ICS 35.240.01
CCS L 70

# T/CCUA

中 国 计 算 机 用 户 协 会 团 体 标 准

T/CCUA XXXX—XXXX

# 版本典藏资源智慧展陈 语音交互技术要求

Smart exhibition of archive collection of publications and culture－Technical requirements for speech interaction

（征求意见稿）

（本草案完成时间：2024/4/3）

**在提交反馈意见时，请将您知道的相关专利连同支持性文件一并附上。**

2024－XX－XX 发布　　2024－XX－XX 实施

中国计算机用户协会　　发 布

# 目　　次

# 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。

本文件由中国计算机用户协会提出并归口。

本文件起草单位：中国计算机用户协会创新技术应用分会、中国科学院自动化研究所、武汉大学、武汉数文科技有限公司、中国国家版本馆。

本文件主要起草人：曾智、张桂煊、王晓光、王少华、刘长明、张从龙。

# 版本典藏资源智慧展陈 语音交互技术要求

## 1 范围

本文件规定了版本典藏资源智慧展陈语音交互技术总体框架、语音交互设计、馆藏资源库、语音处理、服务接口和安全技术要求。

本文件适用于版本典藏资源智慧展陈语音交互相关系统或产品的设计、开发、应用和维护。

## 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/T 21024-2007 中文语音合成系统通用技术规范

GB/T 34083-2017中文语音识别互联网服务接口规范

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

GB/T 36464.2-2018信息技术 智能语音交互系统 第2部分：智能家居

## 3 术语和定义

下列术语和定义适用于本文件。

3.1

**版本典藏资源 archive collection of publications and culture**

由公共文化服务机构予以收藏的，古今中外载有中华文明印记的经典版本资源。

3.2

**智慧展陈 smart exhibition**

运用互联网、物联网、大数据、云计算和人工智能等信息，全面感知展陈场馆环境、展陈内容、展陈设施和观众行为等，实现展陈保护精准化、服务个性化、消费网络化、体验数字化、管理智能化。

3.3

**语音交互 speech interaction**

人类和功能单元之间通过语音进行的信息传递和交流活动。

[来源：GB/T 36464.2-2018，定义 3.1]

3.4

**语音合成 speech synthesis**

通过机械的、电子的方法合成人类语言的过程。

[来源：GB/T 21024-2007，定义 3.1]

3.5

**语音识别 speech recognition**

将人类的声音信号转化为文字或者指令的过程。

[来源：GB/T 21023-2007，定义 3.1]

3.6

**命令字识别 command word recognition**

一种基于语音识别语法的语音识别方式，是在语音识别语法规则限定的范围内，对于给定的语音输入，语音识别引擎给出语音识别语法覆盖范围内的文本或拒识作为识别结果。

[来源：GB/T 34083-2017，定义 3.3]

3.7

**连续语音识别 continuous speech recognition**

识别任意的连续语音，并给出相对应的文本。

注：连续语音识别不限制用户说话的词汇、内容和方式，用户可以以任意说的形式输人语音。

[来源：GB/T 34083-2017，定义 3.4]

## 4 系统架构

版本典藏资源智慧展陈语音交互技术对展陈场景下输入设备采集到的语音信号流进行处理，向应用输出语音控制指令，应用得到语音控制指令后进行业务逻辑处理，然后应用业务处理模块根据指令要求向输出设备输出语音或文字信息，从而给用户显示处理结果，系统架构如图 1 所示。

