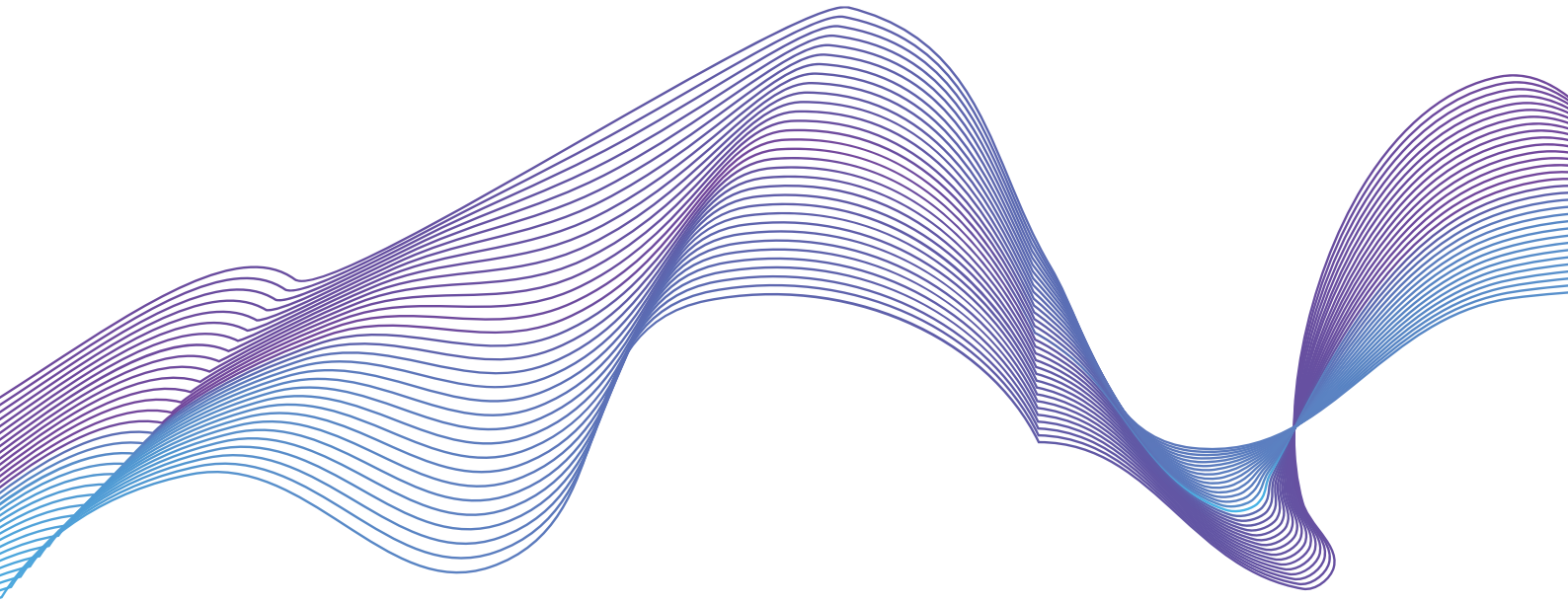
中国声纹识别产业发展白皮书

（2023 年）



《中国声纹识别产业发展白皮书》编委会

2024 年 3 月

编制声明

编委会

目录

面向应用需求的
声纹识别技术创新

现在针对声纹识别的研究非常多, 论文数量也在逐年增加, 可以说是百花齐放, 这是一个非常好的事情。由于我接触产业相对比较多, 过程中也发现一些问题, 比如有的声纹系统上线之后, 过一段时间就面临被下线, 或者应用单位反映不满足业务要求等等, 这些现象可能会对声纹的研究有一些打击, 对声纹的发展也有一定的影响。

我对这些问题进行了思索。我们语音界的老前辈,包括吴宗济先生等,把语音技术叫言语工程, 顾名思义言语工程它的中心词是工程, 也就是说我们的研究内容一定要面向应用, 解决应用所面临的问题。当然一些基础的、理论方面的研究也很重要, 这个研究积累需要更长的时间, 像专委会的主任党建武教授, 北大的吴玺宏教授, 还有社科院李爱军教授等, 在相关的领域做研究一做就是十年、二十年甚至超过三十年, 积累时间很长, 这种理论研究非常耗时耗力。而应用方等不了这么长的时间, 那就需要我们在做研究时更接近实际需求, 所以我们提出叫场景创新或者应用创新。

根据与业界的沟通交流和实践, 我总结了 5 个方面建议, 供各位同行参考:

一、建议加大技术安全方面的研究

音频伪造问题有两个方面, 一个是 spoof (模仿) 。简单来说, 这里攻击人要想让声音更接近目标人的声音, 要解决 anti-spoofing 的问题。第二个问题是 disguise (掩盖) 。我们发出的声音要和当前人的声音远离, 涉及到 recover 的问题, 即要进行声纹的恢复。在这两个方面的研发也出现了很多不如意的地方, 比如现在的 spoof 或者 anti-spoof, 很多方法很依赖数据, 在可解释性方面做得不够, 这就导致换个数据, 结果就不好了, 所以要不停地搜集数据, 但是新的攻击又总是会不停地出现, 这就会应接不暇; disguise 方面的研究与此类似, 利用变声器等改变声音的设备也很多, 此次孙蒙老师团队在刚刚举行的极棒 (Geek Pwn) 比赛中获得了第二名, 这说明有了初步的进步, 但我觉得还是任重道远, 因为我们无法穷尽各种变声器。这是从技术上来讲, 同样, 很多应用单位提出的实际需求都很大。

二、建议加大数据安全和隐私保护方面的研究

《个人信息保护法》和刚刚颁布的国标《声纹识别数据安全要求》值得关注。我们不仅仅要在应用开发过程中遵守法律和符合标准, 我想建议的是我们通过研究思考、通过与应用需求单位的充分沟通, 能够在遵循法律以及符合数据要求标准的前提下, 设计出能更好避免隐私泄露、保护个人隐私的一些技术方案 (比如高安全、弱隐私的动态声纹密码方案) , 再加上我们的技术实现手段, 从而做出可以得到认可、可以持续的产品。

三、建议增加提高用户体验方面的研究

我们还发现，有些系统应用不久被下线，也有的是由于用户体验的问题。比如说，时变问题，我们的声音会随着年龄的变化而变化，早中晚也会有变化，这种变化如何去适应？现在呢，很多应标的技术公司对这个问题都没有很好地解决；还有低语识别问题，即开会的时候如果用低语（悄悄话），能否准确识别身份；还有短语音问题，尤其是在文本无关的声纹识别过程中，你能否用比较短的几个音节把身份确定，这不管是在确认还是辨认的任务中都是很高的要求，若话说得太长了则体验不好，而且在有的场景下，若技术不能覆盖太短语音的时候，声纹识别就远远不能满足解决实际问题的需要。

四、建议加强对开集识别问题的研究

声纹识别的任务跟语音识别相比更是个开集的问题。语音识别的（建模和识别）目标是音素，对于特定的语言来讲其实它是固定的；声纹则不是，不同的应用甚至是同样的应用，它的目标用户集都是动态变化的，因此讲它是开集问题。我们发现的一个非常重要的现象是，我们在一些固定的闭集上测试效果很好的方法，包括 DNN，当用户集发生巨大变化的时候，其效果跟我们在闭集测试完全不同，下降很快。那么我们的声纹识别研究也应该关注这个问题，因为这个问题不解决的话，我们的成果可能更多只是反映在文章上而不是在应用中能解决问题。

五、建议加强非完整信息下多特征深度融合的研究

我们现在很多研究在做多模态的融合，声纹也好，人脸也好，指纹也好，更多是在决策级或者分数级，但是有些情形如果能做得更好，会大幅提高用户体验，提高安全度。比如说我们的人脸不一定要对得那么准，我的声音也可能有一半没听到或没听清，这叫非完整信息，我们能否做到深度的特征融合？能否对有效信息进行精准的提取和进行精准的比对，最后把两种甚至多种不同生物特征在特征层深度融合到一起，做一个最终综合判决？这个如果能解决的话，用户的体验就会非常顺滑，而且由于是多特征的应用，会使准确率、安全度等均有很大的提高。

再给大家分享一下来自业界的一个信号：我们最近发现有些部门、一些大国企的业务口，现在提出来，能不能只通过声纹这一种手段来进行身份认证？这种情况在以前几乎没有遇到过。我想它主要是源于法律和标准的要求在不断严格，以及用户对隐私保护的需求和意识在不断加强。所以我建议就上面几个问题进行深入研究，希望关注声纹识别研究和开发的相关人员能够多多关注。

声纹如果能够面向应用、面向场景，一定大有作为！谢谢大家！

CCF 语音对话与听觉专委会副主任
中国中文信息学会语音专委会主任
郑方

一、环境篇

1.1 AI 安全和隐私监管日益加强

智能语音系统的发展推动着人类生活的进步, 但人机之间实现便捷沟通的同时也暴露了各种隐私和安全问题。例如, 语音助手个人数据隐私保护、语音深度伪造和欺诈等。近年来, 生成式模型和大模型等技术在多个领域取得了显著进展, 为 AI 应用提供了丰富的创新空间。然而, 这种强大的生成能力也为安全领域带来了挑战。继计算机视觉领域频频爆出 AI 换脸、合成诈骗等安全隐患之后, “ 语音 + 安全 ” 的话题也日渐引起业界重视。“ 可信 AI” 成为近两年来人工智能领域的热词。另一方面, 深度学习技术的快速发展在促进语音技术产业应用的同时, 其弊端也开始逐渐凸显。采用深度学习技术建造的语音系统易受攻击、稳定性不高、可解释性和鲁棒性较差等问题日益突出。如何采用可解释的第三代人工智能理论与方法, 发展更少隐私收集的声纹识别技术, 打造可信任的声纹识别应用, 对行业发展意义重大。(参见李荪等编著《AI 智能语音技术与产业创新实践》)

2020 年 2 月, 中国人民银行发布《个人金融信息保护技术规范》, 继 2018 年的《移动金融基于声纹识别的安全应用技术规范》后, 对声纹技术的产业化应用再一次产生了重要的推动和示范作用。该规范 “ 根据信息遭到未经授权的查看或未经授权的变更后所产生的影响和危害 ”, 将个人金融信息按敏感程度从高到低分为 C3、C2、C1 三个等级。其中, “ 动态声纹密码 “ 被列入较低隐私敏感度级别的 C2 级个人信息, 与被列为高隐私敏感度的 C3 级个人信息 “ 用于用户鉴别的个人生物识别信息 ” 区别开来。这是声纹特征首次作为较低隐私敏感因子, 从生物特征识别因子中独立出来, 对行业发展具有重要指导意义。

2022 年 11 月, 国家网信办、工信部、公安部联合颁布《互联网信息服务深度合成管理规定》, 明确要求 “ 深度合成服务提供者和技术支持者提供人脸、人声等生物识别信息编辑功能的, 应当提示深度合成服务使用者依法告知被编辑的个人, 并取得其单独同意 ”, 并 “ 依法自行或者委托专业机构开展安全评估 ”, 此外还 “ 应当在生成或者编辑的信息内容的合理位置、区域进行显著标识, 向公众提示深度合成情况 ” 等。其中就包括 “ 合成人声、仿声等语音生成或者显著改变个人身份特征的编辑服务 ”。

1.2 行为特征识别技术开始兴起

生物特征识别最大的共性是唯一性。人的生理特征都存在唯一性, 每个人都有独一无二的脸、指纹、虹膜等。由于每个人的生物特征具有与其他人不同的唯一性和在一定时期内不变的稳定性, 所以利用生物识别技术进行身份认定相对其他身份认证技术是安全且准确的。但也正是由于生理特征的不可撤销性, 生物特征信息一旦被泄露、大量的带有唯一性的生物特征数据被盗取, 基于生理特征的身份识别系统将彻底崩溃。这也是生理特征识别方式的真正 ‘ 痛点 ’。

近两年来, 随着疫情防控常态化, 基于人脸等静态生理特征的生物识别技术得到大量应用, 但也进一步引致了公众对于个人隐私泄露的担忧。因此, 仅提取较少人体特征数据即可完成身份

识别、并可体现用户认证意愿的生物识别技术日益受到重视。一方面, 国内主流生物识别厂商的研发路线开始从单纯追求快速、准确, 向更注重隐私、安全演进。另一方面, 区别于人脸等静态生理特征, 基于声纹、步态等行为特征的动态生物识别技术开始兴起。(见倪光南院士发起、田霞主编《2021 网信自主创新调研报告》)

1.3 新的场景需求源源不断激发

在埃森哲发布的《人工智能成熟之道: 从实践到实效》研究报告中, 针对来自 17 个行业的 250 家中国企业开展了AI 成熟度调研评估。结果显示, 大部分中国企业仍处于应用 AI 的试验阶段, 仍需加大规模化应用力度以推动企业持续转型和全面重塑。从行业分布来看, 高科技行业的 AI 成熟度较为领先, 银行业与公共服务业的 AI 成熟度也保持了较高起点, 但近几年的增速有所放缓。而保险与零售、汽车与工业、资源与能源、通讯与媒体等行业的 AI 应用成熟度则大幅提升。例如, 保险与零售行业借助 AI 进一步提升客户与员工体验, 工业企业也已在设计开发与生产制造的各个环节使用 AI 技术。总体来看, 虽然 AI 在不同行业的应用重点和成熟度存在明显差异, 但行业差距正在不断缩小。

上述现象在声纹应用领域同样开始显现。在金融行业的示范带动下,第一波应用效应开始“外溢”, 新的行业对声纹应用的需求正在源源不断被激发出来。从创新的需要, 到安全的需要, 再到提质、降本、增效, 成为企事业单位应用声纹技术的新的驱动力。

1.4 无障碍环境建设立法实施

今年 9 月 1 日,《中华人民共和国无障碍环境建设法》正式实施, 该法律对无障碍信息交流、无障碍社会服务等提出了明确要求,进一步坚定了社会各界推动无障碍环境建设的方向和信心。目前, 我国现有残疾人约 8500 多万, 截至 2021 年底 60 岁及以上的老年人已有 2.67 亿, 而且老年人口数量还将继续增长。加强无障碍环境建设, 消除公共设施、交通出行、信息交流、社会服务等各领域有形或者无形的障碍, 保障残疾人、老年人能够平等充分参与社会生产生活, 是保障残疾人、老年人权益, 促进我国人权事业发展的内在要求和重要体现。党中央、国务院也高度重视无障碍环境建设。2020 年 9 月, 习近平总书记在湖南考察并主持召开基层代表座谈会时明确指出, 无障碍设施建设问题, 是一个国家和社会文明的标志, 我们要高度重视。2021 年、2022 年习近平总书记在考察冬奥会和冬残奥会筹办工作时, 又对做好无障碍环境建设提出了明确要求。

