ICS 35.240.99
CCS TP39

团　　体　　标　　准

T/CES XXX-2023

电力虚拟数字人指标要求和评价规范

Electric power virtual digital human index requirements and evaluation specifications

XXXX-XX-XX 发布　　XXXX-XX-XX 实施

中国电工技术学会　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。

本文件由国网信息通信产业集团有限公司提出。

本文件由中国电工技术学会标准工作委员会能源智慧化工作组归口。

本文件起草单位：国网信息通信产业集团有限公司、福建亿榕信息技术有限公司。

本文件主要起草人：李强、王秋琳、庄莉、梁懿、王燕蓉、陈江海、吕志超、张晓东、李炳森、邱镇、苏江文、陈又咏、林钊、俞成强、宋立华、伍臣周、林闽微、林靖、吴佩颖。

本文件为首次发布。

电力虚拟数字人指标要求和评价规范

1 范围

本文件规定了电力虚拟数字人的指标要求和评价规范。

本文件适用于指导电力虚拟数字人组件服务的设计、研发、评估和验收等工作。

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。

信息技术词汇第28部分：虚拟数字人

3 术语和定义

下列术语和定义适用于本文件。

3.1 虚拟数字人 digital human

简称数字人或虚拟人，是基于现实世界设计或通过计算机虚拟生成，再借助真人或计算驱动，在多模态输出设备呈现的拟人化的虚拟形象或虚拟人物，是通过计算机图形学、计算机视觉、智能语音、人工智能和自然语言处理等多种技术集合生成的虚拟任务，可用于数字内容生成和人机交互，帮助提高内容生产效率和用户体验。