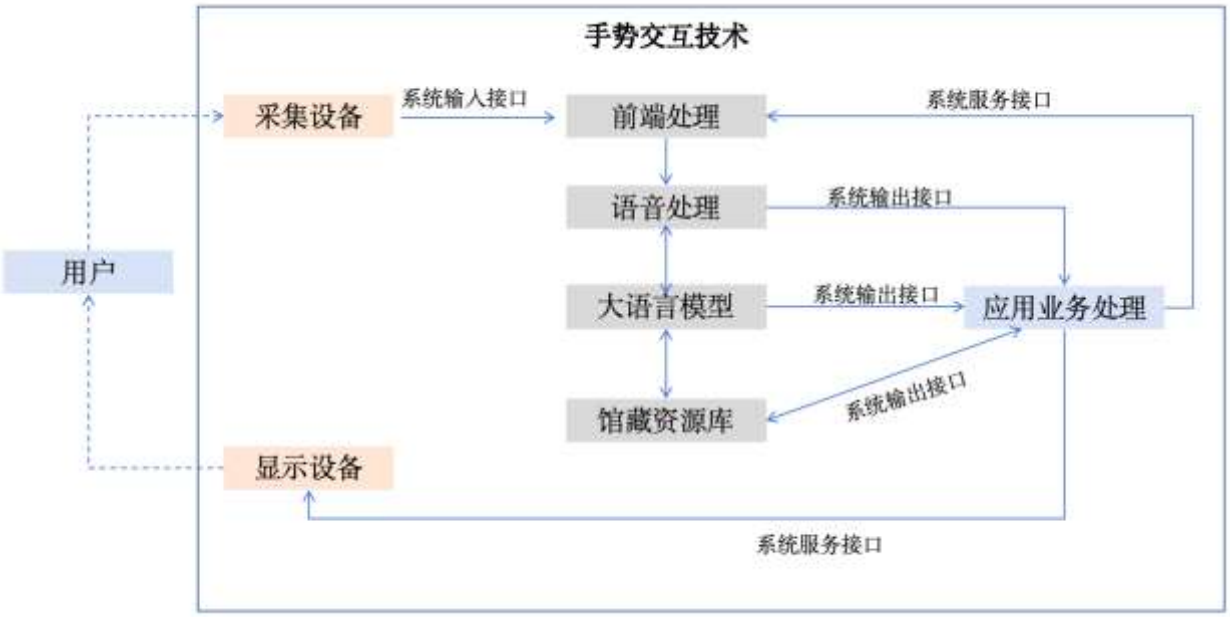


**图 1 系统架构**

版本典藏资源智慧展陈语音交互技术各个模块的接口操作及交互信息流传输方式如下：

a） 采集设备是用户采集用户的语音输入控制指令，包括语音信号输入、输出，所采集的数据通过系统输入接口传输给前端处理模块；

b） 前端处理提供语音唤醒、声源定位、声纹识别、语音增强、格式转换、重采样等功能；

c） 语音处理提供语音识别、语义理解、语音合成、端点检测、语音编解码、全双工交互、情感计算等功能；

d） 馆藏资源库包括系统处理的音频数据和文本数据；

e） 大语言模型作为可选项，可以支持语音识别处理或应用业务处理的智能支持；

f） 语音指令通过大语言模型进行意图识别、清洗、过滤、提取参数，以版本典藏资源库为系统认知，优化组织适合展陈机构相关业务需求指令；

g） 应用业务处理是提供智能应用服务的模块，通过接收语音控制指令，将相关应用服务通过显示设备呈现在用户面前。

## 5 功能要求

### 5.1 交互显示界面

在语音交互界面显示设计方面的相关要求如下：

a） 应具备良好的用户友好性，确保界面直观易用，与用户习惯相符，并始终提供清晰的指引和反馈；

b） 应具备高度的适配性，确保界面能够适配不同的设备和屏幕尺寸，从而在各种环境下都能保持良好的可用性；

c） 应具备出色的可访问性，特别是考虑到不同能力的用户，如视障人士，并为他们提供相应的辅助功能；

d） 整体色调、布局、风格应符合展陈机构展陈物理空间主色调及设计理念。

### 5.2 语音采集设备

#### 5.2.1 语音采集

在语音交互采集设计方面的相关要求如下：

a） 语音采集应具备噪音过滤能力，能够准确识别用户的语音指令，不受背景噪音或多种语言干扰；

b） 语音采集应支持远场拾音，能够在较大展陈空间完成精确定向的拾音；

c） 语音采集设备应具备良好的音频录制和传输能力，如高灵敏麦克风、方向性麦克风等。

#### 5.2.2 语音播报

在语音播报设计方面的相关要求如下：

a） 播报声音应保证语音播报声音清晰流畅，音量适中，语调自然；

b） 播报流程应详细描述播报内容的编排和格式要求，确保信息传达的准确性；

c） 播报声音宜提供至少 2 种语言播报。

#### 5.2.3 输入输出

在语音交互输入输出设计方面的相关要求如下：

a） 交互设计应包括采集、预处理、识别等完整处理流程；

b） 语音输出的生成流程应包括合成、调整、播放等关键环节；

c） 输入和输出应具备健全的错误处理机制，确保在输入或输出过程中出现错误时能够及时处理并给予用户明确的反馈；

d） 输入和输出宜支持实时语音交互处理。