信息无障碍日益成为全社会共识, 针对老年人与残障群体的无信息障碍建设对语音技术需求迫切。2020 年 11 月, 国务院办公厅印发《关于切实解决老年人运用智能技术困难的实施方案》, 对金融服务提出要打造大字版、语音版、民族语言版、简洁版等适老手机银行 APP, 提升手机银行产品易用性和安全性。2021 年 3 月,人民银行印发《移动金融客户端应用软件无障碍服务建设方案》, 从用户视图、开发设计、使用辅助、运营保障等方面提出移动金融 APP 无障碍服务建设的具体要求, 如在用户视图方面应结合用户特殊需求提供“关怀模式”“语音模式”“民族语言模式”等便捷模式。

二、技术篇

2.1 声纹识别算法研究进展

近些年来, 声纹识别的算法模型发展迅速, 经历了从 i-vector 到 d-vector 再到 x-vector 等一系列发展过程。传统的 i-vector 模型并不区分说话人空间和通道空间, 而是将这两个空间合并起来形成一个总体变化空间, 采用类似于主成分分析的因子分析方法, 使用 T 矩阵将高维的高斯超向量进行降维并提取出能代表说话人信息的低维总变化因子（i-vector）, 然后在低维的 i-vector 空间里应用线性判别模型来进行通道补偿, 进而分离说话人信息和通道信息。可见, 传统 i-vector 模型的本质就是一种线性降维模型。

首先需要做的就是将原始录音文件转换成声学特征, 一般选择传统的声学特征梅尔频谱倒谱系数（MFCC）, 然后使用高斯混合模型（GMM）建模, 因为从目标用户那里收集大量的录音数据非常困难, 我们使用大量的非目标用户数据来训练一个 GMM, 这个 GMM 可以看作是对语音的表征, 但是又由于它是从大量身份的混杂数据中训练而成, 它又不具备表征具体身份的能力。因此我们需要基于目标用户的数据在这个混合 GMM 上进行参数的微调即可实现目标用户参数的估计, 学术界基于此目标提出了GMM-UBM 即通用背景模型。在这里得到的 i-vector 同时包含说话人 speaker 和信道 channel 的信息, 可以使用线性区分分析（LDA）来减弱 channel 的影响。当 i-vector 利用线性变换进行降维时, 难于保留原始数据中的非线性特征。因此, 科学界们想研究一种更好的非线性变换方法来将高维的高斯超向量降维得到说话人的低维总变化因子。由此基于深度学习的 d-vector 和 x-vector 应运而生。

近年来, 深度学习在语音识别领域中的成功应用鼓励着研究者将它运用到声纹识别中去。2015 年以前的声纹识别论文中几乎看不到深度神经网络（DNN）的存在, 如果涉及 DNN, 也只是用于代替 i-vector框架中的GMM 模型去计算统计量或是提取瓶颈层特征等诸如此类的非本质性、非变革性的工作, 直到 Google 提出了d-vector 的概念和框架。主要分为训练、注册和验证三个阶段。将上述阶段整合得到了一个简单且全新的架构 —— 端到端架构（End-To-End）, 在此架构中, 我们用 DNN 来建模作为说话人语音的特征表示, 并使用这个同样的 DNN 来做注册和验证工作, 这套架构的最终得到接受和拒绝损失 loss 来训练整个网络。

x-vector 是当前声纹识别领域最主流的基线模型框架, 一开始基于时延神经网络（TDNN）, TDNN 来自 1989 年的论文。原理假设整个结构简化为输入层、2 个隐藏层、输出层为识别出 B\D\G 的音素。TDNN 对于语音的特征学习是非常适合的, 因为语音与图像不同, 天然就具备时序特点, TDNN 网络随着时间的推进, 可以从不同时序的语音数据中抓取更多与说话人身份密切相关的独有的表征。TDNN 结构的优势在于, 其相对于长短期记忆人工神经网络可以并行化训练, 又相对于 CNN\DNN 增加了时序上下文信息。这在日常生活中我们可以体会, 当我们接到一个陌生来电时, 如果对方只说一句 “ 喂 ”, 我们可能愣住, 一时猜不出是哪位朋友的声音。但当对方继续说了几句之后, 我们的大脑快速反应, 很容易辨识出来是哪个熟悉的声音。

2.2 音频防伪算法研究进展

近年来, 抗攻击鉴伪算法和系统成为了业界关注的焦点。我们可以大致将系统的核心模块分为两个阶段, 当一段语音送入系统后, 首先进行声学特征提取, 然后送入到神经网络模型进行判定。在前端的特征提取的方法, 业界采用了很多尝试和实验, 算法包括恒 Q 倒频系数 （CQCC）、线性频率倒谱系数(LFCC)、逆梅尔倒谱系数(IMFCC)、快速傅里叶变换(FFT)、常数 Q 变换(CQT)、离散余弦变换 (DCT) 、自适应滤波 (LMS) 、Log-CQT 等。

而后端的模型, 在 ASVspoofing 的主办方的基线模型采用了经典的 GMM, 在 ASVspoofing2019 宣布结果后, 很多企业、高校在 Interspeech2019 国际会议上将自己的实现方案公布出来。从评价抗攻击系统的性能指标 t-DCF 和 EER 两项综合来看, 特征提取使用 CQT+ 网络模型 LightCNN, 得到的结果在 TTS/VC 攻击和 replay 攻击上都达到很好的效果。这里提到的 LightCNN 是在经典神经网络 CNN 的基础上做了改造, 加入 MFM (Max-Feature-Map) 激活函数。从国际竞赛结果中有一些有趣且有价值的关键性发现值得思考:

首先, 在这个 “ 盾 ” 的系统里, 前端特征提取算法的选择显得至关重要, 不同的特征提取对系统的性能影响差异非常大, 可以推测有可能是伪造的语音信息包含特殊的一类特征, 对于这类特征, 某种特征提取算法不一定能很有效地抽取出来, 或者只能抽取少量的特征, 而另外一些特征提取算法类似于 CQT 却能将伪造信息最大限度地提取, 所以影响了结果 , 也即伪造信息对特征提取的算法相对敏感, 但对后端模型不一定如此敏感。CQT 可以被视为一组有着对数间隔的滤波器, 它和小波变换类似, 具有可变的时间和频率分辨率, 相较传统的离散傅里叶变换 (DFT) 而言, 能提供更佳的信号分辨能力, 在 ASVspoof 2015 的合成语音检测任务中表现出优秀的检测性能。在特征提取的特征中, 研究发现 “ 功率谱 ” 蕴含有非常有价值的信息, 频谱可以用多种算法进行提取, 如 CQT、FFT、DCT 抽出。另外值得思考的是, 可否将不同的特征提取特征进行串联? 作为后端模型的多通道输入?

其次,在一些实验研究中,增加静音检测 VAD 后,系统的等错误率(EER)性能结果反而变差了,这也是一个有趣的现象。很有可能的原因是在静音的时间段内反而包含了语音伪造的关键信息。增加静音检测 VAD, 反而把有用的信息抹杀掉, 从而导致系统的抗攻击性能的下降。

再次, 基于 ASVspoofing 仿真的物理攻击 (PA) 录音重放的语音数据训练出的系统的有效性值得质疑, 在俄罗斯语音技术中心 (Speech Technology Center, STC) 发表的研究中提出针对仿真模拟的 PA 语音训练的抗攻击系统并不能有效检测真实的电话信道语音的环境中的攻击语音。

最后,攻击的种类繁多,从大类来讲也有逻辑攻击(LA)和 PA,可以为 LA 和 PA 设计不同的方案,这在学术研究中是可行的, 但在工程实践中, 在防伪语音检测的系统可能和生产系统进行联动, 针对 LA 和 PA 可能使用不同的前后端算法和参数配置, 但商业趋势和技术趋势是用一种网络结构去解决两大类攻击类型。更加前瞻的展望是, 可以利用多任务学习 (MTL) 的技术, 可以将防攻击的任务集成到主任务中, 例如通过一套模型输出多个结果, 就可以同时进行说话人识别和语音防伪检测, 或者同时进行语音转写和语音鉴伪, 而避免了串联或并联多个子系统, 这种基于多任务学习的深度集成的方案可以预见是未来重要的研究方向和商业趋势。

生成式人工智能的发展, 使得伪造语音往往更加贴近真实说话人语音, 且更新换代快, 对使用语音的相关系统造成威胁, 以伪造语音进行欺诈的事件也频频发生。采用传统方法进行检测, 对多样化的伪造语音泛化能力低、可解释性弱、更新扩展性差, 离实际需求差距很大。

2024 年 1 月, 由清华大学与得意音通联合研发的基于类脑感知和决策的伪造语音检测方法正式获得国家专利授权 (ZL 2023 1 1379225.8) 。该项专利技术针对语音相关技术中的伪造语音检测算法过于依赖数据、缺少对多样化伪造语音的泛化性、检测结果缺少可解释性等问题, 充分应用第三代人工智能 “ 知识引导与数据驱动 ” 相结合的策略, 提高了检测方法的普适性、泛化性、可解释性和可扩展性等。

2.3 工程化难点及技术进展

2.3.1 基于电话信道、实时音频流的声纹识别

首先, 当基于麦克风的文本相关或文本提示声纹识别较为成熟后, 随后就是基于电话信道、实时音频流的声纹识别技术需要进一步攻关。基于电话信道的声纹识别目前还面临着许多挑战如:

1)噪声和采样率影响: 电话信道噪声及环境噪声的叠加,电话采样率较专业收音设备采样率低, 多以 8kHz 为主, 同时由于电话信道多为对话语音, 角色分离的准确率不高, 这几方面因素都对声纹识别准确率造成影响。

2) 实时流处理难度高: 电话信道的声纹识别使用场景大多数为实时对话, 需处理实时流, 需要从核心网设备或呼叫中心服务器同步语音流, 并与元数据对应, 实施难度大。

3) 被动采集涉及隐私保护问题: 基于电话信道的声纹识别可实现无感知注册及验证, 但会涉及隐私保护问题。此外, 被动采集声纹信息, 音频质量不可控也是难点。

4) 跨信道训练与预测: 由于基于电信信道中文的大数据集的缺乏, 模型的训练可能基于非电话信道数据, 而模型的预测为电话信道数据, 导致精度的下降。

2.3.2 提升超大规模声纹辨认性能

近些年, 公安部正在规划将声纹识别技术纳入公共安全防治举措的方案, 并开展声纹采集设备选型。各地公共安全领域相关部门也在加大声纹采集力度。与此同时, 声纹数据库建设工作和建库规范也开始提上日程。但在类似公安的声纹库场景里, 就属于典型的 1:N 声纹辨认, 实验和实际项目应用中都发现, 当 N 呈现上升时, EER 和 topN 准确率、搜索效率、预测结果的响应速度等性能都会急剧下降。因此在 N 的数量达到万人规模时, 是不是声纹识别的性能就几乎是不可用的? 能否达到商用级别的要求? 但像公共安全、金融保险的行业, 声纹库的规模一定是在百万级或以上的。这目前是业界需要解决的难点。

网络信息显示: 某厂商声纹比对系统提供 1:N 大库检索比对, 支持千万库容建设。在使用的

实际数据测试的实验中, 用 159449 句语音, 与 12782 个说话人进行约 10 亿次比较, 154027 条语音对应的实际说话人直接命中 top1, 也就是说 top1 的直接命中率为 96.6%。根据该性能表现, 在较好的测试数据集下（声音噪音、信道、有效语音时长、采样率比较理想的情况下）, 一万人的平均返回比中排名约为 1.5 位, 十万人返回排名约为 9.7 位。

2.3.3 多模态多任务联合识别

2020 年初春, 一场新型冠状病毒肺炎疫情如同一个黑天鹅效应, 让整个世界范围、各个行业措手不及。疫情来临后, 人脸识别的很多场景都遇到了麻烦, 用户佩戴口罩, 只露出眼眉等部位, 人脸识别系统就无法识别出来。如果每一个通过闸机的用户都得摘掉口罩才能识别, 又增加了感染的风险。疫情之下, 促使了非接触、多模态技术的蓬勃发展。单个识别技术如人脸识别对于光照强弱、口罩遮挡、表情变化、尺度变化、设备采集角度等常见问题有局限性, 精度无法达到某些场景下商业要求。且人脸识别广泛应用后, 个人隐私数据被各类系统广泛采集, 仅凭单一识别技术存在漏洞和安全风险, 特别是涉及金融支付、用户认证等。

疫情影响下, 在电梯、门禁、闸机、取款设备等多种场景下都提出了非接触需求, 多模态技术融合后的产品形态将会明显提升用户使用体验, 也就是在这些场景下, 声纹＋人脸等联合进行识别成为重要趋势。

2.3.4 多说话人分离

在有多个说话人的场景, 如何运用人工智能技术把不同说话人甄别并分类出来, 此类需求一般被叫做“话者分离”或者“说话人分离”, 所用到的核心技术, 学术界一般称为多说话人分割聚类或说话人日志（Speaker Diarisation）, 该技术包含两个过程, 说话人分割（Speaker Segmentation）和说话人聚类（Speaker Clustering）。

说话人分割: 获取一段音频的语音信息流,检测出哪些帧是有语音的,哪些帧是没有语音的(无声音或只有背景声）; 接着检测出说话人的变更点, 然后根据变更点把音频进行分割。由于说话人变更点检测不准, 目前更多的是采用固定窗长（如 1.5 秒）和固定窗移（如 0.75 秒）分割。

说话人聚类: 对分割后的每一段语音,提取特征,进行聚类,得到每一段的语音属于哪个人说的。也可以理解为声纹识别的 N:N 分支。

多说话人分割聚类技术目前在国内发展方向比较多但杂, 有多种不同的技术路线, 包括模块化和端到端方案, 可应用在电话信道、现场环境与网络信道, 但根据专利表格中国内已有的专利去相关企业官网搜索, 相应技术的落地应用场景少于专利数量且不够成熟。

目前多说话人分割聚类技术还存在的挑战有:

1. 人数限制: 无论是说话人分割还是聚类, 在实际应用上都是有人数限制的, 目前 1-2 人的分

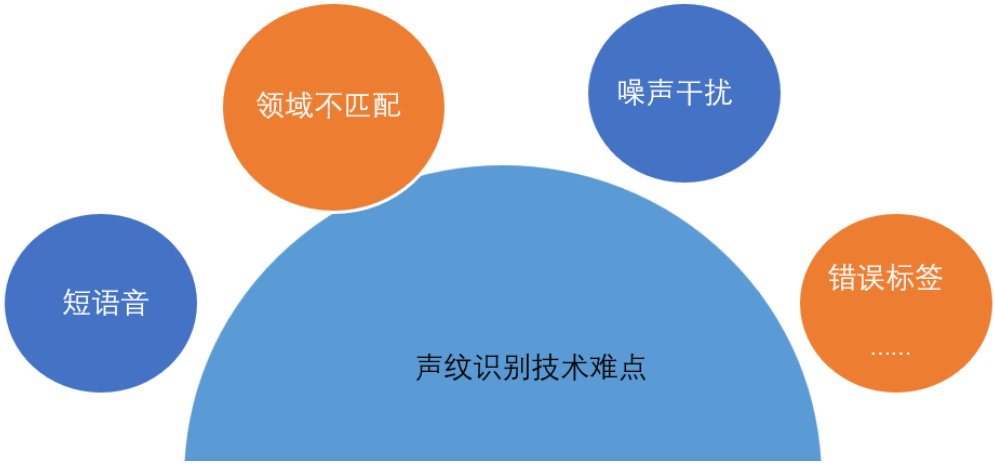
割聚类相对成熟, 但是 3 人以上的分离聚类并不成熟, 会导致准确率下降。

2. 人声重叠: 在同一音频中若出现人声重叠情况, 即两人同时说话的现象, 会导致音频不能准确分离, 片段分割不准确, 技术性能大幅下降。

3. 建模方法: 说话人分割聚类的建模方法多样化, 哪种方式效率最高, 结果最稳定且能达到最优效果, 仍需实验和摸索。

2.4 前沿挑战及技术进展

近几年来, 声纹识别技术的发展迅速, 在工业界的应用不断增多, 包括智能手机、智能音箱、智能电视、智能客服等场景, 实现 1:1 确认或 1:N 辨认任务, 为用户提供身份认证或个性化的服务。在公安刑侦等领域, 大规模声纹检索系统也进入实战应用, 对声纹识别的精度和响应速度提出更高要求, 测试指标包括 Top-N 命中率和百 / 千万级声纹库比对时间 (秒级响应)。