3.2 电力虚拟数字人系统 electric power digital human system

利用人机界面、通信设施和数据管理系统等资源提供的服务来实现虚拟数字人能力的应用程序。

3.3 文本驱动 text-driven

支持不同类型的文本到最终数字人视频（包括语音和画面）的驱动方式。

3.4 音频驱动 audio-driven

支持语音输入以及音乐输入，实现对数字人表情及动作的控制的驱动方式。

3.5 视频驱动 video-driven

基于计算机视觉和计算机图形学等技术，通过摄像头捕捉真人的动作，实现对数字人表情及动作的控制的驱动方式。

3.6 模型 model

对客观现实某些方面的抽象。按照对象所需呈现的结构或动势等，通过专用软件完成对象每个表面拓扑结构的数学表示，从而在虚拟三维空间中塑造出的物体。

3.7 渲染 rendering

对一个虚拟场景进行处理，得到符合人类知觉（视觉、听觉、触觉等）输出的过程。

本文件中特指经由虚拟数字人模型数据生成一张或多张二维图像的技术，可具体分为实时渲染技术和离线渲染技术。

3.8 交互 interaction

一种行为，它由一些特定背景下为实现特定目的的对象之间交换的消息组成。

本文件中特指用户与虚拟数字人进行交流互动的行为。

3.9 精细度 fineness degree

描述虚拟数字人模型中各视觉要素的粒度或精度是否能够充分满足渲染需求。

3.10 电力 electricity

本文电力是指电力系统，是由发电、输电、变电、配电和用电等环节组成的电力生产与消费系统。

4 符号和缩略语

下列符号和缩略语适用于本文件。

2D：二维（2-dimensional）

3D：三维（3-dimensional）

AI：人工智能（artificial intelligence）

5 电力虚拟数字人分类

5.1 概述

电力虚拟数字人的分类方法包括但不限于从人物图形资源维度、人物风格、互动形式、用途4个维度进行划分。

5.2 按照人物图形资源维度分类

人物图形资源维度可以分为2D数字人和3D数字人。

a） 2D数字人：人物形象为平面形象，人物图形内容只包含水平的X轴向与垂直的Y轴向的信息的数字人系统。

b） 3D数字人：人物形象为立体形象，人物图形内容包含*X*、*Y*、*Z*三个轴向的信息的数字人系统。

5.3 按照人物风格分类

人物风格可以分为写实风格、科幻风格、个性化风格。

a） 写实风格：人物形象为写实形象(与真人形象相近，具有逼真的人眼立体视觉效果)的数字人系统。

b） 科幻风格：人物形象具有未来感或超现实感（包括机械元素、发光效果或异形的身体结构）、呈现出一种高科技或神秘的气质的数字人系统。

c） 个性化风格：人物形象可以根据品牌形象、文化背景或故事情节来创建独特的数字人形象，以体现特定的价值观或情感表达。

5.4 按照互动形式分类

互动形式可以按照响应时间、驱动方法进行分类。

a） 按照互动形式的响应时间，虚拟数字人分为实时交互和非实时交互虚拟数字人。

1）实时交互虚拟数字人是指以三维实时引擎为技术途径进行构建，可利用不同的驱动方式与其进行实时互动，对于 AI 的生成速度和质量有所要求；

2）非实时交互虚拟数字人是指以传统影视技术为基础进行构建，其运作流程主要依据目标文本对应生成虚拟数字人语音和动画，并合成呈现给用户，但无法进行实时驱动，主要驱动方式包括但不限于通过文本驱动、音频驱动、视频驱动等。

注1:同一个虚拟数字人可支持一种或多种驱动方式，可根据主要驱动方式进行归类。

注2:与实时交互虚拟数字人相比，非实时交互虚拟数字人更加关注资产质量的维度，如生成的图像质量、资产的质量、模型面数、拓扑合理性、是否有法线贴图等。

b） 按照互动形式中所涉及的驱动方法，虚拟数字人分为智能驱动虚拟数字人和真人驱动虚拟数字人。

1）智能驱动虚拟数字人是指通过前置性对声音、动作等内容数据进行标样、整理和学习,使虚拟数字人智能系统对外界输入的多模态信息能够进行自动读取、解析及识别，实现虚拟数字人智能化信息处理与传输，从而决策后续的输出文本、驱动模型生成相应的语音与动作，完成与用户的互动:

2）真人驱动虚拟数字人是指在实现虚拟数字人从静态到动态的转变过程中，需要通过视频监控、动作捕捉等系统提取真人的关键数据信息，将真人的表情、动作实时呈现在虚拟数字人形象上，完成与用户的互动。

5.5 按照用途分类

用途可以按照服务对象、所具有的身份特征进行分类。可以分为电力虚拟客服、电力虚拟主播、电力虚拟讲解员。

a） 电力虚拟客服：主要应用于电力企业客户服务领域，帮助电力企业实现以客户为中心的、智能高效的人性化服务的数字人系统。支持实时回答用户关于电费、用电安全、电力设施维护等方面的问题，提供24小时不间断的服务。

b） 电力虚拟主播：主要应用于电力知识的普及和宣传领域，通过生动有趣的方式向公众传递电力知识，提高公众的电力安全意识和节能意识。

c） 电力虚拟讲解员：主要应用于电力宣传领域，支持对电力展览馆、电力公司展厅、电力科普活动、电力安全与应急演练等详细解说的电力虚拟数字人系统。

d） 电力虚拟专家：主要应用于电力实时指导与支持、安全风险降低以及员工素质提升的电力虚拟数字人系统。

e） 电力虚拟培训师：主要应用于电力科研教学、电力学员培训的电力虚拟数字人系统。

5.6 按照业务分类

按电力业务场景进行分类。可以分为电网调度数字人、设备巡检数字人、客户服务数字人。

a） 电网调度数字人：主要应用于电量统计与分析、电网实时监控、故障预测与诊断、负荷预测与优化、辅助决策支持、培训与模拟、智能报表生成等场景，提高电网运行的效率、安全性和智能化水平。

b） 设备巡检数字人：主要应用于高低压配电室巡检、园区巡检、数据中心巡检、地下综合管廊巡，通过自动化和智能化的方式，提高了巡检的效率和准确性，降低了人力成本。

6 电力虚拟数字人系统

电力虚拟数字人系统可分为图像、语音、动画、交互、多模态输入、多模态输出6个模块。前4个模块与电力虚拟数字人角色本身密切相关，后2个模块用以支撑虚拟数字人驱动与合成显示，如图1所示。