## 6 前端处理

前端处理的相关要求如下：

a） 语音唤醒次数按交互次数计算，应达到 95%以上；

b） 从唤醒到设备给出反馈的时间，响应时间不大于 1.5s；

c） 声源定位在 3m 远场拾音条件下定位准确度的平均误差宜小于 5 度。

## 7 语音处理

### 7.1 指令识别

指令识别的相关要求如下：

a） 指令识别准确率应不小于 85%；
b） 指令识别的实时响应不大于 0.2s，以确保流畅的用户交互体验；
c） 指令种类应符合并适用展陈机构各设备设施产品的操作流程及规范；
d） 语音识别应符合 GB/T 36464.1-2020 中 8.1.1 的一般要求。

### 7.2 语音合成

语音合成的相关要求如下：
a） 语音合成的速度应不大于 1s；
b） 语音合成的质量应确保语音的自然度和流畅度，连贯、不生硬；
c） 应提供多种声音选项，以适应不同场景和用户偏好；
d） 语音合成应符合 GB/T 36464.1-2020 中 8.3 的要求。

### 7.3 语音理解与答复

语音理解与答复的相关要求如下：
a） 语音理解与答复的信息内容来源优先以版本典藏资源库为主；
b） 语音理解与答复的速度应不大于 1s；
c） 语音答复应支持动态生成答复，以提供个性化响应；
d） 语音答复应支持利用上下文信息提供连贯和相关的答复。

## 8 大语言模型

为进一步提升语音交互的体验性，在语音交互环节可以根据环境条件选择使用大语言模型，大语言模型的相关要求如下：
a） 大语言模型宜结合声学模型，生成自然流畅的人类语音，包括语音的音调、韵律、语速等因素，以提高语音合成的质量；
b） 大语言模型宜支持端到端训练的方法，将语音识别、自然语言处理和语音合成等任务整合在一个模型中进行训练；
c） 大语言模型应具备数据安全和隐私保护的能力，确保在语音交互过程中用户的个人信息和隐私的不被泄露和利用。

## 9 馆藏资源库

### 9.1 音频数据

为提升语音交互的准确性，需要馆藏资源库作为语音交互智能学习的语料库，音频数据的相关要求如下：
a） 语料库应支持音频数据的文件格式，如 WAV、MP3 等，以确保数据的兼容性和高质量；
b） 音频数据应包含采样率和位深数据，方便特定环境下音频数据文件的适配选择；
c） 音频数据的存储结构、目录组织、备份和恢复等应有明确的管理要求；
d） 音频数据应包含基本的元数据，如标题、作者、时长等关键信息，元数据符合版本典藏资源元数据规范；
e） 语音识别的输人音频数据格式应符合 GB/T 34083-2017 中表 1 的要求；
f） 语音合成的输出音频数据格式应符合 GB/T 34145-2017 中表 2 的要求。