虽然声纹识别在受限场景已取得比较好的效果, 但在实际应用仍面临着噪声干扰、短语音、错误标签、领域不匹配等诸多挑战, 如上图所示。尤其是完全放开的领域, 等错误率 (EER) 还很高, 如 CN-Celeb 测试集高达 10% 以上, 还很难实用。

针对挑战问题, 很多研究者提出了针对性的改进方案, 包括 On-the-fly 扩增 (即在线扩增)、文本到语音 (TTS) 扩增、迁移学习、ECAPA-TDNN、SE-Block、带 Margin 的损失函数、多任务学习、带标签修正的概率线性区分分析 (PLDA) 等方法。

下表对近几年前沿技术做了梳理和汇总:

问题分类	前沿技术
数据扩增	* 加噪扩增 (MUSAN, RIR) 00, 极大提升性能, 但对未知噪声仍较差。 * 变说话方式]0 *On-the-fly 扩增 0 *Voice Conversion (VC) 扩增 0 *TTS 扩增]00
抗噪鲁棒性 (前后端优化、对抗学习)	* 后端分类器: * SNR-invariant PLDA0 * SNR-dependent mixture of PLDA 0 * 语音增强 (+ 说话人识别): * DNN-based binary masking0 * VoiceID Loss(前后端联合优化)0 * 语音分离 (BNF)+ 联合训练 0 * 对抗学习: * 文本无关 (厦大)0 * 文本相关 (微软)0 * 差异对齐 (成对输入): * invariant representation learning(IRL)00 * within-sample variability-invariant loss0
网络改进	* ECAPA-TDNN0: * ResNet 残差设计 * 跨层聚合 * SE-Block 通道增强 * Res2Net0

问题分类	前沿技术
损失函数 (Margin Loss)	* AM-Softmax0 * AAM-Softmax0 * AM-Centroid Loss0: 缩小类内间距, 拉大类间中心距离 * DAM-Softmax 0: 动态 margin * Angular Prototypical (AP) loss0
新的 Embedding	* xi-vector0: 结合 x-vector 和 i-vector 的优点, 提升泛化性能
多任务学习 (利用音素信息)	* 段级别说话人识别 + 帧级别音素识别 0 * 梯度反转层 (Gradient Reversal Layer, GRL) 0 * 段级别: 抑制音素信息 * 对于无法对齐数据, 音素分类可采用伪标签 0
领域自适应 0	*** 无监督方法:** * CORAL * Feature-based domain adaptation * Model-based domain adaptation * CORAL+ **有监督方法:** * 监督线性差值法 (LIP) * 相关对齐差值法 (CIP) * 基于 CORAL 的监督参数调整法 * 基于 CORAL+ 的监督参数调整法

问题分类	前沿技术
损失函数 (Margin Loss)	* AM-Softmax0 * AAM-Softmax0 * AM-Centroid Loss0: 缩小类内间距, 拉大类间中心距离 * DAM-Softmax 0: 动态 margin * Angular Prototypical (AP) loss0

2.5 研究型数据集建设

2.5.1 声纹数据集建设过程

说话人声纹数据集的建设包括以下五项:

第一: 数据采集

声纹数据的采集需要收集大量的声音样本, 样本应涵盖不同的人员、年龄、性别和口音。除说话人本身的因素, 还需要考虑声学环境, 在复杂的声纹模型应用场景中采集真实的数据需要对环境做声学分析。

采集声纹数据的信息需要采用特定设备, 包括专业的声纹采集器、麦克风阵列等, 设备需要具备高灵敏度、低噪声、抗干扰能力强等特点, 以确保采集的声纹信息质量可靠, 满足模型测试指标要求的采样率、音频信道等要求。

第二: 数据标注

针对采集的音频数据进行标注处理, 需依据不同的指标要求规范做数据标注。以数据标注平台实现标注任务管理,以人工+机器辅助模式完成高质量数据标注。为每个语音样本进行标注, 并与相应的说话者身份进行关联, 这样可以在后续应用中识别和验证说话者身份。通常, 首先将语音样本分割为较短的片段, 以确保每个片段都包含清晰的发音和声纹特征; 然后, 建立适当的数据结构和数据库来方便检索、使用、组织和存储数据; 最后是对构建的声纹特征数据集进行验证和质量控制, 确保数据的准确性和可靠性, 这可以通过人工审查或自动化的方式来实现。

第三: 数据安全管理

即建设声纹数据集管理平台来进行科学有效的数据安全管理, 保证数据的安全性、模型应用的稳定性。数据管理平台可有效提高声纹模型应用的效率。针对不同模型的应用目标, 数据管理平台提供数据分析、数据整理、数据挖掘等能力。对声纹数据依据不同的应用指标进行分类, 快速迭代声纹模型的优化和测试能力。

第四: 数据质量管理

因声纹数据需满足声纹识别技术应用的匹配性和高合格率, 声纹数据的质量管理是构建准确和可靠声纹模型的重要步骤。针对数据的质量管理有如下几种方式:

数据收集规范: 在收集声纹数据时, 制定明确的规范和准则, 包括录制环境、录制设备、录制指导等, 以确保采集的数据质量和一致性。

数据预处理: 对声纹数据进行必要的预处理操作, 如去除噪声、时长归一化、数据增强等, 以提高数据的质量和准确性。

数据标注与验证: 对采集的声纹数据进行标注, 可通过人工验证或自动化算法进行, 以确保标注正确性。

数据验证和筛选: 对采集的声纹数据进行验证和筛选, 剔除质量不佳或不符合标准的数据。使用质量标准和准则进行评估, 如语音清晰度、发音准确性指标等。

数据平衡和多样性: 确保声纹数据集有足够的样本覆盖不同说话者、性别、年龄、语言等变化, 以提高模型的泛化能力和鲁棒性。

数据存储和管理: 建立合适的数据存储和管理系统, 确保数据的安全性、完整性和易于检索。备份数据并定期检查数据质量。

数据监控和更新: 持续监控声纹数据的质量, 及时检测和修复可能存在的问题。定期更新数据集, 以反映新的语音变化和说话者的变化。

第五: 数据合规性

声纹数据涉及个人私密信息, 因此在处理声纹数据时需要严格遵守隐私和数据保护法规, 确保数据的隐私和合规性。在收集声纹数据时, 向说话人明确表达数据收集的目的, 确保参与者明确知晓数据用途, 并取得他们的合法和明确的同意。声纹数据信息特征上尽量降低数据的风险, 对声纹数据进行去标识化或匿名化处理, 以防止个人身份的识别。建立安全的数据存储系统, 并采取适当的物理和技术措施来保护声纹数据, 防止未经授权的访问、泄露或滥用。在共享或转移声纹数据时, 确保符合适用法律和隐私保护要求, 并与共享方或接收方签订合适的协议。

2.5.2 研究型声纹数据集建设现状

声纹识别的模型性能受到不同发音个体的器官生理特性、环境等影响, 使得声纹识别的性能发生改变。配合声纹识别的应用场景和性能要求, 声纹识别的数据集建设尤为重要。建设声纹数据集是一个复杂且持续的过程, 说话人生理体征的变化对声纹识别的性能也会产生影响, 为提高声纹识别的准确性, 数据集的多样性和代表性非常重要, 因此收集具有广泛覆盖性的样本非常关键。在声纹识别技术中, 建设大规模、高质量的声纹数据集可以为模型的训练提供更多的样

本, 有助于提高模型的性能和鲁棒性。

声纹数据集在行业内的建设和发展现状主要表现在以下几个方面:

第一: 数据规模不断扩大

随着人们对声纹识别技术的需求不断增加, 声纹数据集的规模也在不断扩大。这些大规模的声纹数据集能提供更丰富的语音特征和更准确的识别结果,有助于提高声纹识别技术的性能。

第二: 数据质量不断提高

为了提高声纹识别技术的准确性和可靠性, 声纹数据集的质量也在不断提高。这包括提高数据的清晰度、降低背景噪音、标准化语音采集设备等方面的工作, 以确保声纹数据集的质量可靠。

第三: 数据标注精度提升

为了使声纹识别技术更好地应用于实际场景中, 需要高精度的标注数据。因此, 声纹数据集的标注精度也在不断提升。这需要投入大量的人力物力进行数据标注和校对工作, 以确保标注数据的准确性和可靠性。

第四: 数据多样化发展

由于人们的语音特征在不同场景、不同条件下会有所不同, 因此声纹数据集也在向着多样化的方向发展。这包括采集不同年龄、不同性别、不同口音的语音数据, 模拟不同环境下的语音采集情况等, 以提高声纹识别技术的适应性和可靠性。

第五: 数据安全和隐私保护

随着人们对隐私问题的关注度不断提高, 声纹数据集的安全和隐私保护也成为了一个重要的研究方向。这需要采取有效的技术手段和管理措施, 确保声纹数据集的安全性和隐私保护能力。

第六: 标准化和公开化趋势

为了促进声纹识别技术的交流和发展, 声纹数据集的标准化和公开化趋势也越来越明显。标准化有利于不同研究机构之间的比较和评估, 公开化则可以促进技术的交流和进步。

2.5.3 常用的研究型声纹数据集

在研究和学术界, 声纹数据集的建设和发展是推动声纹识别技术进一步发展的重要基础。学术研究机构、大学和研究实验室积极建立声纹数据集以促进算法研究和技术创新。

以下是一些目前常用的学术研究型数据库信息:

VoxCeleb1/VoxCeleb2: 数据集最初由牛津大学计算机科学系创建, 并由康奈尔大学和谷歌 DeepMind 共同合作维护和更新。VoxCeleb1 是一个大型的开放式语音数据库, 包含来自名人和社交媒体的数千个说话者的语音片段。VoxCeleb2 增加了更多的说话者和语音片段, 以弥补原始数据库中的一些不足。

TIMIT: 数据集是由德州仪器、麻省理工学院和斯坦福研究院 (SRI International) 共同创建。TIMIT 是一个经典的英语语音数据库, 包含多个说话者的语音和文本。它被广泛用于语音识别和说话者识别的研究。

VCTK: 数据集由爱丁堡大学发起并开发, 是一种包含多种口音和语音变体的多说话者语音数据库。它主要用于多说话者语音合成和说话者识别领域的研究。

CN-Celeb: 数据集由清华大学语音和语言技术中心 (CSLT) 发布, 是一个中国知名人物的声纹数据库。它涵盖了来自社交媒体平台的数千个说话者的语音片段, 可用于研究多说话人和多模态声纹模型技术研究。

SITW (Speakers in the Wild) : 数据集由斯坦福研究所 (SRI International) 和布宜诺斯艾利斯大学计算机学院合作创建, 是一个包含来自真实世界环境的多说话者语音片段的数据库。它用于评估和比较不同系统在实际场景下的性能。

AISHELL-DMASH: 数据集由北京希尔贝壳科技 (AISHELL) 与昆山杜克大学合作创建, 是一个在真实家居场景下录制, 跨时间域并针对不同年龄的说话人进行多设备同时采集的声纹数据集。它用于研究家居场景下多设备的声纹识别技术研究。

声纹数据集的建设和发展正处于积极的发展阶段。随着声纹识别技术的广泛应用以及对数据质量和隐私保护的不断要求, 声纹数据集的建设和管理将继续受到关注和重视。然而, 也需要注意确保数据的合规性、隐私保护和安全性, 以满足相关法规和标准的要求。

2.6 相关赛事综述

2.6.1 CNSRC 2022

CNSRC 2022全称为CN-Celeb Speaker Recognition Challenge 2022中国明星声纹识别挑战赛。¬¬ 该竞赛是由 Odyssey 2022 组委会发起, 由清华大学、厦门大学和北京希尔贝壳科技有限公司联合承办的说话人识别竞赛。赛事时间在 2022 年 2 月至 6 月。赛事研讨会在声纹领域国际会议 Odyssey 2022 中召开。竞赛的核心目的是验证当前说话人识别技术在实际复杂场景下的真实可用性, 并甄选出适应于实际应用场景的有效算法。

该竞赛共设定了两个任务: 说话人确认 (Speaker Verification) 和说话人检出 (Speaker Retrieval) 。前者验证测试语音是否属于某一声称说话人, 后者从 50 万背景语音中检出目标说话人的 10 句发音。每个任务依据训练数据不同, 又分为固定赛道 (Fixed Track) 和开放赛道 (Open Track) 。前者仅允许使用 CN-Celeb 作为训练集, 目的是验证算法先进性; 后者可利用任何数据

进行训练, 目的是验证当前技术所能达到的性能上界。

该竞赛吸引了 132 支海内外队伍参赛, 来自上海交通大学、国音智能、北京理工大学、腾讯、山西大学等单位的参赛队伍取得佳绩。

赛事官方: http://cnceleb.org/competition

赛事研讨会: http://cnceleb.org/workshop

2.6.2 VoxSRC 2022

VoxSRC 2022 全称 VoxCeleb Speaker Recognition Challenge 2022 大规模说话人识别挑战赛。该竞赛由牛津大学、谷歌、韩国科学技术院、亚马逊科技、以及韩国 Naver 公司联合组织。赛事时间在 2022 年 7 月至 9 月。赛事研讨会在语音领域国际会议 INTERSPEECH 2022 中召开。该竞赛旨在研究现有的说话人识别方法在 “in the wild” 场景下的识别性能。竞赛所提供的官方数据集 VoxCeleb1 和 VoxCeleb2 来自 YouTube 的名人采访、新闻节目、脱口秀和辩论等场景。

该竞赛共设定了四个赛道。前两个赛道统称为有监督的说话人识别（Fully Supervised Speaker Verification）, 不同之处在于: 第一个赛道仅允许使用 VoxCeleb2 数据集作为训练集, 而第二个赛道训练数据集是开放的、不受限的。第三个赛道是半监督领域自适应（Semi-Supervised Domain Adaptation）; 它是 2022 年新引入的一个赛道, 赛事数据包括一个大规模带标签的源领域数据集、一个大规模无标签的目标领域数据集和一个少量带有标签的目标领域数据集, 模拟实际应用的领域自适应问题。第四个赛道是说话人日志, 使用 VoxConverse 评测集, 探索多人对话场景下的说话人日志技术。

