线的反射率不同，从获取的视频中提取相应的活体信息，如图**??**所示。这项技术由于无需特殊硬件设备、且具有较高的准确率，在手机等移动端使用较为方便。其缺点在于要求视频拍摄过程稳定，闪光带来的用户体验需要得到提升，同时户外强光也会带来较大干扰[**?**]。

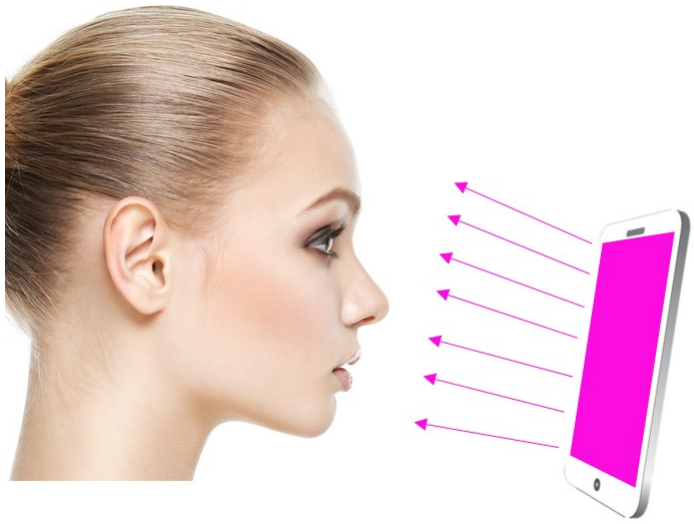


图 1-9 光线活体

1.3.2 声纹识别中的反欺诈

随着声纹识别的应用越来越广泛，声纹识别的安全性要求也越来越高。因此声纹识别研究中，必须了解声纹识别的攻击技术，才能进一步进行防范。在声纹识别应用中，主要的攻击方法分为四大类：模拟(Impersonation)、录播(Replay)、语音合成(Speech Synthesis)和语音转换(Voice Conversion)。

模拟攻击通常来自两种情况：双胞胎和专业模仿人员。双胞胎主要是因为生物信息相对比较接近，所以存在声纹接近的可能。曾经有报道说有双胞胎弟弟用自己声音破解了哥哥在银行的声纹密码。有论文 [**?**]提到这个问题，从他们的实验结果来看，就算是双胞胎还是存在一定的差异性，能进行区分。对应专业模仿人员，也有数据 [**?**]显示具有声纹攻击性，但是差异比双胞胎大。这一类攻击最大的问题就是可供训练和测试的数据很有限，小数据上能进行防范，但是对于其真正的攻击情况目前认为是不确定的。

录播攻击，一般是用录音机对特定人进行录音，然后播放出来从而达到声纹认证通过的效果。在ASVspoof2017 [**?**]比赛中，通过参赛人的比赛结果以及对往届的比赛结果来看，录播攻击识别难度比语音合成、

语音转换攻击识别难度更高。录播攻击其中涉及到录音环境、播放环境、录音设备等干扰，会一定程度对识别是否是攻击造成一定难度。

语音合成攻击，一般分为两类，一个是拼接方式，一个是TTS语音生成方式。针对TTS，机器识别准确一般比较高，因此不会造成比较大的威胁。但是语音拼接，这种技术方式，针对我们人耳能准确识别，但是对于机器而已，他们比较大可能认为是真实说话声音。

语音转换是把一个人说话的录音改变音色，频率等转换为具有另外一个人说话特点的技术 [**?**]。一般会用语音转换和TTS技术进行结合进行攻击，这样能比较方便有效，相对攻击成本比较低。这种攻击方式通常都能识别出来，机器生成的毕竟存在一定模式规律。但是也会存在一种风险，这种防止攻击的鲁棒性不一定很高。

针对这一系列的攻击方法，通用的抗攻击框架是在声纹识别前增加一个前置模块，用来专门识别是否是攻击。目前业界对这些攻击方式能较为准确识别，但是需要不断积累足够大的样本，从而提高更高的防护能力。