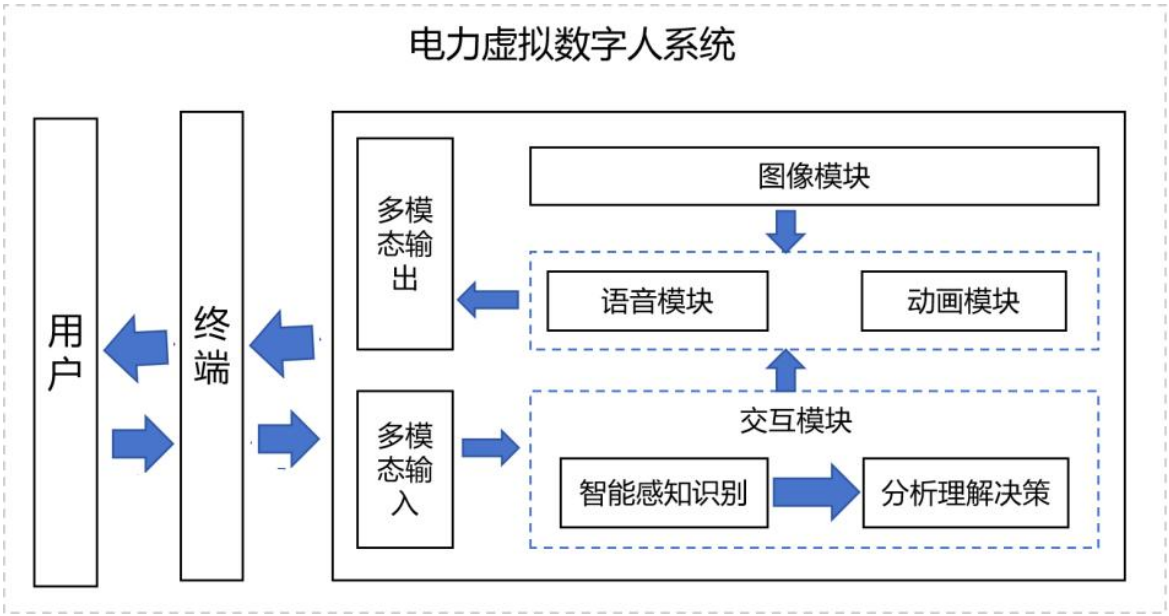

图1 电力虚拟数字人系统框架

a） 图像模块：提供电力虚拟数字人的人物图像。可以是2D、3D数字人等。

b） 语音模块:可生成与文本相对应的电力虚拟数字人角色语音。

c） 动画模块:可基于文本生成相对应的虚拟数字人角色动画，包括面部表情和肢体动作。

d） 交互模块:可使电力虚拟数字人具备识别、感知、分析和理解决策能力，即利用语音识别、自然语言处理、智能感知与识别、智能分析与决策等技术对用户输入的信息进行识别，同时通过意图理解做出后续决策，以决定电力虚拟数字人后续进行的语音和动画播放，驱动电力虚拟数字人开始新一轮的交互。

e） 多模态输入:用于接收用户输入信息。

f） 多模态输出:用于向用户呈现输出信息。

7 电力虚拟数字人指标要求和规范性描述

7.1 图像

7.1.1 人物特性

系统应体现电力虚拟数字人的设定、外形风格等任务特性信息。

其规范性描述，应包括以下信息:

a） 说明数字人的身份设定，如性别、身高、体型、职业定位等;

b） 说明数字人外形的风格，如写实风格、科幻风格、个性化风格等。

7.1.2 完好性

指系统提供的数字人形象的完好程度。若出现下列情况中任意一种或几种则视为有破损。对于2D人物形象:

a） 存在严重的扭曲;

b） 存在严重的马赛克;

c） 存在明显跳帧;

d） 其他破损情况。

对于3D人物模型:

a） 存在严重的变形;

b） 存在严重的穿插;

c） 存在没焊接的点;

d） 存在破面;

e） 其他破损情况。

注:此处的“焊接”特指将3D模型中的多个顶点结合成一个点的过程。

7.1.3 精细度

指系统提供的数字人形象的精细程度，为3D数字人专用指标。

其规范性描述，应包括以下信息:

a） 人物模型的面数；

b） 人物面部细节建模情况，如眼球，晶状体，睫毛，口腔内部结构及毛发等；

c） 人物面部布线结构；

d） 人物贴图分辨率；

e） 人物身体及服饰绑定层级和复杂度；

f） 人物身体、面部及服饰的模型点数；

g） 人物身体、面部及服饰的骨骼数量。

7.1.4 形象舒适性

指系统提供的数字人形象让用户生理上感到舒适的程度。该指标为主观性评估指标，用户根据看到的数字人形象质量,在李克特量表中给出一个主观评分评价质量优劣,1最差~5最优,具体评分规则见表1。

序号	评测维度	描述	1	2	3	4	5
1	好感度	你喜欢该形象的设计吗？	十分不喜欢	不太喜欢	一般	比较喜欢	十分喜欢
2	自然度	该形象是否自然？	十分不自然	不太自然	基本自然	比较自然	十分自然
3	使用意愿	你愿意使用该形象为你服务吗？	十分不愿意	不太愿意	一般	比较愿意	非常愿意

表1 形象舒适性主观评分细则

7.2 语音

7.2.1 发音准确度

指系统从文字合成语音过程中的发音准确程度。发音不准确的表现包括漏音吞音、多余发音、音素错误、音调错误等，相应的性能指标包括发音字准确率和发音句准确率，计算方法如下。

所示:

$$R_{wc} = (1 - \frac{N_{ew}}{N_w}) \times 100\%$$

式中：

N_{w}——文本总字数，单位为个；

N_{ew}——发音错误字数（多种发音错误字数之和），单位为个；

R_{wc}——发音字准确率。

$$R_{\mathrm{sc}} = (1 - \frac{N_{\mathrm{es}}}{N_{\mathrm{x}}}) \times 100\%$$

式中：

N_x——文本总句数，单位为个；

N_{es}——发音错误句数，单位为个；

R_{sc}——发音句准确率。

7.2.2 韵律准确度

指系统语音合成过程中的韵律准确程度。韵律包括停顿断句、音高、音长、音量、重音位置焦点位置等因素，对应了焦点发音、问句语调、感叹句语调等自然发音规律，此处只考察停顿断句，具体可参考ACL“黄金”标准分词文件。其指标的计算方法如下所示:

$$\mathrm{R_{pc}} = \frac{N_{\mathrm{pc}}}{N_x} \times 100\%$$

式中：

N_{pc}——停顿正确用例数，单位为个；

N_x——总用例数，单位为个；

R_{pc}——韵律准确率。

7.2.3 语音逼真度

指系统合成的语音清晰、自然，能够模拟真实的人类语音，让用户生理上感到舒适。该指标为主观性评估指标，用户根据听到的数字人声音质量，在李克特量表中给出一个主观评分评价质量优劣，1最差~5最优，具体评分规则见表2。