该竞赛吸引了上百只支海内外队伍参赛,来自美国 ID R&D 实验室、快商通、上海交通大学、微软、中科院、昆山杜克大学、腾讯、杜克大学、韩国 GIST 等单位的参赛队伍取得佳绩。

赛事官网: http://mm.kaist.ac.kr/datasets/voxceleb/voxsrc

赛事研讨会: https://www.robots.ox.ac.uk/~vgg/data/voxceleb/interspeech2022.html

2.6.3 FFSVC 2022

FFSVC 2022 全称 Far-Field Speaker Verification Challenge 2022 远场说话人验证挑战赛。该竞赛由昆山杜克大学、南加州大学、新加坡国立大学和北京希尔贝壳科技有限公司联合组织。赛事时间在 2022 年 4 月至 7 月。赛事研讨会在语音领域国际会议 INTERSPEECH 2022 中召开。该竞赛聚焦在远场单通道场景下的说话人识别任务。该赛事所提供的官方数据集 FFSVC2020 是 AISHELL-DMASH（中文普通话麦克风阵列家居场景语音数据库）的一部分。该数据集中包含不同距离、不同设备的录音, 特别针对于远场说话人识别任务。

该竞赛共设定了两个赛道。第一个赛道是有监督的远场说话人识别任务, 参赛者需使用带有

说话人标签的 FFSVC2020 和 VoxCeleb1&2 数据集来构建远场说话人识别系统。第二个赛道是半监督远场说话人识别任务, 是 2022 年新设定的赛道, 参赛者需使用带有说话人标签的领域外 VoxCeleb1&2 数据库和不带说话人标签的领域内 FFSVC20 数据集搭建远场说话人识别系统, 来模拟更贴合实际的数据条件。

该竞赛吸引了多支海内外队伍参赛, 来自国音智能、西北工业大学、华为、中兴等单位的参赛队伍取得佳绩。

赛事官网: https://ffsvc.github.io/

赛事研讨会: https://ffsvc.github.io/workshop/

2.6.4 SASV 2022

SASV 2022 全称 Spoofing-Aware Speaker Verification 2022 伪造语音感知的说话人识别挑战赛。该竞赛由韩国的市立汉城大学、延世大学、Naver 公司、法国 EURECOM 研究院、东芬兰大学等多个研究机构共同组织。赛事时间在 2022 年 1 月至 3 月。赛事研讨会在语音领域国际会议 INTERSPEECH 2022 中召开。该赛事旨在评测说话人识别系统和伪造语音检测系统的集成技术方法, 进一步提升说话人识别系统应对闯入攻击的鲁棒性。该赛事所提供的数据集是 VoxCeleb 和 ASVspoof 2019。ASVspoof 2019 数据集特别针对于伪造语音感知的说话人识别任务, 包含了伪造语音、真实语音及其标签, 以及相应的说话人标签。

该竞赛是 2022 年首届举办, 共设定一个赛道, 其评价标准是说话人识别的准确率和假冒闯入攻击语音的检出率,目的是鼓励研究者构建集成说话人识别和伪造音检测两项技术的联合系统。

该竞赛吸引了 23 支海内外队伍参赛。来自美国 ID R&D 实验室、昆山杜克大学联合 OPPO 团队、韩国汉阳大学、香港中文大学、台湾国立大学、三星等单位的参赛队伍取得佳绩。

赛事官网: https://sasv-challenge.github.io

赛事结果: https://sasv-challenge.github.io/challenge_results

2.6.5 CSSD 2022

CSSD 2022 全称 Conversational Short-phrase Speaker Diarization Challenge 2022 对话短语音说话人日志挑战。该竞赛由中国科学院声学研究所、西北工业大学、新加坡 A*STAR 信息通信研究所、上海交通大学以及 Magic Data 联合主办。赛事时间在 2022 年 7 月至 9 月。赛事研讨会在语音领域国际会议 ISCSLP 2022 中召开。该赛事聚焦在对话短语音场景下的说话人日志技术。该赛事所提供的数据集是由 MagicData-RAMC 开源的中文对话语音数据集。

该竞赛设定了一个赛道, 使用 Conversational-DER (CDER) 评估指标来度量句子级的说话人日志精度, 来验证当前主流技术在对话短语音场景下的说话人日志性能。

该竞赛吸引了多支队伍参赛, 来自上海交通大学 - 思必驰联合实验室、上海声通信息科技、西北工业大学联合传音控股、亚信科技、浙江大学等单位的参赛队伍取得佳绩。

赛事官网: https://magichub.com/competition/sec-competition/

赛事研讨会: http://www.iscslp2022.org/

三、场景篇

3.1 从技术到场景

声纹技术应用在金融行业的示范带动下，第一波应用效应开始“外溢”，新的行业对声纹应用的需求正在源源不断被激发出来。从创新的需要，到安全的需要，再到提质、降本、增效，成为企事业单位应用声纹技术的新的驱动力。

作为生物识别技术的一种，声纹识别技术在应用分类上同样分为声纹确认（1:1 认证）和声纹辨认（1:N 搜索）两大类，分别对应不同的应用场景，详见下表。在声纹确认模式下，技术主要用于安全访问验证，确保只有声纹信息匹配的个体才能进行某些敏感操作或访问特定资源。这为金融交易、电子支付等场景提供了强大的安全保障。而在声纹辨认模式下，技术主要用于快速搜索和识别大量声纹数据库中的特定声纹。例如，语音助手能够通过声纹辨认迅速识别并响应用户指令，大大提升了交互体验。

表 1 声纹识别技术应用分类

应用分类	分类说明	适用场景
声纹确认 （1:1 认证）	给定一段只含一名说话人的语音和一个说话人的声纹模型，判断该段语音是否是该说话人所说的声纹识别方式。	已知用户身份认证等
声纹辨认 （1:N 搜索）	给定一段语音和一组候选说话人的声纹模型，判断该段语音是哪位说话人所说的声纹识别方式。	黑名单检测、语音助手等

根据声音采集信道的不同，声纹识别技术按照信道类型可分为电话信道与网络信道两种，详见下表。

表 2 声纹识别技术信道类型

信道类型	技术说明	渠道举例
电话信道	声音传输的媒介是电话，声音采样率为 8k。	电话银行
网络信道	声音传入的媒介是网络，声音采样率一般不小于 16kHz。	手机银行等各类 APP

而在具体采集模式上，又可以分为固定文本、动态数字和自由说三种，详情见下表。

表 3 声纹识别技术应用模式

应用模式	技术说明	技术特点
固定文本	声纹注册时需用户录入指定内容的声纹，在声纹认证时需用户说出与注册时完全相同的内容。语音内容一般为 10 个字以内。	同等录音有效时长情况下声纹比对准确率最高，但容易受录音攻击。
动态数字	声纹注册时需用户录入指定内容的声纹，在声纹认证时需用户说出与注册时相同的内容，但顺序可以变化。例如注册时用户录入 0-9 的数字，认证时通过用户说出系统给出的 8 位随机数字来进行认证。	同等录音有效时长情况下声纹比对准确率居中，通过验证时使用随机数字方式，防攻击效果好。
自由说	用户可以通过任意的内容的声纹来进行声纹注册和认证。自由说对用户的有效语音长度有要求，一般需要 10-20 秒。	同等录音有效时长情况下声纹比对准确率最低，需通过录音时长进行弥补。对语音内容没有要求，适用于无感识别场景。

从上述各种分类中可以看出，不同的业务场景使用需求会对应不同的应用模式，声纹识别与人脸识别等其它生物特征识别技术相比，原子服务种类相对较多。应用单位在实际技术开发应用时，需综合评估技术成熟度和行内应用需求情况等因素，选择合适的推进路线，以满足各渠道不同场景差异化适配的使用需求。

与传统静态密码 100% 精确比较后判断是否匹配成功的方式不同，声纹识别技术是通过判断声纹的相似度来决定匹配结果，其本质上属于近似计算而不是精确计算。也就是说，声纹识别的相似度阈值设置，实际是在安全性与易用性之间进行取舍，寻找合适的平衡点。例如，若调高相似度阈值，表现为安全性更高，但容易把应该匹配上的说话人误判为匹配失败，易用性变差。若调低相似度阈值，表现为易用性较好，但容易把不应该匹配上的说话人误判为匹配成功，降低了系统安全性。此外，受噪音干扰、声音时变、跨信道差异等因素影响，声纹识别技术的准确率也会有所波动。

3.2 金融科技

金融科技创新给移动端身份认证带来了新的思路和途径，目前各种解决方案呈现出爆发式增长。在移动金融安全风险防范方面，相关监管政策也在逐步完善。在法律和行业规范日益严格、公众个人隐私被愈发重视，全球疫情下“非接触式服务”普及等诸多因素的叠加之下，如何在兼顾便捷和安全的同时进行精准客户身份识别、如何验证用户反映的是本人真实意愿和真实操作而非黑客或诈骗集团刻意“冒充”，同时符合监管合规要求，是目前业界亟需解决的问题。

商业银行作为金融科技创新的引领者，在声纹识别等新技术上提前布局，并在平衡技术安全性和易用性关系的基础上，持续推动声纹识别技术的建设和应用。

金融领域主要应用场景举例：

1) 登录、支付、转账、取款——采用声纹识别技术，自动匹配用户个人身份信息，完成登陆、支付、转账、取款的身份验证，一般采用文本相关 / 提示的方式，即 8 位随机动态数字串或者固定文本。

2) 业务核身——采用声纹识别技术，在业务沟通中完成用户身份核验，在自动匹配业务办理的信息，进行比对，完成业务办理的身份核验，一般采用文本无关方式，如开卡开户。

3) 信贷风控——采用声纹识别技术，在信审环节对用户身份进行识别，并查验是否为黑中介 (黑名单用户)，完成信审身份审核，采用文本无关的方式。

4) 金融反洗钱——采用声纹识别技术，在判定出疑似洗钱行为后对用户进行电话远程身份验证以及自动对用户信息核对，完成可疑用户身份核验，采用文本无关的方式。

5) 内部培教 OA 系统核身——金融机构内部系统安全登录、内部话术考核身份核身。

由于声音信号处理的复杂性，声纹识别在实际应用中具有较高的技术门槛，需要相当的技术积累。商业银行为满足其对产品功能完善、性能优异、成熟度高、快速上线的要求，大多通过采购业界成熟商用软件产品的方式来实现技术能力的快速构建。同时，为适应银行内部系统的架构

设计和服务能力要求，具备较强科技实力的银行一般采取“平台自建，算法集成”的建设策略。

已有部分银行业务场景通过应用声纹识别技术，增强用户身份认证，加强风险防控，简化业务流程，提升业务满意度。例如，工行在对客服务和风控领域开展声纹识别技术应用，对客服务应用在手机银行登录，风控领域则应用在电话银行信用卡申请反欺诈环节，是国内首家将声纹技术应用在反欺诈领域的银行；建行主要在对客服务领域开展应用，场景较多样，如手机银行 APP 登陆、转账等功能的辅助增强认证并全面推广，以及在 ATM 取款辅助认证场景进行了局部试点；招行则在手机银行利用声纹识别进行信用卡临时调额时的身份确认。

表 4 部分银行声纹识别技术应用场景

名称	场景说明	信道类型	应用模式
工商银行	电话银行信用卡申请使用声纹识别进行反欺诈应用	电话信道	自由说 1:N 搜索
建设银行	1. 手机银行使用声纹技术进行登录 2. 利用声纹技术在手机银行转账、ATM 二维码取款上进辅助身份认证	网络信道	动态数字 1:1 认证
招商银行	手机银行利用声纹识别进行信用卡临时调额时的身份确认	网络信道	动态数字 1:1 认证
浦发银行	手机银行使用声纹技术进行登录	网络信道	动态数字 1:1 认证
广发银行	1. 手机银行使用声纹技术进行登录	网络信道	动态数字 1:1 认证
贵阳银行	2. 利用声纹技术在手机银行进行本人本行同名转账	网络信道	动态数字 1:1 认证
西安银行	手机银行使用声纹技术进行登录	网络信道	动态数字 1:1 认证

声纹识别技术在上述银行业务场景的成功应用, 将逐步形成示范效应, 进一步扩大该技术在商业银行的场景应用。

目前, 银行主要应用了网络信道动态数字 1:1 认证方式, 电话信道自由说 1:N 搜索方式则只有个别银行成功应用。为此, 部分商业银行通过探索建立配套的工作机制, 从客户体验和算法适配等方面着手, 收集客户使用反馈和生产运营指标, 进而根据不同使用场景下的差异化需求, 灵活调整声纹识别的使用方式, 更好地平衡安全性和易用性之间的关系。同时, 针对高安全性场景的应用, 从业务控制上采用多因子认证的方式, 避免使用声纹作为单一的认证手段。例如, 建设银行等在手机银行业务场景中, 除了声纹识别外, 还会要求输入密码进行双重校验。

3.3 公共安全

随着互联网智能手机和智能软件的普及,涉及声纹鉴定与识别的语音案件数量也在不断增加。为了应对这一趋势,公安系统面临重大考验,需要遏制网络诈骗犯罪并维护人民群众的合法权益。在这个背景下, 声纹识别技术在反电信诈骗方面发挥着无可替代的作用, 成为公安应对新形势下声纹识别、鉴定以及布控需求的关键技术。

通过对诈骗分子声纹的研究, 可以精准定位海量电话数据场景下的诈骗通话, 快速发现有害诈骗信息。具体过程如下: 结合声纹识别、声纹聚类技术, 利用采集标注的诈骗人有害话音集合, 提取并存储此类人员的声纹特征, 建立声纹库, 新的通话接入时, 经过声纹提取及声纹比对, 可以实现目标通话中诈骗声纹的检出和发现, 如图所示:

因此, 公安领域声纹解决方案是一个全方位的一体化方案, 包括声纹采集设备、智能音频分析平台和云计算管理平台等。这个方案提供从声纹数据采集、存储分析到应用管理的一系列功能, 可应用于各种场景, 如室内半开放复杂声场环境下的高保真语音及声纹采集、声纹识别、多语言语音转录等。

智能音频分析平台是该解决方案的核心组成部分, 由声纹数据库、声纹识别引擎、语音识别引擎及语音转录引擎组成。声纹数据库集声纹数据管理、清洗、比对于一身, 是实现声纹识别、语音识别与转录功能的基础。高性能的声纹数据采集系统则是获取高质量声纹数据的关键, 对提高声纹识别的准确性起着至关重要的作用。现有的声纹数据采集系统能实现自适应降噪, 多通道采集及声源分离, 在复杂场景下达到高保真拾音。

云计算管理平台可以灵活地与公安声纹实战平台、声纹鉴定平台、司法审讯平台等对接。其中,

声纹实战平台以声纹识别系统为核心, 联合海量数据库, 针对公安领域深度优化, 提供声纹大数据检索核心功能。通过声纹比对, 该平台能有效锁定嫌疑人员, 广泛应用于重点人员监控、反电信诈骗、案件侦破、身份核验等场景。

