



团 体 标 准

T/UNP 255—2024

录音 AI 语音智能分析系统技术规范

Technical specification for recording AI voice intelligent analysis system

2024 - 11 - 07 发布

2024 - 11 - 07 实施

目 次

前言 II

引言 III

1 范围 1

2 规范性引用文件 1

3 术语、定义和缩略语 1

 3.1 术语和定义 1

 3.2 缩略语 1

4 系统框架 1

5 功能要求 2

 5.1 文字识别功能 2

 5.2 语义理解功能 2

 5.3 关键词提取与检索 3

 5.4 说话人识别功能 3

 5.5 音频质量修正功能 3

 5.6 音频修复 3

 5.7 音频格式转换 4

6 性能要求 4

 6.1 准确性 4

 6.2 实时性 4

 6.3 稳定性 4

 6.4 高效性 4

 6.5 可扩展性 4

 6.6 用户友好性 5

7 安全要求 5

 7.1 系统安全 5

 7.2 数据安全 6

8 运维要求 6

 8.1 人员管理 6

 8.2 性能监测与优化 6

 8.3 安全漏洞排查与修复 6

 8.4 功能扩展与升级 7

9 评价与改进 7

参考文献 8

前 言

本文件按照GB/T 1.1—2020《标准化工作导则 第1部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利。本文件的发布机构不承担识别专利的责任。

本文件由武汉铃声科技有限公司提出。

本文件由中国联合国采购促进会归口。

本文件起草单位：武汉铃声科技有限公司、励为（武汉）技术有限公司、武汉新塔医疗科技有限公司、湖北国威时代信息技术有限公司、武汉锐佳讯科技有限公司。

本文件主要起草人：郑伟莲、卢蹁跹、余汉汉、李嘉茵、胡彬。

引 言

为助力中国企业参与国际贸易,推动企业高质量发展,中国联合国采购促进会依托联合国采购体系,制定服务于国际贸易的系列标准,这些标准在国际贸易过程中发挥了越来越重要的作用,对促进贸易效率提升,减少交易成本和不确定性,确保产品质量与安全,增强消费者信心具有重要的意义。

联合国标准产品与服务分类代码(UNSPSC, United Nations Standard Products and Services Code)是联合国制定的标准,用于高效、准确地对产品和服务进行分类。在全球国际化采购中发挥着至关重要的作用,它为采购商和供应商提供了一个共同的语言和平台,促进了全球贸易的高效、有序发展。

围绕UNSPSC进行相关产品、技术和服务团体标准的制定,对助力企业融入国际采购,提升国际竞争力具有十分重要的作用和意义。

本文件采用UNSPSC分类代码由6位组成,对应原分类中的大类、中类和小类并用小数点分割。

本文件UNSPSC代码为“43.23.34”,由3段组成。其中:第1段“43”为大类,表示“信息技术广播和电信”,第2段为中类,“23”表示“软件”,第3段为小类,“34”表示“实用程序和设备驱动程序软件”。

录音 AI 语音智能分析系统技术规范

1 范围

本文件规定了录音AI语音智能分析系统的系统框架、功能要求、性能要求、安全要求、运维要求、评价与改进。

本文件适用于录音AI语音智能分析系统的设计与建设。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

GB 18030 信息技术 中文编码字符集

GB 21023 中文语音识别系统通用技术规范

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件。

3.1.1

语音识别 speech recognition

将人类的语音信号转化为文字或指令的过程。

[来源：GB/T 36464.1—2020, 3.7]

3.2 缩略语

下列缩略语适用于本文件。

CRC: 循环冗余校验 (Cyclic Redundancy Check)

IDS: 入侵检测系统 (Intrusion Detection System)

IPS: 入侵防御系统 (Intrusion Prevention System)

IP: 网际互连协议 (Internet Protocol)

MD5: 信息摘要算法 (MD5 Message-Digest Algorithm)

RT0: 恢复时间目标 (Recovery Time Objective)

RPO: 数据恢复点目标 (Recovery Point Objective)

SSL: 安全套阶层 (Secure Socket Layer)

TL: 传输层安全 (Transport Layer Security)