序号	评测维度	描述	1	2	3	4	5
1	语音语调	整体语音是否标	十分标准	比较标准	基　本	个别标准	十分不标准

		准？			标准		
2		发音吐字是否清晰？	十分清晰	比较清晰	基本清晰	不太清晰	十分不清晰
3		断词断句、停顿是否恰当？	十分恰当	比较恰当	基本恰当	不太恰当	很不恰当
4		语气语调是否自然？	十分自然	比较自然	基本自然	不太自然	十分不自然
5		重读发音是否得当？	十分恰当	比较恰当	基本恰当	不太恰当	很不恰当
6		语速表达是否恰当？	十分恰当	比较恰当	基本恰当	不太恰当	很不恰当
7	流畅连贯度	语音表达是否流利？	十分自然	比较自然	基本自然	不太自然	十分不自然
8	情绪饱满度	按照文本语义和内容，情绪表达是否恰当？	十分恰当	比较恰当	基本恰当	不太恰当	很不恰当
9	拟人舒适度	声音拟人程度是否和真人一样?	完全无法区分	比较相似，与真人语音有席位区别	基本相似	不太一样	完全不一样
10		聆听该声音时，感受是否愉悦？	十分愉悦	比较愉悦	一般	不太愉悦	十分不愉悦
11		你愿意使用该声音为你服务吗？	十分愿意	比较愿意	一般	不太愿意	十分不愿意

表2 语音逼真度主观评分细则

7.3 动画

7.3.1 动作契合度

指系统中数字人动作与当下语境的契合度。电力虚拟数字人动作类型及契合度体现见表3。

序号	动作类型	契合度体现
1	嘴唇动作	● 作为发音器官，嘴唇能够根据输入语言信息（语音或文本）自动生成嘴唇动画参数。 1）嘴形满足单帧时刻同步音素发音的需求； 2）嘴形能满足前后若干相邻帧时刻音素对当前时刻嘴形的影响（耦合性） ● 作为语义表达渠道，嘴唇动作能够自主地根据内心表达需要（比如：情绪或意图）生成合理的嘴型。 1）基础的嘴唇单元包括嘴唇上拉、急剧嘴角上拉、嘴角收紧、嘴角下拉、下唇下拉、下唇上推、噘嘴、嘴角拉伸、嘴唇外翻、嘴唇收紧、嘴唇挤压、张嘴和吸唇。
2	眉毛与眼皮动作	● 作为非语义表达渠道，眉毛与眼皮动作能够自主地展示模仿真实人类的生理需求（如：眨眼）：在数字人说话时，眉毛与眼皮动作符合语音的时序韵律特征。

		● 作为语义表达渠道，眉毛与眼皮动作能够自主地展示与内心状态一致的情绪或意图。
3	眼球动作	● 作为生理需求，眼球能够自主地模仿真实人类的眼球旋转； ● 作为语义表达渠道，眼球动作能够自主地反映出内心状态（比如:情绪或意图）。
4	头旋转动作	● 作为非语义表达渠道，头的旋转动作能够自主地表达模仿真实人类的生理动作:在数字人说话时，头的旋转动作符合语音的时序韵律特征； ● 作为语义表达渠道，头的旋转动作能够自主地表达符合场景需求的语义信息，比如:点头和摇头。 1）基础的头旋转的动作单元包括头左转、头右转、头向上、头向下、头左倾斜、头右倾斜、头前倾、头后仰、头上下摆动、头左右摆动、头上扬再左/右倾。
5	上身肢体动作[包括躯干关节（旋转）动作、大臂、小臂和手掌]	● 作为韵律节奏动作，在数字人说话时，上身肢体动作符合语音的时序律特征； ● 作为指示功能性动作，数字人能够自主地通过上身肢体动作表达人物关系、空间位置、时间顺序、抽象概念等的作用； ● 作为符号功能性动作，数字人能够自主地通过上身肢体动作比划出实体的属性或行为，来描绘对应的实体或动作； ● 作为比喻功能性动作，数字人能够自主地通过上身肢体作构建一个空间，来表示一个抽象性的概念； ● 作为操作功能性动作，数字人能够自主地通过上身肢体动作有效且自然地操作物体，模拟真实人类进行相关生产生活。
6	下身肢体动作（包括大腿、小腿和脚掌）	● 作为平衡功能性动作，数字人能够模拟真实人类下半身肌肉对抗地球重力，有效且自然地维持身体的平衡。 ● 作为位移功能性动作，数字人能够模拟真实人类下半身肢体动作（如:走、跑、跳等方式）实现身体位移，并可进行多种行动方式的自然切换，同时也能够体现出内心的情绪状态（如:不同情绪状态下，走路/跑步姿态略有区别）。
7	全身动作	● 多模态动作协调一致，全身（包括嘴唇、眉毛与眼皮、眼球、头旋转、上身肢体和下身肢体）共同协作完成表达功能或履行某种功能。