基于音素检索技术、声纹识别技术及关键词检索技术的声纹鉴定平台, 结合公共安全及司法鉴定领域身份鉴定业务需求, 提供一套完整的软硬件一体的解决方案。该方案利用人工智能技术和专业的数字化频谱, 辅助声纹识别专家快速比对检材和样本的声纹信息, 实现对语音文件说话人的识别认定, 为声纹的实时识别和快速鉴定提供了可靠的技术基础。

在公安领域的应用中, 声纹识别技术可用于重点人员布控、侦查破案、反电信欺诈、治安防控、司法鉴定、审讯室建设、网络身份认证等多个方面。

公共安全主要应用场景举例:

1) 重点人员布控: 通过建立重点人员声纹数据库, 在特定情况下一旦发现重点人员或黑名单人员的声纹信息, 即进行预警, 有效预防事态发生。

2) 侦查破案: 利用声纹识别技术的海量筛查优势, 进行“案查人”、“人查案”、“案查案”与“人查人”等多种排查方式, 缩小侦查范围, 提高办案效率。公安领域要求声纹数据库的声纹比对系统能够提供 1:N 大库检索比对, 同时要支持千万库容建设。

3) 反电信诈骗: 利用声纹鉴定技术对电信诈骗等案件中的涉案语音进行个体、团伙的识别, 确定犯罪嫌疑人身份, 为侦查破案、案件诉讼提供技术支撑。

4) 治安防控: 利用“语种识别”、“内容识别”、“声纹特征识别”等声纹综合分析技术, 对重点人员进行布控, 一旦出现立即进行关注控制。

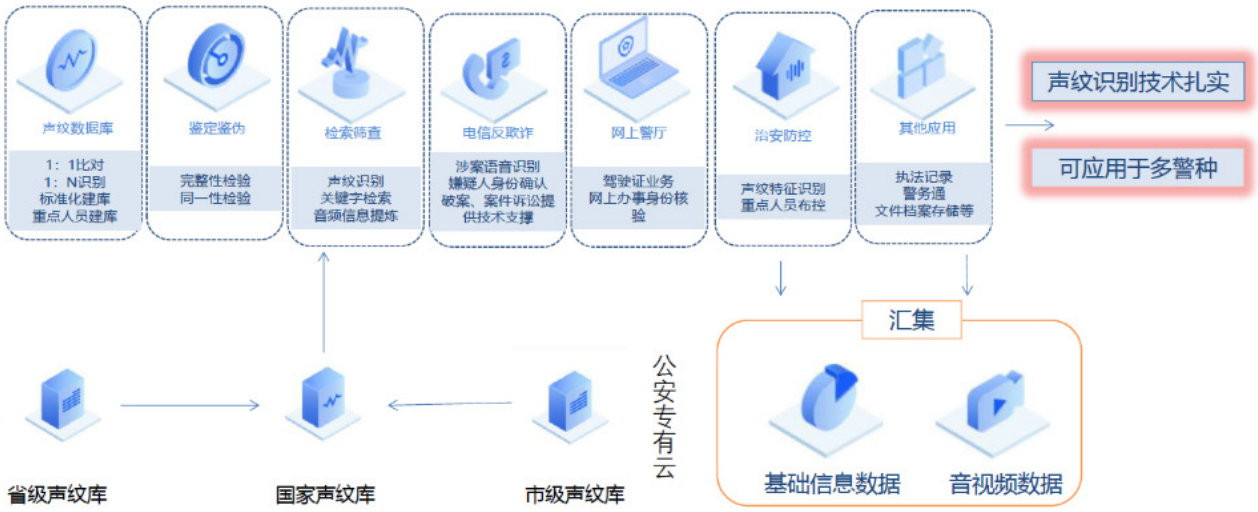
5) 身份认证: 在监狱亲情电话应用中, 通过采集犯人家属的声纹信息, 可有效鉴别家属身份的合法性。在司法社区矫正应用中, 通过识别定位手机位置和呼叫对象说话声音的个人特征, 系统就可以快速地自动判断被监控人是否在规定时间出现在规定场所, 有效解决人机分离问题。

不过总的来说, 基于上述场景的应用目前仍然处于探索阶段, 其效果有待进一步观察和验证。

3.4 政务民生

3.4.1 政务场景

目前大部分政府机构及企业核心应用仍然在私有数据中心或者内网部署, 由于承载的业务大多涉密度高, 因此必须对用户身份进行准确识别, 才能保证各级用户按照规定的权限存取数据, 不发生越权访问事件,并且要确保其在任何位置、任何时间访问网络时,都能准确标示其身份信息。随着政府企业移动化办公带来安全边界的消失, 身份认证对于企业安全重要性日益增强, 一旦身份认证系统被攻破, 系统的所有安全措施将形同虚设。



如某公司“声纹+”政企服务解决方案可为政府机构及企业提供智能可信的远程身份认证服务，适用于各种政企服务平台、线上办事大厅、内部 OA 系统、档案数据系统查询权限分级等场景。具有高活体检测、高认证强度、防假冒攻击、真实意愿等优势，可有效提升平台系统访防伪能力和反向追查能力。依托于清华大学语音和语言技术中心及得意音通信息技术研究院强大的技术支撑，得意音通亦可为地方省域搭建可信身份平台及声纹数据库基础设施，全力支持我国身份认证云建设。

政务领域主要应用场景举例：

1) 政务服务平台身份核验：利用声纹身份认证代替账户密码、指纹人脸等传统认证方式进行系统登录、业务办理等操作。

2) 数字档案系统权限分级：利用声纹身份认证按等级权限进行相应操作。

3) 门禁考勤设备身份核验：融合声纹及人脸两种生物识别指令，提高身份核验准确率。

3.4.2 民生场景

目前我国正在加速进入老龄化社会，社保金管理存在冒领现象，且现有解决冒领问题的方法成本高、效率低。行动不便或身在异地的老人领取养老金由原来的到场验证升级成为远程认证解决方案。看似是为广大老年群体提供了便利，但实际上基于人脸识别技术的远程社保认证解决方案所采用的全程无监督采集认证模式，这种做法强调体验而弱化了安全，如不能有效的防止假冒攻击，无疑是给冒领社保金埋下更深的隐患。

某公司“声纹+”社保解决方案，采用声纹+人脸多模态融合技术，解决了单一生物特征所具有的局限性,为地方社保机构提供安全可信的远程身份及生存核验。支持多种通信设备远程核验，有效防止国家养老金的冒领和流失，实现社保管理工作的提质增效，保障社保养老金精准发放。对医护人员、患者及患者家属等人员进行身份认证服务，满足医院日常办公以及开处方、诊疗、手术等业务中的人员身份认证需求，医护责任明确、患者医院身份真实，减少纠纷，业务可信。

社保领域主要应用场景举例:

1) 多设备远程生存认证: 用户使用电话、手机、PC 电脑等各类通信设备远程进行身份及生存状态核验。

2) 补贴发放身份核验: 适用于高龄、低保、困难残疾人员及重点残疾人员护理补贴等发放的身份核验 。

3) 实名问诊 / 医保购药: 病人实名建档、挂号、看病及取药报销及医生远程开具处方等环节的实名身份核验。

3.5 教育与医疗

3.5.1 教育考试场景

在教育领域,技术的进步正在不断地改变我们的教学方式和学生的学习体验。声纹识别技术,作为一种创新性的身份认证手段, 在提高教学质量、评估效果和个性化学习以及保障考试的公平性和安全性等方面为教育行业的发展注入更多创新智慧。

增强考试与出勤管理

随着我国教学改革的推进,招生规模和数量逐年扩大,考生人数递增,代考现象难以避免。目前,大多数考场仍采用身份证、准考证等传统认证方式, 人工比对存在主观性和不一致性, 认证结果难以把控。因此, 引入多样化智能识别比对系统是必要的, 结合声纹识别技术可有效防止学生逃课、替课, 解决老师耗时长、耽误上课时间的问题, 且声纹采集设备成本低, 可减轻学校支出。虽然高校采用智能身份认证方便了学生, 但也存在新问题, 如设备昂贵导致学生排队打卡耽误时间、指纹和人脸破解门槛低等。

1) 考生身份核验: 在考试中, 确保考生的身份是防止作弊的关键环节。声纹识别技术可以用于考生身份核验, 通过比对预先存储的声音样本确认考生的身份, 有效防止替考等作弊行为。这种身份核验方式具有高度的准确性和安全性, 为考试公平提供了有力保障。

2) 线上课程打卡: 在线教育和远程学习的普及使得学生的出勤率成为一个挑战。通过声纹识别技术进行线上课程打卡, 学生只需通过语音识别即可完成签到, 既方便又准确。这种方式能够确保学生按时参加课程, 提高出勤率和学习效果的管理效率。

语言学习与评估

1) 语音助手: 对于语言学习, 尤其是对于非母语的学习, 准确的发音至关重要。声纹识别技术可以帮助学生在学习语言时得到及时的反馈, 确保发音准确。语音助手能够识别学生的发音, 并提供个性化的纠正建议, 有效提高学生的口语和听力技能。

2) 智能评估: 在语言测试中, 传统的评分方式可能受到主观因素的影响。声纹识别技术能够

提供自动化的语音评分, 客观地评估学生的发音、语调和节奏, 使评估过程更加公正和高效。

辅助教学与个性化学习

1) 辅助课堂教学: 教师可以使用声纹识别技术来辅助课堂教学。例如, 在朗读教学中, 教师可以通过声纹识别技术对学生的朗读进行实时评估和指导, 帮助他们改进发音和表达技巧。此外, 教师还可以利用声纹识别技术进行互动式教学, 增强学生的学习兴趣和参与度。

2) 个性化学习: 声纹识别技术可以用于实现个性化学习。通过分析学生的学习风格和特点, 教育机构可以为他们提供定制化的学习资源和辅导, 满足学生的个性化需求。这种基于声纹识别技术的个性化学习能够更好地激发学生的潜力, 提高学习效果。

3.5.2 游戏防沉迷场景

根据中消协发布的报告, 青少年近视和网游消费问题日益严重, 强制实名游戏防沉迷机制存在明显缺陷。为解决这一问题, 声纹识别技术可被应用于青少年防沉迷系统, 提供更可靠的认证方式。

1) 注册与登录环节: 在注册和登录环节, 传统的青少年防沉迷系统依赖于身份证信息和密码进行认证。然而, 冒用身份证或密码被窃取的情况时有发生。声纹识别技术的引入可有效解决这一问题。通过采集用户的声纹信息, 系统可进行身份验证, 确保注册和登录环节的安全性。

2) 游戏内语音对话: 针对游戏类网站, 声纹识别技术可以获取游戏语音对话中的声纹信息, 准确判断使用者的身份。当检测到用户为青少年时, 系统自动开启 “ 青少年模式 ”, 对使用时段、服务功能和在线时长进行限制。例如,每天使用时间不超过 40 分钟,晚上 10 点至早上 6 点无法使用, 禁止直播和同城浏览功能, 限制充值、提现和打赏等操作。

3) 全流程认证: 声纹识别技术可以与青少年防沉迷系统结合, 实现全流程认证。在游戏过程中,系统定期采集用户声纹信息,与注册时的声纹信息进行比对,确保使用者身份与注册身份一致。一旦发现异常, 系统可自动退出 “ 青少年模式 ”, 并提醒家长或管理员进行处理。

4) 防止绕过系统: 针对部分用户绕过防沉迷系统的情况,声纹识别技术可增加额外的安全层。即使 App 卸载后重新安装或关闭弹窗提示, 系统仍可通过声纹信息进行身份验证。此外, 家长设置的密码被猜到或冒用身份证的情况也能得到有效遏制。

3.5.3 智慧医疗场景

在医疗领域, 医疗问题频发的今天, 非法挂号、医保冒用、医疗器械操作失误等问题屡见不鲜。而声纹识别技术以其独特的优势, 为医疗行业提供了有效的解决方案。

优化医疗服务流程

1) 预约挂号: 通过声纹预约挂号, 可以有效制约号贩子的非法挂号行为, 确保患者能够公平

地获得医疗服务。这一措施大大提高了挂号的公正性和安全性。

2) 医保防伪: 声纹识别技术为医保防伪提供了有力支持。通过声纹识别, 可以准确鉴别患者身份, 防止医保冒用现象的发生, 确保医保资金的安全使用。

3) 一体化流程: 以声音作为身份信息, 可以实现挂号、住院、缴费等流程的一体化。这将大大简化患者的就医流程, 提高医疗服务的效率。

提升医疗安全与质量

1) 医疗器械权限管理: 利用声纹识别技术为医疗器械加权限, 可以避免误操作, 降低医疗事故的风险。这一措施有效提高了医疗工作的安全性和准确性。

2) 跨省共诊与支付: 通过声纹识别技术, 可以实现跨省共诊及跨省支付。这将打破地域限制, 让患者享受到更为便捷的医疗服务。

3) 智能录入及医疗纠纷辅助鉴定: 采用 “ 智能识别 + 人工审核 ” 的模式, 医生通过语音识别录入病例档案, 解放了医护人员的双手和双眼, 提高工作效率。后期如遇到医疗纠纷时, 通过语音时间戳对涉事人员的语音样本进行采集和分析判断涉事人员的身份,通过对语音中的关键词,语气、语速等内容的分析, 判断涉事人员在纠纷中的态度和情绪状态, 近一步为鉴定结果提供依据。

4) 病灶识别: 声纹识别技术还可以用于准确识别病人的情绪及声音病灶。这一创新功能结合传统望闻问切的诊疗方式, 有助于医生更全面地了解患者状况, 从而减少医疗问题的发生。

3.6 消费物联网

物联网时代, 智能语音将成为最为符合应用场景的人机交流模式, 有望成为每个智能硬件的 “ 标配 ”。但目前市场上的语音助手普遍不具备身份识别的能力, 无法做到个性化的语音交互。也正因为这一点, 使得语音助手存在一些侵入风险。黑客和欺诈者可能会通过语音命令应用程序, 以打开网站, 购买, 甚至关闭警报系统和解锁门。这些安全风险也来自于语音助手用户密码薄弱和缺乏身份验证。

如某公司为智能设备、物联网、车联网等提供以声纹技术为核心的智能语音综合解决方案, 将语音识别、声纹识别、语音情感识别、语音鉴伪技术有机结合在一起, 在自动识别和理解语音命令的同时进行无感身份认证, 使智能设备不仅可以准确识别语音内容, 更可以基于说话人的身份提供个性化服务。