4 系统框架

系统框架结构如图1所示。



图 1 系统框架图

5 功能要求

5.1 文字识别功能

- 5.1.1 具备高效的实时语音转文字能力，支持在录音进行的同时生成文字内容，便于用户及时查看和处理。
- 5.1.2 转换后的文字应具备良好的格式规范，包括标点符号的准确使用、段落划分清晰等，方便阅读和理解。
- 5.1.3 支持对大量录音文件进行批量语音转文字操作，提高工作效率。
- 5.1.4 系统应准确提取需要处理的录音，且支持下列全部或部分语音识别功能：
- a) 中文语音识别服务；
 - b) 多语种识别：
 - 1) 英语；
 - 2) 汉语；
 - 3) 韩语；
 - 4) 俄语；
 - 5) 德语；
 - 6) 法语；
 - 7) 希腊语；
 - 8) 波兰语；
 - 9) 葡萄牙语；
 - 10) 西班牙语；
 - 11) 土耳其语；
 - 12) 意大利语；
 - 13) 波斯语；
 - 14) 日语。
 - c) 多语种混读识别；
 - d) 个性化识别；
 - e) 自定义语法；
 - f) 识别结果多候选。
- 5.1.5 文字输出结果为汉字内码、系统所定义的符号时，应满足 GB 18030 的要求。

5.2 语义理解功能

5.2.1 意图识别

准确判断语音中表达的意图并标记，例如请求、建议、命令等，为后续的响应和处理提供依据。

5.2.2 上下文关联

理解语音内容的上下文关系，对前后语句进行综合分析，并获得准确的语义理解结果。

5.2.3 逻辑分析

对语音中的逻辑关系进行分析，如因果关系、转折关系等，辅助用户更好地把握语音的核心含义。

5.3 关键词提取与检索

5.3.1 动态更新关键词

随着新的录音内容不断输入，系统应自动更新和优化关键词库，确保关键词提取的准确性和全面性。

5.3.2 模糊检索

支持模糊关键词检索，用户输入的关键词不完全准确，系统应通过相似性匹配找到相关的录音内容。

5.3.3 多维度检索

支持用户不同的检索需求，提供基于时间、说话人、音频属性等多维度的检索方式。

5.4 说话人识别功能

5.4.1 说话人聚类

对大量录音文件中的说话人进行聚类分析，将相似说话人归为一类，并支持用户对特定人群的语音进行集中分析。

5.4.2 说话人跟踪

支持在连续的录音中持续跟踪特定说话人的语音，确保不会出现说话人混淆的情况。

5.4.3 说话人身份验证

结合声纹识别技术、语音密码技术等身份验证技术，对说话人的身份进行准确验证，提高录音分析的安全性和可靠性。

5.5 音频质量修正功能

5.5.1 噪音去除

5.5.1.1 系统应自动检测音频中的各种噪音类型，如环境噪音（如风声、雨声、交通噪音等）、设备噪音（如电流声、风扇噪音等），并根据噪音的特性进行自适应降噪处理。通过先进的信号处理算法，在不影响语音质量的前提下，最大限度地去除噪音，确保语音更加清晰可辨。

5.5.1.2 对音频进行多频段分析，针对不同频段的噪音采用不同的降噪策略。

5.5.2 音频增强

系统的音频增强功能应符合以下要求：

- a) 音量均衡：自动调整音频的音量，使其在不同的录音片段中保持相对稳定的音量水平，避免出现音量忽大忽小的情况；
- b) 频率：对音频的频率响应进行补偿，改善音频的音色和清晰度；
- c) 动态范围压缩：对音频的动态范围进行压缩，使音频中的轻声和大声部分能清晰可闻。

5.6 音频修复

5.6.1 音频剪辑修复

系统应自动检测并修复音频中存在剪辑痕迹或者不连续的音段。通过分析音频的波形和频谱特征，采用插值算法或者信号重建技术，使音频恢复到连续、自然的状态。

5.6.2 音频失真修复

系统应自动检测并修复音频中谐波失真、互调失真等失真现象。通过分析失真的类型和程度，采用相应的信号处理算法进行修复，使音频恢复到原始的状态。

5.7 音频格式转换

5.7.1 多格式支持

系统应满足不同用户对音频格式的需求，支持多种常见的音频格式，如MP3、WAV、FLAC等，并支持在这些格式之间进行自动转换。

5.7.2 格式优化

在进行音频格式转换的过程中，系统应对音频进行优化处理，通过调整音频的码率、采样率等参数，提高音频的质量和兼容性。

6 性能要求

6.1 准确性

6.1.1 文字识别准确率应符合 GB 21023—2007 中系统分类指标要求的规定。

6.1.2 各功能分析准确性应符合以下要求：

- a) 文字识别准确性：按 GB 21023—2007 中系统分类指标要求的规定；
- b) 语义理解准确性：标准普通话识别准确率 $\geq 95\%$ ；外语识别准确率 $\geq 80\%$ ；普通话夹杂部分方言识别准确率 $\geq 70\%$ ；
- c) 说话人识别准确性：多说话人场景下，5 人以内录音识别准确率 $\geq 90\%$ 。