表3 电力虚拟数字人动作类型及契合度体现

7.3.2 动作舒适性

指系统中的数字人动作让用户生理上感到舒适的程度。该指标为主观性评估指标，用户根据看到的数字人形象质量，给出一个主观评分评价质量优劣，1最差5最优，具体评分规则见表4。

序号	评测维度	描述	1	2	3	4	5
1	口型匹配度	口型与发音匹配吗？	完全不匹配	不太匹配	基本匹配	比较匹配	完全匹配
2	面部表情自然度	面部表情是否自然？	十分不自然	不太自然	基本自然	比较自然	十分自然
3	肢体动作自然度	肢体动作是否自然？	十分不自然	不太自然	基本自然	比较自然	十分自然

表4 动作舒适性主观评分规则

7.4 交互处理

7.4.1 虚拟数字人交互逻辑架构

虚拟数字人交互逻辑由多模态输入、感知与理解、多模态输出、人工四个模块组成，系统基本架构如图2所示。

a) 多模态输入:该模块用于接收用户的输入信息，支持文字、语音、图像、触控等多种输入方式；

b) 感知与理解:该模块通过语音感知、视觉感知等技术对多模态输入信息进行处理和感知，使用自然语言处理等技术理解用户意图,并根据用户当前意图决定数字人后续的语音和动作；

c) 多模态输出:该模块用于将交互结果通过语音、动画或其他形式进行输出；

d) 人工:该模块用于实时接入用户的人工请求，为用户提供人工服务。人工可选择性采纳感知与理解模块的处理结果，即数字人获取输入后，将输入信息转人工处理，不进入智能感知与理解模块；或数字人将输入信息进行感知与理解后，交由人工处理。