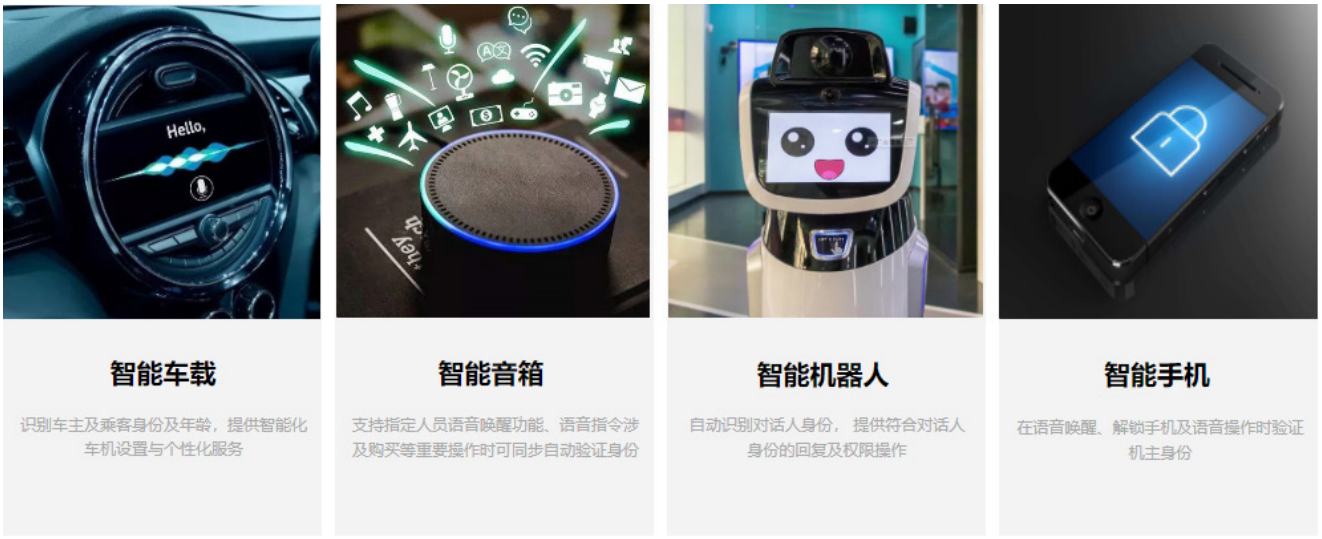
消费物联网领域主要应用场景举例:

1) 智能车载: 识别车主及乘客身份及年龄, 提供智能化车机设置与个性化服务;

2) 智能音箱: 支持指定人员语音唤醒功能、语音指令涉及购买等重要操作时可同步自动验证身份 ;

3) 智能机器人: 自动识别对话人身份， 提供符合对话人身份的回复及权限操作;

4) 智能手机: 在语音唤醒、解锁手机及语音操作时验证机主身份。



3.7 工业物联网

在工业生产中, 质检是一个至关重要的环节, 它能够确保产品的质量和安全性。然而, 传统的质检方法通常需要人工进行, 效率低下且容易出错。为了解决这个问题, 声纹识别技术被引入到工业质检领域。

声纹识别技术通过采集和分析产品的声音信号, 检测其是否存在异常或缺陷。在工业生产线上, 每个产品都会发出独特的声音, 这些声音与正常状态下的声音存在差异。通过声纹识别技术, 可以快速准确地检测出异常声音, 从而判断出产品是否存在问题。

具体来说, 声纹识别技术在工业质检中的应用包括以下几个方面:

1) 声音采集: 使用专业的声音采集设备, 采集生产线上的声音信号。这些信号可以包括产品在生产过程中的声音、机械运转的声音等。

2) 特征提取: 对采集到的声音信号进行特征提取, 提取出与产品质量相关的特征。这些特征可以包括声音的频率、幅度、波形等。

3) 声纹比对: 将提取出的特征与标准声纹库中的特征进行比对,判断产品是否符合质量要求。如果发现异常声音特征, 则说明产品存在质量问题。

4) 结果输出: 将质检结果输出到控制系统中, 对不合格的产品进行筛选和剔除。同时, 也可以将质检结果记录下来, 用于后续的质量分析和改进。

典型工业声纹应用场景举例:

1) 发动机故障诊断: 传统的发动机故障诊断主要依赖于工程师的技术能力, 通过对发动机噪声的强度进行分析, 可以大致判断出发动机部件的故障。工业声纹技术可以进一步提高故障诊断的准确性和效率。

2) 电机故障检测: 当电机发生故障时, 传统的维护方式需要人工听电机发出的声音来判断故障类型, 这种方式既耗费人力又无法保证及时检测到故障。基于声信号的声纹识别系统可以识别出电机异响及各种类型的故障, 如线圈破碎和定子线圈短路等。

3) 水轮机状态监测: 水轮机是水电站中的关键设备, 其运行状态直接影响到整个水电站的发电效率和安全性。通过安装工业声纹传感器, 可以实时监测水轮机的运行声音, 并借助声纹识别技术对声音进行特征提取和分析。这有助于及时发现水轮机的异常声音, 如轴承磨损、齿轮松动等问题, 从而及时采取相应的维修措施, 避免设备损坏和发电中断。

4) 水泵房噪声控制: 水泵房是水电工程中用于供水、排水的重要设施, 但同时也是噪声污染的主要来源之一。工业声纹技术可以用于水泵房噪声的监测和分析, 帮助工程师识别噪声源并采取相应的降噪措施。通过优化水泵房的设计和运维方式, 可以有效降低噪声对周围环境的影响, 提高水电工程的环保性能。

5) 管道泄漏检测: 水电工程中的输水管道如果发生泄漏, 不仅会造成水资源的浪费, 还可能对周围环境造成损害。工业声纹技术可以用于管道泄漏的监测和定位。通过采集管道沿线的声音信号, 利用声纹识别技术对声音进行特征提取和比对, 可以快速准确地定位泄漏点, 为及时修复泄漏提供有力支持。

随着技术的不断进步和市场的日益成熟, 工业声纹技术在设备故障检测与预测性维护、安全监控与入侵检测和工业自动化和智能制造等具体领域场景得到了广泛的应用。通过对设备运行声音的实时监测和分析, 及时发现设备异常声音, 从而预测可能发生的故障, 并采取相应的维护措施。这大大提高了设备的运行效率和使用寿命, 降低了维护成本, 通过与其他工业设备和系统的集成, 工业声纹技术可以实现自动化控制、智能调度等功能, 提高生产效率和质量。

然而, 尽管工业声纹技术的应用前景广阔, 但仍然存在一些挑战和问题需要解决。例如, 声纹识别技术的准确性和稳定性需要进一步提高, 以适应各种复杂的工业环境; 同时, 如何降低声纹识别设备的成本, 使其更加普及和实用, 也是当前需要解决的问题。

四、产品篇

4.1 身份验证类

主要包括声纹登录、支付、锁控等软 / 硬件产品，对应于 1:1 身份认证场景，是目前最主流的声纹识别应用产品形态。

4.1.1 “ 动态声纹密码 ” 可信身份认证系统

产品形态：

软件产品服务 / 可提供 SDK 或 API 接口调用

产品简介：

结合声纹、语音以及动态密码等技术，来对用户身份进行确认，以提高身份认证的安全性。需要认证时，系统会随机产生一组动态码（如 6 位或 8 位数字）要求用户朗读，系统对用户读出的声音进行语音识别并将识别的内容与发出的动态码数字进行比对，同时系统对用户的发音进行声纹比对，两种认证手段都通过时才判断通过。

应用方向：

基于动态声纹密码的 “ 声纹 +” 移动金融解决方案为银行、证券、保险等金融机构提供全流程声纹识别服务，适用于移动客户端、呼叫中心、线下智能机具等场景的客户声纹确认、黑名单客户辨认，身份假冒导致的欺诈交易、冒名账户开立，降低账户盗用导致金融机构的客户赔付及风控成本，提高客户服务满意度，保障金融领域交易和业务的安全性。

产品认证：

“声纹识别系统”于 2019 年 10 月被国家市场监管总局、中国人民银行列入我国首批金融科技产品认证目录，是第一个进入该目录的生物特征识别产品品类；2019 年 11 月得意音通声密保产品获得中金国盛认证中心颁发的“移动金融技术服务认证”证书，是我国首个获得该牌照的声纹识别产品。

应用案例：

声密保产品目前已经广泛应用于金融业（包括 30 余家银行及中国银联、中国互金协会等）、社保生存认证（贵州社保、内蒙古社保、江苏社保等）、电子政务（国家信息中心、中国政务服务平台等）、公安（贵州省、陕西省等）等。

4.1.2 声纹智能门锁

声纹智能门锁是智能门锁领域的“新物种”，它结合了声纹识别技术与传统密码等技术，是“动态声纹密码”可信身份认证产品的硬件演进，填补了居家级智能门锁市场上的“声控”空白，将声纹识别应用产品成功延伸到 C 端。

动态声纹密码技术是声纹智能门锁的核心技术之一。每次解锁时，门锁都会根据用户的语音特征和内部算法动态生成一个新的声纹密码。这个密码具有极高的独特性和不可复制性，使门锁能够准确地识别并验证用户身份，确保只有经过授权的人才能进入家中。

同时，声纹智能门锁还支持远程控制。用户可以通过手机应用程序随时随地查看门锁的状态和访问记录。无论身处何地，都可以轻松掌控家门的开关状态，确保家庭的安全。此外，该门锁还采用先进的加密技术，确保用户的声纹数据在传输和存储过程中得到充分保护。用户无需担心个人隐私泄露的风险。

最后，声纹智能门锁可与智能家居系统无缝集成，实现智能家居设备的互联互通。用户可设置语音指令控制门锁的开关状态，也可以通过智能家居系统查看门锁的状态和访问记录。

4.1.3“ 声纹 +” 门禁系统

该系统通过声纹识别技术对用户进行身份验证，只有通过验证的用户才能进入特定区域或访问特定资源。

产品简介：

通过终端设备的摄像头 + 麦克风，一次性采集通行人员的声纹和人脸双生物特征信息，与数据库进行比对，根据判定结果控制门禁设备通行。一次性完成用户的双特征身份比对及红外测温，同时克服光线、噪音等干扰因素，让身份认证顺畅自然，准确率更高。

应用方向：

智能安防，具体场景有：政企办公区、高档住宅区、学校宿舍、酒店客房、银行金库、科研实验室、军事器材库等。

4.2 音频分析类

主要包括真假声纹鉴定、1:N 声纹检索、声音故障诊断等场景，是目前市场需求较为旺盛但技术成熟度仍有待打磨的声纹识别应用产品形态。

4.2.1 “声纹+”音频鉴伪平台

“声纹+”音频鉴伪平台是一款基于声纹技术的音频鉴伪平台，通过先进的声纹技术和模型算法，实现对音频文件的准确鉴伪。它具有庞大的模型库、高精度和稳定性的模型算法，以及强大的扩展性和灵活性，为音频鉴伪提供了全新解决方案，为音频数据的真实性和安全性提供了有力的保障。

该平台拥有庞大的模型库，涵盖了由数千人、数百台不同型号手机采集的近百万条重放语音数据。这些数据经过严格的采集和处理，建立了高精度、高稳定性的模型库，为后续的鉴伪工作提供了坚实的基础。

在模型库的基础上，平台进行了深入的模型算法训练。通过不断的优化和调整，平台成功地提高了鉴伪的准确性和稳定性。无论是对真实语音还是对伪造语音，平台都能够快速、准确地识别出其真实性和来源。

此外，“声纹+”音频鉴伪平台还具备强大的扩展性和灵活性。随着技术的不断进步和数据量的不断增加，平台可以持续更新和优化模型库和算法，进一步提高鉴伪的准确性和效率。

4.2.2 声纹鉴定工作站

声纹鉴定工作站是专为司法领域“语音同一性认定”而设计的语音鉴定设备，集成了语音采集、分析和比对等多种功能，可帮助司法机关、公安部门等对语音证据进行科学、准确的鉴定，为案件调查和审判提供有力支持。

声纹鉴定工作站采用先进的声学和语音学技术，能够对声音进行高精度的采集、存储和分析。它能够提取出说话人的语音特征，如音调、音色、语速等，并与其他声音样本进行比对，以确定它们是否来自同一人。

该工作站还配备了强大的数据库管理系统，可以存储和检索大量的语音样本。用户可以通过简单的操作，对语音样本进行查询、比对和鉴定，大大提高了工作效率和准确性。

此外,声纹鉴定工作站还具有高度的可定制性,可以根据不同用户的需求进行个性化配置,例如,可以配置不同的麦克风、采样率、分析算法等,以满足不同场景下的需求。

4.2.3 智能听诊器

智能听诊器是一款由清华大学研发团队引领的科技产品,通过先进的声纹技术和智能化功能,既可为医生提供更准确的诊断依据和治疗方案,也可为慢病患者提供便携的居家监测手段。

智能听诊器采用了先进的声纹技术,能够捕捉并分析人体内部器官发出的声音信号,包括心音、肺音等。这种技术利用了声波在人体内部传播的特性,通过捕捉和分析声波的反射、折射等变化,能够判断出人体内部器官的状态和异常。这种声纹技术具有高精度、高敏感性和非侵入性的特点,为医疗诊断提供了新的手段和工具。

智能听诊器还具备智能化功能,能够自动识别和分析患者的声音数据,为医生提供更准确的诊断依据。通过与医疗设备的连接和数据共享,医生可以快速了解患者的病情和病史,为患者提供个性化的治疗方案。

此外,智能听诊器还采用了先进的信号处理技术和降噪技术,能够消除环境噪音和其他干扰因素,确保声音数据的准确性和可靠性。同时,该产品还具备便携性和易用性,方便医生在临床实践中随时使用。

4.2.4 工业声纹检测系统

工业声纹检测系统是一款针对工业场景中声音异常检测和故障预警的高效解决方案。该系统利用声音识别技术和深度学习算法,对工业设备运行过程中的声音进行实时监测和分析,有效识别出异常声音并即时预警,帮助企业及时发现和解决潜在的设备故障,保障生产安全和稳定。

系统核心功能包括实时声音采集、异常声音检测、故障预警和数据分析等。通过高灵敏度的声音传感器和阵列麦克风,系统能够实时采集设备运行过程中的声音信号,并进行快速处理和分析。利用深度学习算法,系统可以自动识别出异常声音,如轴承损坏、电机故障等,并即时发出预警信号,提醒工作人员及时检查和维修。同时,系统还可以对采集的声音数据进行存储和分析,帮助企业了解设备运行状况和性能,为设备维护和优化提供数据支持。

工业声纹检测系统的优势在于非接触式监测、高精度识别和实时预警。该系统无需接触设备即可进行声音监测,有效避免了传统接触式监测方法对设备的干扰和损伤。同时,系统采用先进的深度学习算法,能够快速准确地识别出异常声音,提高了故障预警的准确性和及时性。此外,系统还支持远程监控和数据分析,方便企业进行集中管理和远程控制。

工业声纹检测系统是工业生产中不可或缺的智能监测设备,能够帮助企业实现高效的声音异常检测和故障预警,保障生产安全和稳定。

4.3 语音助手类

指将声纹识别技术和语音识别、自然语言理解等技术相结合，以替代纯文字交互功能、具备个性化交互能力的 AI 助手，在某些行业也被称作智能客服。

4.3.1 “ 一句话解决问题 “ 金融级智能语音助手

产品简介：

将语音识别、声纹识别、语音情感识别、语音鉴伪技术有机结合在一起，在自动识别和理解语音命令的同时进行无感身份认证，使智能语音系统不仅可以准确识别语音内容， 更可以基于说话人的身份提供个性化服务。

应用方向：

如，当用户发起转账、支付、缴费等交易请求时，原有的交互流程从验证发起人身份、识别目标账户、输入金额等至少三步缩减为只需一步，交互时长由至少需数十秒减少为只需几秒钟。这种多种先进而复杂的人工智能技术的协同和融合，不仅大大提高了交互效率，降低了使用门槛，而且在保证同等安全级别的前提下，有力提升了用户体验，可谓 “ 无需腾出手，只要动动嘴 ”。