### 9.2 文本数据

为提升语音交互的准确性，需要馆藏资源库作为语音交互智能学习的语料库，文本数据的相关要求如下：

a） 语料库应支持文本数据的文件格式，如 TXT、XML 等，以确保数据的兼容性和可处理性；

b） 文本数据应包含字符编码标准，如 UTF-8，以确保支持多语言内容；

c） 文本数据的存储结构、目录组织、版本控制等应有明确的管理要求；

d） 文本数据应包含基本的元数据标准，如标题、作者、关键词等关键信息，元数据符合版本典藏资源元数据规范。

## 10 服务接口要求

### 10.1 系统输入接口

#### 10.1.1 语音采集

功能：语音交互采集通过本接口获取输入设备采集的语音数据，通常在语音输入起始后调用。

接口输入：语音数据、时长、语音数据大小。

接口输出：正确执行返回 0，否则返回非 0 值。

#### 10.1.2 添加语音指令

功能：添加一条语音指令，通过指令可执行用户输入语音的相关业务请求。

接口输入：语音指令。

接口输出：正确执行返回语音指令 ID，否则返回非 0 值。

#### 10.1.3 删除语音指令

功能：删除一个语音指令。

接口输入：语音指令 ID。

接口输出：正确执行返回 0，否则返回非 0 值。

#### 10.1.4 设置语音指令执行函数算法

功能：配置语音指令相关执行函数算法，实现相关指令的业务需求交互。

接口输入：指令执行函数对象 ID、函数算法 ID。

接口输出：正确执行返回 0，否则返回非 0 值。

### 10.2 系统输出接口

#### 10.2.1 语音播报

功能：将语音数据通过输出设备进行播报。

接口输入：语音数据 ID。

接口输出：正确执行返回 0，否则返回非 0 值。

#### 10.2.2 音频文件下载

功能：音频文件流下载，可供输出设备缓存到本地播报。

接口输入：音频文件 ID。

接口输出：正确执行返回音频文件流，否则返回非 0 值。

#### 10.2.3 文字合成语音

功能：将文字字符串合成音频文件流。

接口输入：文本字符串。

接口输出：正确执行返回音频文件流，否则返回非 0 值。

#### 10.2.4 语音识别文字

功能：将音频文件流识别提取文本内容。

接口输入：音频文件。

接口输出：正确执行返回文本内容，否则返回非 0 值。

#### 10.2.5 获取大语言模型修正语音意图

功能：使用大语言模型识别语音文本内容，修正用户的请求意图，提升匹配率。

接口输入：文本字符串。

接口输出：正确执行返回文本内容，否则返回非 0 值。

#### 10.2.6 知识图谱知识库查询

功能：根据查询语句检索知识图谱知识库内容，用于反馈版本典藏资源相关的语音知识问答请求。

接口输入：查询语句字符串。

接口输出：正确执行返回知识库三元组内容，否则返回非 0 值。

#### 10.2.7 获取数据资源文件流

功能：根据版本典藏资源文件 ID 获取文件流信息。

接口输入：版本典藏资源文件 ID。

接口输出：正确执行返回数据资源文件流，否则返回非 0 值。

## 11 安全要求

### 11.1 安全和合规

在安全和合规方面，具体要求如下：

a） 语音交互应具备完善的数据加密安全策略和措施；

b） 语音交互应符合所有相关法规、标准和行业规范，并确保合规；

c） 语音交互应具备高可用性设计，包括故障检测、恢复机制、灾备方案等；

d） 语音交互应支持对用户隐私和敏感信息保护的安全措施；

e） 语音交互应支持定期进行安全审计。

### 11.2 访问控制

在访问控制方，具体要求如下：

a） 访问控制应支持用户认证机制，如密码策略、双因素认证等；

b） 访问控制应支持角色的不同访问控制策略；

c） 访问控制应具备访问日志的记录、存储和审查信息。

### 11.3 网络安全

在网络访问方面，具体要求如下：

a） 网络方面应具备网络防火墙和侵入检测能力，及时检测和阻止恶意攻击和未经授权的访问，提高系统的安全性和稳定性；

b） 语音交互模块之间的通信应采用安全的加密协议，如 SSL/TLS，以确保数据在传输过程中的保密性和完整性；

c） 语音交互相关核心模块应定期进行安全扫描和实时监控。