6.2 实时性

6.2.1 在录音过程中支持实时进行语音转文字、语义分析、说话人识别等操作，且延迟时间应控制在可接受范围内。其中，语音转文字的延迟不应超过 3 s，确保系统在实时场景（如直播会议记录）中的可用性。

6.2.2 对录音质量（如音量异常、噪音过大）、分析结果（如关键信息提取、情感倾向判断）等应实时反馈给用户，确保用户及时调整录音设备或获取重要信息。

6.3 稳定性

6.3.1 系统应长时间稳定运行，在连续工作 24 h 甚至更长时间的情况下，不会出现崩溃、卡顿或性能明显下降的情况。

6.3.2 对外部干扰（如网络波动、电磁干扰）具有较强的抵抗力。在网络信号较弱或者存在一定电磁干扰的环境下，系统仍然能正常录音、分析，保证基本功能不受影响。

6.4 高效性

6.4.1 从用户触发录音操作到系统开始正常录音的启动时间不应超过 1 s，应及时捕捉到需要记录的声音。

6.4.2 对已录制的音频文件，系统进行语音转文字、语义分析等操作的处理速度不应超过音频市场的 1.5 倍。

6.5 可扩展性

6.5.1 支持随时添加新的功能模块，如支持新的语言种类、新的语义理解场景。

6.5.2 应不断扩展储存容量，支持与外部存储设备的连接或灵活接入云存储服务。

6.5.3 应支持通过增加服务器节点来扩展系统容量和提升系统计算、数据处理、并发处理等能力。

6.5.4 应支持通过升级硬件配置来提升单个服务器节点的性能。

6.5.5 各个模块应分工明确，负责系统的一个特定功能或服务，便于进行模块的独立升级和维护。

6.6 用户友好性

6.6.1 界面设计

6.6.1.1 应采用简洁的界面设计风格，不应使用过多复杂的元素和混乱的布局，各个功能模块分区明确，用户能快速识别并找到所需功能的入口。

6.6.1.2 选择舒适的色彩搭配方案，不应使用过于刺眼或暗淡的颜色。应采用柔和的色调来降低视觉疲劳，提高用户长时间使用系统的舒适度。不同功能模块应通过不同的颜色进行区分，增强视觉引导性。

6.6.1.3 界面上的文字字体大小应适中，易于阅读。重要信息和提示，应采用较大的字体或加粗、变色等方式突出显示，方便用户快速获取关键内容。

6.6.2 操作流程

6.6.2.1 提供一键启动录音、一键停止、一键播放等一键式操作功能，减少用户的操作步骤。

6.6.2.2 在用户进行操作时，系统应提供智能引导，帮助用户顺利完成复杂的任务。

6.6.2.3 支持快捷键操作，支持用户通过键盘快捷键快速执行常用功能。

6.6.3 反馈与完善

6.6.3.1 在用户进行操作时，系统应检测是否有异常情况并及时给予反馈，帮助用户快速定位问题并解决。

6.6.3.2 提供多种技术支持渠道，如在线客服、电子邮件、电话等，确保用户在遇到问题能及时获得帮助。

7 安全要求

7.1 系统安全

7.1.1 系统防护

7.1.1.1 系统应部署专业的防火墙，对网络流量进行实时监测和过滤。防火墙规则应根据系统的实际需求进行精细配置，阻止未经授权的外部网络访问系统内部资源，防止内部敏感信息泄露到外部网络。

7.1.1.2 配备先进的 IDS/IPS 系统，实时监测网络活动、分析网络数据包，确保能及时发现并识别各种类型的网络攻击行为，如端口扫描、恶意软件传播、拒绝服务攻击等。一旦检测到潜在的入侵行为，系统应立即采取相应的防御措施，如阻断攻击源 IP 地址的访问、发出警报通知管理员等。

7.1.2 系统警示

7.1.2.1 建立全面的安全审计机制，对系统中的所有用户活动进行详细记录和跟踪，并在发生安全事件时能进行事后追溯和分析。审计记录应包括用户登录时间、操作内容、操作结果等关键信息。

7.1.2.2 实时监控系统的运行状态、网络流量、用户行为等关键指标，并对异常情况发出警示。监控系统应设置合理的阈值，当指标超出正常范围时自动触发警报。

7.1.3 系统检测

7.1.3.1 系统启动之前，应对其运行环境进行全面的安全性检测，包括操作系统的安全性配置、网络连接的稳定性和安全性、硬件设备的正常运行状态等。

7.1.3.2 对系统的关键组件和服务进行可靠性检测，确保系统在正常运行过程中不会因组件故障或服务异常而导致数据丢失或系统崩溃。

7.1.4 系统恢复

7.1.4.1 制定完善的灾难恢复计划，在系统遭受意外事件（如自然灾害、硬件故障、网络攻击等）导致系统瘫痪或数据丢失时的恢复策略和流程。灾难恢复计划应包括数据备份策略、系统恢复步骤、RTO 和 RPO 的设定等内容。