随着使用门槛的降低，“一句话”语音模式的应用场景和适用人群还将不断扩大：一方面从转账、支付、缴费等特定场景逐步过渡到更多通用场景，另一方面也可以从老年人逐渐覆盖至各年龄层到更广泛的特殊群体，例如青少年，以及视障人士等特殊群体。

4.3.2 智能音箱语音助手

智能音箱语音助手是一款结合了人工智能和声纹技术的语音交互产品。通过识别用户的语音指令，该助手可以完成各种智能家居控制、查询信息、播放音乐等任务。而声纹技术的应用，使得智能音箱语音助手能够更准确地识别不同用户的身份，提供更加个性化的服务。

在智能音箱语音助手中，声纹技术的应用主要体现在声纹识别方面。通过对用户的语音信号进行采集和分析，智能音箱可以提取出用户的声纹特征，从而判断出用户的身份。这种技术的应用，使得智能音箱能够根据不同用户的偏好和习惯，提供个性化的服务。

例如，当用户通过智能音箱语音助手发出“播放音乐”的指令时，智能音箱会先识别出用户身份，然后根据用户平时的音乐喜好，推荐适合的音乐。这样，每个家庭成员都可以通过智能音箱获得自己喜爱的音乐，提高了用户体验。

此外，声纹技术还被广泛应用于智能音箱的声纹认证功能中。通过采集用户的语音样本并存储在云端，智能音箱可以在需要验证用户身份的场景下，与云端存储的声纹样本进行比对，从而判断用户的身份。这种认证方式具有很高的安全性，可以有效防止非法入侵和误操作。

4.3.3 老人居家安全呼叫器

老人安全 AI 呼叫器是一款智能化的设备，通过声纹技术识别老年人的声音特征，能够准确判断老年人的身份和需求。当老年人遇到紧急情况或需要帮助时，只需按下呼叫器上的按钮，设备就会立即接收到信号并发出警报。同时，设备还会将老年人的声音特征与预先存储的数据进行比对，确保准确识别老年人的身份，避免误报或漏报。

利用声纹技术不仅提高了识别准确率，还具有非侵入性和隐私保护的特点。老年人在使用过程中无需进行繁琐的操作，只需简单按下按钮即可发出警报。同时，声纹技术也不会泄露老年人的个人信息，确保了他们的隐私安全。

除声纹技术外，老人安全 AI 呼叫器还具备其他智能化功能。例如，设备可以与智能手机等设备连接，方便子女或亲属随时了解老年人的情况。此外，设备还具备远程控制功能，子女或亲属可以通过手机应用程序远程设置警报阈值、查看警报记录等，为老年人提供更加全面的安全保障。

老人安全 AI 呼叫器利用先进的声纹技术为老年人提供更安全、更便捷的生活体验。通过准确识别老年人的声音特征和身份，该设备能够及时发出警报并通知相关人员进行处理，确保老年人的安全和健康。

4.4 声纹采集类

严格来说, 声纹采集并不足以支撑起一个独立的产品品类, 因其是所有声纹类产品和服务的前置要求, 且目前绝大部分的声纹采集通过手机等智能终端的内置麦克风就可以实现, 并不需要更专业的设备。不过, 鉴于在某些特定场景下, 需求方对于声纹采集质量提出了较高要求, 有厂家也相应做了一些工作。因此本报告中将此部分单列。

4.4.1 声纹采集终端

声纹采集终端通常采用专业的声音采集硬件, 采取标准化的接口设计, 同时嵌入语音预处理算法, 自动对输入的语音进行降噪、去混响等处理, 以获取清晰的语音信号, 确保后续识别的准确性和高效性。此外, 这类设备一般还支持多种不同的输入方式, 如麦克风、线路输入等, 以满足不同场景下的需求。

除基本的声纹采集功能外, 声纹采集终端也可以具备其他多种功能, 如语音录制、语音比对、语音转换等。用户可根据实际需求选择不同的功能模块, 实现个性化的声纹识别服务。此外, 声纹采集终端还支持与其他系统的集成和数据交换, 如公安、司法机关的信息化系统等, 为用户提供全方位的声纹识别解决方案。该设备采用方便与其他系统的集成和数据交换。

2020 年 2 月 13 日,《个人金融信息保护技术规范》(标准编号: JR/T 0171-2020) 由中国人民银行正式发布, 即日起实施。

该标准 “ 根据信息遭到未经授权的查看或未经授权的变更后所产生的影响和危害 ”, 将个人金融信息按敏感程度从高到低分为 C3、C2、C1 三个等级。其中, 2018 年移动金融声纹标准中定义的 “ 动态声纹密码 “ 被列入较低隐私敏感度级别的 C2 级个人信息, 与被列为高隐私敏感度的 C3 级个人信息 “ 用于用户鉴别的个人生物识别信息 ” 区别开来。

2021 年 1 月 22 日,《远程声纹识别应用技术规范 第 1 部分: 身份验证》由 IIFAA (互联网金融身份认证联盟) 发布。由得意音通、蚂蚁金服等起草。

5.2.2 公安应用标准

2010 年 12 月 2 日,《安防生物特征识别应用术语》(标准编号 GA/T 893-2010) 由公安部发布。这是我国第一个关于声纹识别应用的行业标准。

2018 年 5 月 14 日, 全国安全防范报警系统标准化技术委员会 (简称安标委, 秘书处设在给公安部第一研究所) 下设的人体生物特征应用分委员会 (SAC/TC100/SC2) , 投票通过了声纹识别标准化体系建设 12 项标准中的 3 项, 进入起草阶段, 它们是:《声纹数据采集的技术要求》、《声纹数据质量评价标准》、《声纹数据建库要求》, 由得意音通与清华大学牵头, 对规范我国未来声纹身份认证具有重要意义。

经专家讨论, 此 3 项标准现已合并为 1 项国家标准《公共安全 声纹识别应用 第 1 部分: 采集识别建库技术要求》, 目前已进入报批稿阶段, 即将由国家标准委和国家市场监督管理总局联合发布。

5.2.3 电信应用标准

《电信行业语音声纹库技术要求》由中国电信研究院、中国信息通信研究院、得意音通等共同发起, 由中国通信标准化协会归口, 是声纹识别在电信领域应用的首个标准规范。

5.2.4 平台应用标准

由中国信息通信研究院牵头的中国人工智能产业发展联盟 (AIIA) 发布。

5.3 数据标准

2021 年 6-7 月, 国家标准《信息安全技术 声纹识别数据安全要求》启动了试点修订工作, 由全国信息安全标准化技术委员会 (简称 “ 信安标委 ”) 归口。

该标准是基于 GB/T 35273《信息安全技术 个人信息安全规范》中对个人生物识别信息的定义, 并且在对《信息安全技术 生物特征识别信息保护要求》的细化与补充基础上, 结合《信息安全技术 网络数据处理安全规范》和《个人信息保护法》对数据处理安全保护内容, 提出的声纹识别数

据安全要求。

该标准由得意音通牵头制定, 中国电子技术标准化研究院提供技术支撑, 国家工业信息安全发展研究中心、清华大学、厦门大学、百度网讯、科大讯飞等众多单位参与, 遵循《信息安全技术个人信息安全规范》, 以保护用户信息和隐私信息为出发点, 从数据安全保护的角度对使用以及处理声纹识别数据的应用系统提出安全要求。可为提升声纹识别产品和服务安全性, 防范个人信息和隐私信息安全风险提供重要支撑。

5.4 评测标准

2014 年 8 月 18 日, 公安部颁布了《安防声纹确认应用算法技术要求和测试方法》(标准编号 GA/T 1179-2014) , 并于同年 10 月 1 日实施。这是我国首次就声纹识别的技术评测制订标准。

2019 年 9 月 11 日, 公安部颁布了《声纹自动识别系统测试规范》(标准编号 GA/T 1587-2019) , 并于同年 11 月 1 日实施。该标准由全国刑事技术标准化技术委员会 (SAC/TC 179) 归口。

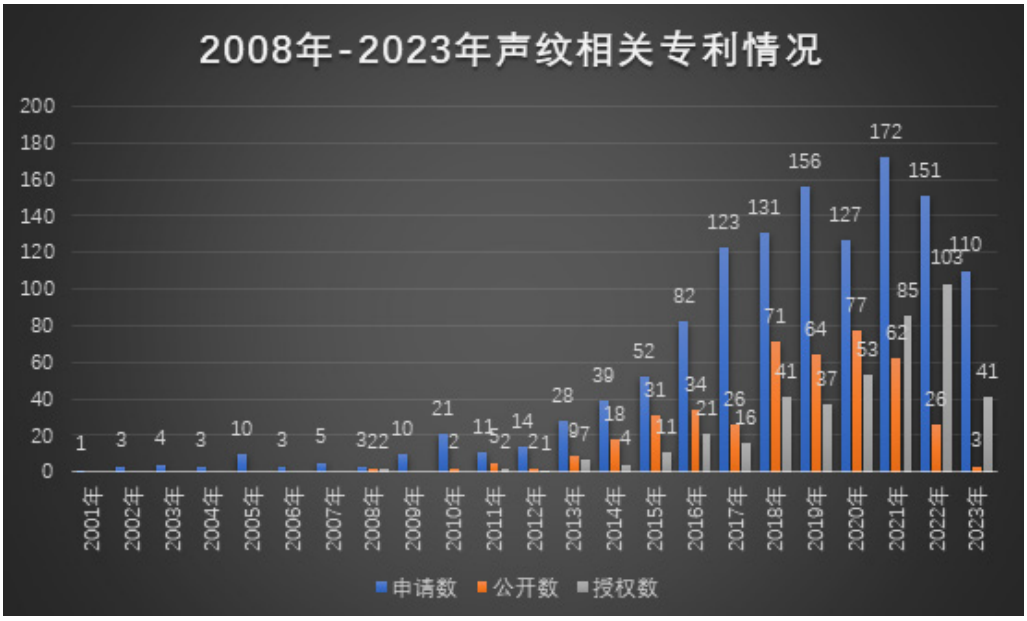
2021 年 6 月 29 日, 中国支付清算协会发布《移动金融基于声纹识别的安全应用评估规范》, 针对声纹识别在金融领域的应用, 对于 “ 声纹服务器评估 ” 和 “ 客户端软件评估 ” 分别给出了评估规范, 并根据不同测试项给出了具体的声纹语音样本库要求。

研发路线开始从单纯追求快速、准确, 向更注重隐私、安全演进。另一方面, 区别于人脸等静态生

六、行业篇

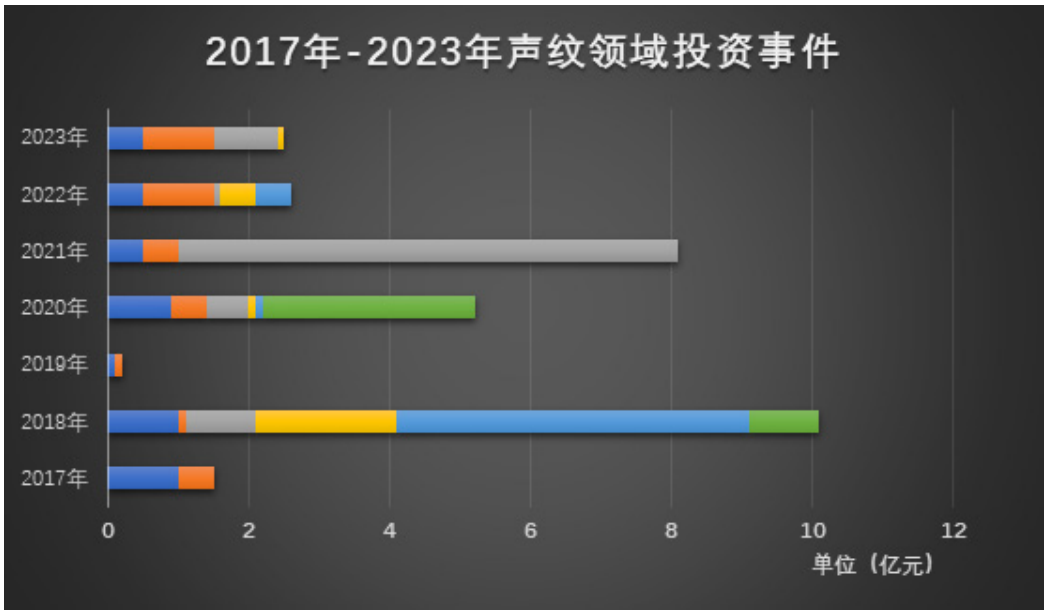
6.1 本领域专利情况

从 2008 年至今, 声纹相关专利的申请、公开及授权数整体呈增长趋势, 专利申请数于 2021 年专利申请数到达峰值, 继而开始回落; 声纹相关专利公开数于 2020 年到达峰值, 继而开始回落; 声纹相关专利授权数于 2022 年到达峰值, 继而开始回落。