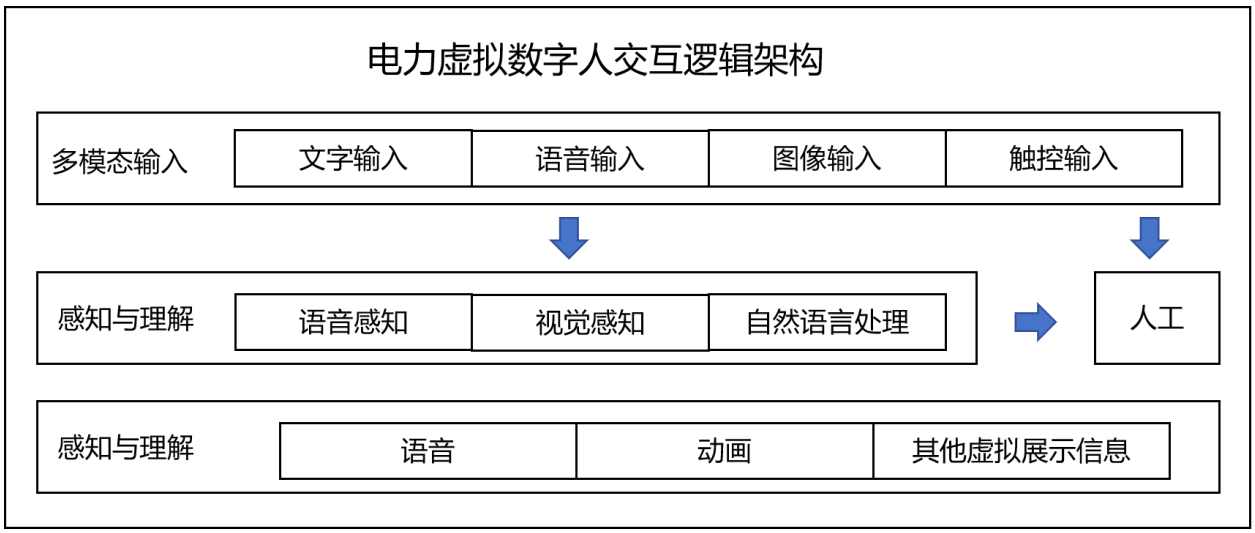

图2 电力虚拟数字人交互逻辑框架

7.4.2 交互指标

7.4.2.1 语音识别准确性

指系统对用户进行语音识别的性能表现，通过计算识别结果中每个单词的错误情况，评估数字人的语音识别准确性。包括字错误率（WER）、句准确率。

字错误率是衡量语音识别系统性能的关键指标。它表示实际错误的单词数除以总的单词数。其指标的计算方法如下所示：

$$WER = （S + D + 1）/N$$

式中：

S——替换掉的字符数量；

D——删除掉的字符数量；

I——额外插入的字符数量；

N——原字符串的长度。

句级准确率是指识别正确的句子数与总句子数的比值。通过对比数字人的识别结果与原始句子，统计识别正确的句子数量，然后除以总句子数，即可得到句级准确率。

7.4.2.2 交互任务执行时间

指系统与用户进行一轮交互的平均响应时间。其指标的计算方法如下所示：

$$\overline{T} = \frac{\sum_{i=1}^{N}(Ts_i - Te_i)}{N}$$

式中：

i——表示第i轮交互；

Ts_i——系统开始反馈的时间点，单位为秒（s）；

Te_i——用户输入信息结束的时间点，单位为秒（s）；

N——测试总次数；

$\overline{T}$——平均交互响应时间，单位为秒（s）。

7.4.2.3 用户交互体验满意度

本指标为主观性评估指标,用户基于对参评系统的交互体验进行满意度评估,参照表5从对话交互准确度、风格一致性两个方面分别给出一个主观评分。

序号	评测维度	描述	1	2	3	4	5
1	交互准确度	咨询问答情景下,系统是否能够准确地解答用户问题?	完全不准确	不太准确	基本准确	比较准确	完全准确
2		任务执行情景下,系统是否能准确地联系上下文帮助用户完成任务?	完全不准确	不太准确	基本准确	比较准确	完全准确
3		系统是否能准确识别打断?	完全不满意	不太满意	基本满意	比较满意	完全满意
4	风格一致性	在交互过程中,系统中的数字人是否保持前后风格一致,符合设定人物性格特征?	完全不一致	不太一致	基本一致	比较一致	完全一致

表 5用户体验主观性评估指标

7.5 多模态输入

7.5.1 多模态输入方式

用于考核系统支持的输入方式种类，包括但不限于文字、语音、图像、触控等。

7.6 多模态输出

7.6.1 视频合成实时率

指系统的视频合成实时率，即视频合成耗时与输出视频时长比值。

7.6.2 流畅度

用于考核系统生成数字人视频的流畅度，主要通过视频帧率，即FPS值(单位:帧/秒)来评估。

7.6.3 画面准确率

用于考核系统生成固定帧数视频时画面的准确率,若出现跳帧、卡顿等错误均视为画面不准确。

7.6.4 音视频匹配度

用于考核系统生成固定时长(单位:s)视频时音视频的匹配度，若出现口型多余、缺失，音频提前、延迟等错位均视为音视频不匹配。

7.6.5 多模态输出方式

用于考核系统支持的输出方式种类，包括但不限于手机、电视、投影、LED显示、裸眼立体、VR、AR显示等。

团 体 标 准

电力虚拟数字人指标要求和评价规范

T/CES 133—2024

2024 年 4 月第一版

*

北京西城区莲花池东路 102 号天莲大厦 10 层

邮政编码：100055

网址：http://ces.org.cn/html/category/17060132-1.htm

电话：010-63256990　63256997