7.1.4.2 定期对灾难恢复计划进行测试和演练，确保计划的有效性和可行性。测试应模拟各种可能的灾难场景，验证系统在不同情况下的恢复能力。

7.1.5 访问控制

7.1.5.1 采用强密码策略，要求用户设置复杂密码，并定期更换密码。使用多因素身份验证机制，如结合密码、短信验证码、指纹识别、动态令牌等多种方式，提高用户身份验证的安全性。

7.1.5.2 根据用户的角色和职责，分配相应的访问权限。权限管理应遵循最小权限原则，用户仅被授予完成其工作任务所需的最小权限集。

7.2 数据安全

7.2.1 数据采集

采集用户的语音数据或其他相关信息之前，应以清晰、易懂的方式向用户告知数据采集的目的、范围、使用方式以及可能存在的风险，并获得用户的明确同意。

7.2.2 数据传输

在网络中传输的录音数据及系统数据应进行加密处理，采用 SSL/TLS 协议等高强度的加密算法，确保数据在传输过程中不被未经授权的第三方窃取或篡改。

7.2.3 数据存储

7.2.3.1 系统存储平台实施安全管理，定期更新操作系统和数据库管理系统的安全补丁、配置安全的用户权限、启用日志记录和审计功能，并对存储数据的服务器和存储设备进行用户权限管理，确保只有经过授权的管理员才能访问和管理数据存储系统。

7.2.3.2 建立完善的数据备份策略，定期对数据进行全量和增量备份，并将备份数据存储在异地的安全位置。实施数据备份和恢复验证机制，定期验证备份数据的完整性和可恢复性，确保在数据丢失或损坏时能快速恢复数据。

7.2.4 数据完整

7.2.4.1 使用 CRC 校验、MD5 校验等数据校验算法对采集到的数据进行实时校验，确保数据在传输和存储过程中不发生错误或丢失。

7.2.4.2 采用数字签名、消息认证码等技术手段，对传输的数据进行签名或认证，接收方在收到数据后应验证数据的完整性和真实性。

8 运维要求

8.1 人员管理

8.1.1 组建一支具备多领域专业知识的运维团队，包括系统管理员、网络工程师、安全专家、数据分析师等，并制定详细的岗位职责说明书，规定每个岗位在系统运维中的具体任务、工作流程和目标。

8.1.2 针对可能出现的安全事件和系统故障，定期开展应急响应培训和演练。培训内容包括应急响应流程、安全事件分类与分级、应急处理工具的使用等。演练模拟真实的安全事件场景，如网络攻击、数据泄露等，检验和提升运维团队的应急响应能力和协同作战能力。

8.1.3 持续开展安全意识教育活动，提高运维人员对系统安全重要性的认识，并建立安全规范遵守监督机制，对运维人员的日常操作进行监督和审计。

8.2 性能监测与优化

8.2.1 应持续监测系统的性能指标，如服务器负载、网络带宽占用、内容发布速度等，及时发现性能瓶颈。

8.2.2 应采用性能优化技术，如缓存技术、负载均衡、数据库优化等，提高系统的响应速度和处理能力。

8.2.3 应对系统进行压力测试和模拟高负载场景，确保系统在极端情况下能稳定运行。

8.3 安全漏洞排查与修复

8.3.1 定期进行安全漏洞扫描，包括对系统软件、服务器配置、网络连接等方面的检查，及时发现潜

在的安全风险。

8.3.2 应对发现的安全漏洞进行及时修复，更新安全补丁，加强系统的安全防护能力。

8.3.3 应建立安全事件应急响应机制，在发生安全事件时应迅速采取措施，降低损失。

8.4 功能扩展与升级

8.4.1 应及时了解录音 AI 语音智能分析的新技术、新应用和新需求，评估其对系统的影响和价值。

8.4.2 根据用户需求和行业发展趋势，制定系统功能扩展计划。明确新功能的目标、需求、技术方案和实施时间表，确保功能扩展的有序进行。

8.4.3 定期对系统的现有功能进行评估和优化，提高功能的性能和稳定性。

9 评价与改进

依据第 5 章、第 6 章、第 7 章、第 8 章的要求确定系统的评价内容，定期开展系统功能、性能、安全、运维方面的评价，审查不合格项，并有针对性地采取纠偏措施。

参 考 文 献

- [1] GB/T 36464.1—2020 信息技术 智能语音交互系统 第1部分：通用规范
 - [2] GB/T 42447—2023 信息安全技术 电信领域数据安全指南
-