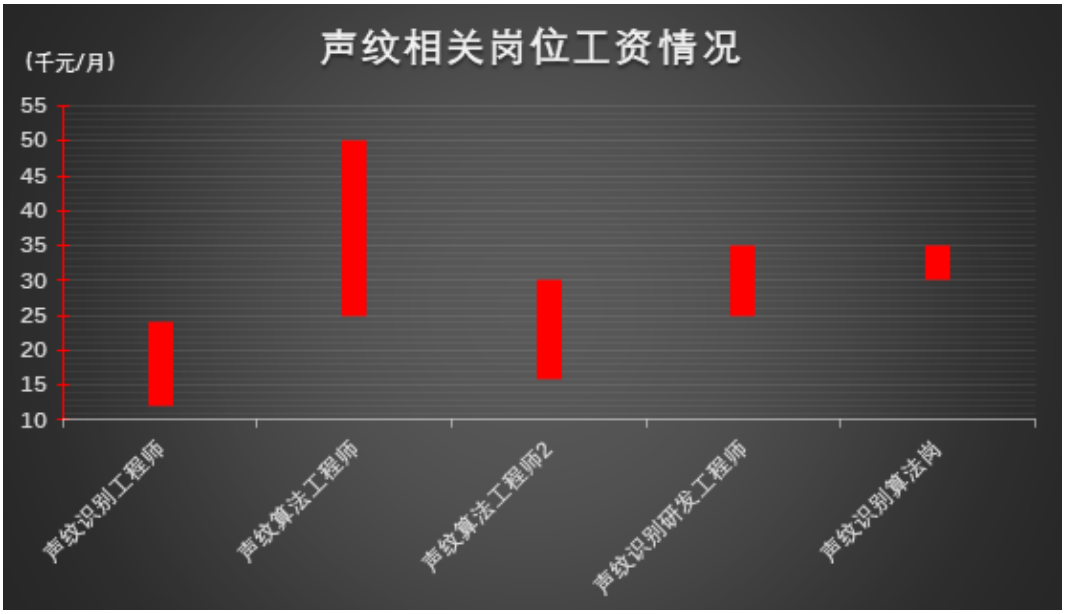
6.2 本领域投资事件

从 2017 年到 2023 年,除了 2019 年的投资事件明显较少,其余年份的融资数量和金额相对平稳。



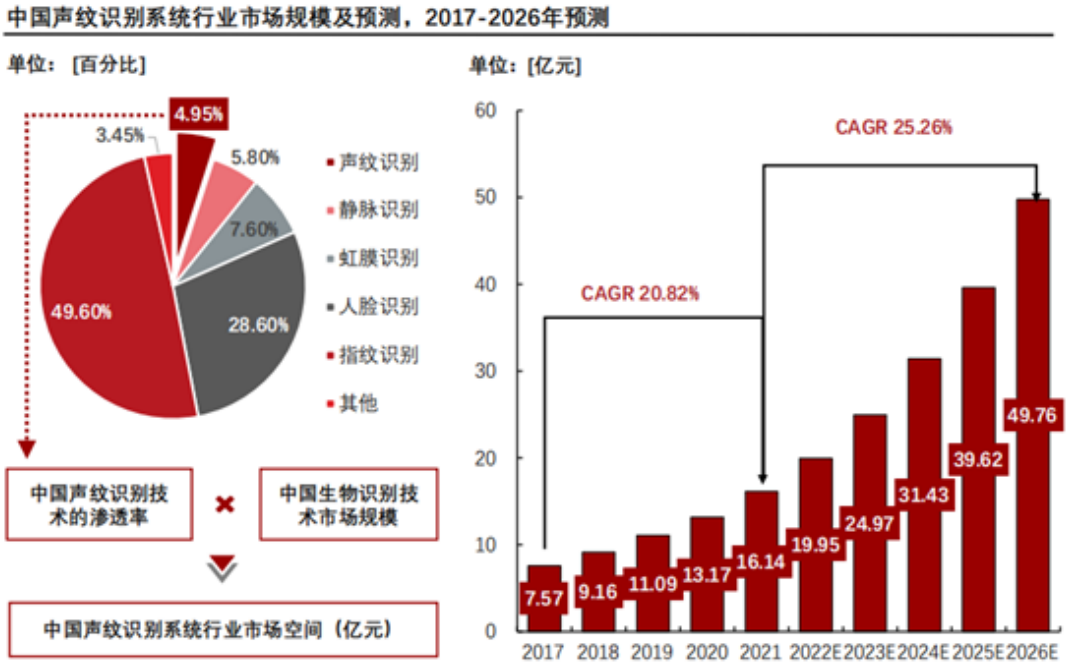
6.3 本领域人才需求

声纹识别算法工程师的工资幅度从 12,000 元 / 月到 50,000 元 / 月不等。



6.4 本领域市场预测

据沙利文头豹研究院, 中国声纹识别系统行业在政策、技术发展和资本的驱动下将迎来新的发展, 市场规模将进一步扩大, 预计 2026 年中国声纹识别系统行业市场规模将达 49.76 亿元,



2021-2026 年 CAGR 为 25.26%。

声纹识别尽管在生物识别市场份额中占比还较小, 但近年来得益于云计算、大数据、物联网、

深度学习等信息技术发展愈发成熟, 以及生物识别从生理特征走向行为特征的发展趋势, 声纹识别有望在金融、公安等领域发挥越来越重要的作用。

此外, 系列政策法规出台, 国家加快 AI 产业引导, 2018 年中国人民银行颁布《移动金融基于声纹识别的安全应用技术规范》是声纹识别迈向产业化的分水岭。得意音通、快商通、声扬科技、捷通华声等声纹识别相关企业先后获得融资, 行业风口可期。在政策、技术发展和资本的驱动下中国声纹识别系统将迎来新的发展。

七、后记

尽管声纹识别技术在某些领域的应用已经相当成熟, 如金融科技与政务民生, 但在其他不少领域的应用却仍然处于起步阶段。一方面, 这种不匹配源自于技术的特性: 声纹识别技术涉及到复杂的生物特征提取和比对, 需要大量的数据和算法支持。因此, 技术的成熟需要时间, 需要大量的研发投入。另一方面, 这种不匹配也来自于市场需求: 尽管有些领域对声纹识别技术的需求非常迫切, 但由于种种原因 (如预算限制、技术门槛高等) , 这些领域的技术应用并不一定能够迅速成熟。

举例来说, 反欺诈、反洗钱、电信诈骗和 AI 诈骗相关场景因涉及到财产和人身安全具有较高的市场价值, 因而在上述场景中格外受到市场的关注。但由于声纹识别技术中的声纹确认 (1:1 认证) 和声纹辨认 (1:N 搜索) 的技术成熟度不同, 导致这些备受关注的场景目前并未产生很好的应用效果。

对于声纹识别技术供应商来说, 如何平衡这种场景关注度与技术成熟度的关系, 是一个巨大的挑战。一方面, 他们需要理解市场的需求, 把握市场的脉搏, 针对关注度高的场景进行研发和应用。另一方面, 他们也需要对技术本身有深刻的理解, 知道哪些技术是可以应用的, 哪些还需要进一步的研究和发展。

而对于行业用户来说, 如何理解和应用声纹识别技术也是一个挑战。他们需要理解技术的成熟度, 知道哪些场景是可以应用的, 哪些还需要进一步的观察和研究。同时, 他们也需要理解市场的动态, 知道哪些是当前关注的热点, 哪些是未来的趋势。技术的发展和应用是一个复杂的过程, 涉及到技术本身、市场需求、政策环境等多个因素。只有深入理解这些因素, 才能更好地推动声纹识别技术的发展和应用。

主要参考文献

[1] 埃森哲,《人工智能成熟之道: 从实践到实效》, 2022.12

[2] A. Afshan, J. Guo, S. J. Park, V. Ravi, A. McCree, and A. Alwan, “Variable Frame Rate-Based Data Augmentation to Handle Speaking-Style Variability for Automatic Speaker Verification,” INTERSPEECH 2020, Oct. 2020, pp. 4318–4322.

[3] B. J. Borgstrom and P. Torres-Carrasquillo, “Bayesian Estimation of PLDA with Noisy Training Labels, with Applications to Speaker Verification,” ICASSP 2020.

[4] CNSRC, http://cnceleb.org/workshop

[5] CSSD, https://magichub.com/competition/sec-competition/

[6] D. Cai, W. Cai, M. Li, “Within-Sample Variability-Invariant Loss for Robust Speaker Recognition Under Noisy Environments, ” ICASSP 2020.

[7] J. S. Chung, J. Huh, et al., "In Defence of Metric Learning for Speaker Recognition", INTERSPEECH 2020.

[8] W. Cai, J. Chen, J. Zhang, and M. Li, “On-the-Fly Data Loader and Utterance-Level Aggregation for Speaker and Language Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 1038-1051, 2020.

[9] W. Chen, J. Huang, T. Bocklet, “Length- and Noise-aware Training Techniques for Short-utterance Speaker Recognition,” INTERSPEECH 2020.

[10] B. Desplanques, J. Thienpondt, and K. Demuynck, “ECAPA-TDNN: Emphasized Channel Attention, Propagation and Aggregation in TDNN Based Speaker Verification,” INTERSPEECH 2020.

[11] C. Du, B. Han, S. Wang, Y. Qian, K. Yu, "SynAug: Synthesis-Based Data Augmentation for Text-Dependent Speaker Verification", ICASSP 2021.

[12] J. Deng, J. Guo, N. Xue, et al. “ArcFace: Additive Angular Margin Loss for Deep Face Recognition,” arXiv: Computer Vision and Pattern Recognition, 2018.

[13] L. Li, D. Wang, C. Zhang, T. F. Zheng, "Improving Short Utterance Speaker Recognition by Modeling Speech Unit Classes," IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol.24, No.6, June 2016.

[14] FFSVC,https://ffsvc.github.io/

[15] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, “Res2Net: A New Multi-Scale Backbone Architecture,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 2, p. 652–662, Feb 2021.

[16] 国家互联网信息办公室 中华人民共和国工业和信息化部 中华人民共和国公安部令第 12 号,《互联网信息服务深度合成管理规定》2022.11

[17] 国务院办公厅,《国务院办公厅印发关于切实解决老年人运用智能技术困难实施方案的通知》, 2020.11

[18] A. Hajavi, A. Etemad, "A Deep Neural Network for Short-Segment Speaker Recognition," INTERSPEECH 2019.

[19] H. Huang, X. Xiang, F. Zhao, S. Wang, Y. Qian, "Unit Selection Synthesis Based Data Augmentation for Fixed Phrase Speaker Verification", ICASSP 2021.

[20] J. Huang and T. Bocklet, “Intel Far-Field Speaker Recognition System for VOiCES Challenge 2019,” INTERSPEECH 2019.

[21] Q. Hong, L. Li, J. Zhang, L. Wan, F. Tong, “Transfer Learning for Speaker Verification on Short Utterances,” INTERSPEECH 2016.

[22] Y. Jung, S. M. Kye, Y. Choi, M. Jung, H. Kim, "Improving Multi-Scale Aggregation Using Feature Pyramid Module for Robust Speaker Verification of Variable-Duration Utterances," INTERSPEECH 2020.

[23] S.M. Kye, Y. Jung, H.B. Lee, S.J. Hwang, H. Kim, "Meta-Learning for Short Utterance Speaker Recognition with Imbalance Length Pairs," INTERSPEECH 2020.

[24] K. Liu, H. Zhou, "Text-independent Speaker Verification with Adversarial Learning on Short Utterances," ICASSP 2020.

[25] K.A. Lee, Q. Wang, T. Koshinaka, “Xi-Vector Embedding for Speaker Recognition,” IEEE SPL, June 2021.

[26] N. Li, M.-W. Mak, “SNR-Invariant PLDA Modeling in Nonparametric Subspace for Robust Speaker Verification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 23, No. 10, pp. 11648-1659, 2015.

[27] Y. Liu, L. He, J. Liu, M.T. Johnson, “Speaker Embedding Extraction with Phonetic Information,” INTERSPEECH 2018.

[28] 李荪、曾然然、殷治纲,《AI 智能语音技术与产业创新实践》, 电子工业出版社, 2021.12

[29] Z. Meng, Y. Zhao, J. Li, and Y. Gong, “Adversarial Speaker Verification,” ICASSP 2019.

[30] 倪光南、严明、田霞等,《2021 网信自主创新调研报告》, 2022.4

[31] X. Qin, Y. Yang, L. Yang, X. Wang, J. Wang, M. Li, "Exploring Voice Conversion based Data Augmentation in Text-Dependent Speaker Verification,” arXiv:2011.10710.

[32] 全国人民代表大会常务委员会,《中华人民共和国个人信息保护法》（2021 年）

[33] 全国人民代表大会常务委员会,《中华人民共和国无障碍环境建设法》（2023 年）

[34] 全国网络安全标准化技术委员会,《信息安全技术 声纹识别数据安全要求》GB/T 41807-2022

[35] SASV, https://sasv-challenge.github.io

[36] D. Snyder, G. Chen, and D. Povey, “Musan: A music, speech, and noise corpus,” arXiv:1510.08484, 2015.

[37] D. Snyder, D. Garcia-Romero, G. Sell, D.Povey, and S. Khudanpur, “X-vectors: Robust dnn embeddings for speaker recognition,” ICASSP 2018.

[38] S. Shon, H. Tang, and J. Glass, “VoiceID Loss: Speech Enhancement for Speaker Verification,” INTERSPEECH 2019.

[39] S. Sreekanth, S. M. Rafi B, K. Sri Rama Murty, and S. Bhati, “Speaker Embedding Extraction with Virtual Phonetic Information,” in 2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP), Ottawa, ON, Canada, Nov. 2019.

[40] F. Tong, Y. Liu, S. Li, J. Wang, L. Li, Q. Hong, “Automatic Error Correction for Speaker Embedding Learning with Noisy Labels," INTERSPEECH 2021.

[41] 头豹研究院,《2022 年中国声纹识别系统产业链分析》, 2022.5

[42] VoxSRC, http://mm.kaist.ac.kr/datasets/voxceleb/voxsrc

[43] F. Wang, J. Cheng, W. Liu, et al. “Additive Margin Softmax for Face Verification,” IEEE Signal Processing Letters, 2018, 25(7): 926-930.

[44] M.-W. Mak, X. Pang, and J. Chien, “Mixture of PLDA for Noise Robust i-Vector Speaker Verification,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 1,

pp. 130-142, 2016.

[45] Q. Wang, K. Okabe, K.A. Lee, and T. Koshinaka, "A Generalized Framework for Domain Adaptation of PLDA in Speaker Recognition", ICASSP2020.

[46] S. Wang, J. Rohdin, L. Burget, O. Plchot, Y. Qian, K. Yu, J. Honza Cernock, "On the Usage of Phonetic Information for Text-independent Embedding Extraction", INTERSPEECH2019.

[47] 王泉,《声纹技术: 从核心算法到工程实践》, 电子工业出版社 2020.9

[48] D. Zhou, L. Wang, K. Lee, et al. “Dynamic Margin Softmax Loss for Speaker Verification,” INTTERSPEECH 2020.

[49] F. Zhao, H. Li, and X. Zhang, “A Robust Text-independent Speaker Verification Method Based on Speech Separation and Deep Speaker,” ICASSP 2019.

[50] J. Zhou, T. Jiang, L. Li, Q. Hong, Z. Wang, and B. Xia, “Training Multi-Task Adversarial Network for Extracting Noise Robust Speaker Embedding,” ICASSP 2019.

[51] X. Zhao, Y. Wang, D. Wang, "Robust Speaker Identification in Noisy and Reverberant Conditions", IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 22, no. 4, pp. 836-845, 2014.

[52] 郑方、程星亮,《声纹识别: 走出实验室, 迈向产业化》,《信息安全研究》2019.2

[53] 中国信通院,《可信人工智能白皮书》, 2021.9

[54] 中国人民银行,《个人金融信息保护技术规范》JR/T 0171—2020

[55] 中国人民银行,《移动金融基于声纹识别的安全应用技术规范》JR/T 0164—2018

免责申明：

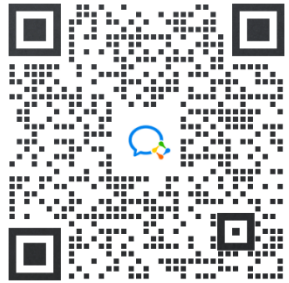
本内容非原报告内容；

报告来源互联网公开数据；如侵权请联系客服微信，第一时间清理；

报告仅限社群个人学习，如需它用请联系版权方；

如有其他疑问请联系微信。

行业报告资源群



微信扫码 长期有效

1. 进群福利：进群即领万份行业研究、管理方案及其他学习资源，直接打包下载
2. 每日分享：6+份行研精选、3个行业主题
3. 报告查找：群里直接咨询，免费协助查找
4. 严禁广告：仅限行业报告交流，禁止一切无关信息

微信扫码 行研无忧

知识星球 行业与管理资源

专业知识社群：每月分享8000+份行业研究报告、商业计划、市场研究、企业运营及咨询管理方案等，涵盖科技、金融、教育、互联网、房地产、生物制药、医疗健康等；已成为投资、产业研究、企业运营、价值传播等工作助手。