ICS 33.160.99
M72

团 体 标 准

T/AIIA 001-2021

支持语音和视觉交互的虚拟数字人技术规范

Technical specifications for virtual digital humans supporting voice and visual interaction

2021-12-30 发布 2021-12-31 实施

深圳市人工智能产业协会 发布

目　　次

目　　次.....I
前　　言.....I
1　范围.....1
2　规范性引用文件.....1
3　术语和定义.....1
4　系统逻辑结构.....3
5　技术要求.....3
5.1　概述.....3
5.2　声学性能要求.....3
5.3　显示性能要求.....4
5.4　语音交互性能要求.....4
5.5　视觉交互识别技术要求.....5
6　测试方法.....5
6.1　语音交互测试要求.....5
6.2　视觉交互测试要求.....6
6.3　声学性能测试.....7
6.4　显示性能测试.....7
6.5　语音交互测试.....7
6.6　视觉交互测试.....8
附录　A.....10
A.1 输入输出要求.....10
A.2 测试集构建方法.....10
A.3 测试场景设置.....12

前　言

本文件按照GB/T 1.1-2020《标准化工作导则 第1部分：标准的结构和编写》给出的规则起草。

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这些专利的责任。

本文件由深圳市人工智能产业协会（Shenzhen Artificial Intelligence Industry Association）标准化委员会提出并归口。

本文件起草单位：深圳市人工智能产业协会、深圳市8K超高清视频产业协作联盟、深圳赛西信息技术有限公司、深圳市优必选科技股份有限公司、深圳市金大智能创新科技有限公司、科大讯飞股份有限公司、深圳市博乐信息技术有限公司、锋睿领创（珠海）科技有限公司、深圳魔耳智能声学科技有限公司、深圳欧博思智能科技有限公司、蓝亚技术服务（深圳）有限公司、深圳酷酷科技有限公司、深圳光子晶体科技有限公司、深圳奥尼电子股份有限公司、东莞市律普电子科技有限公司、杭州汇萃智能科技有限公司。

本文件主要起草人：范丛明、史培宁、杨紫晴、张哲、黄东延、丁万、王茂林、张 斌、朱文臻、何良雨、叶威廷、胡亚莉、郑小霖、郑港、魏贤华、王周余、周柔刚、吕刚、杨诗虹。

本文件为首次发布。

支持语音和视觉交互的虚拟数字人技术规范

1 范围

本文件规定了支持语音和视觉进行交互的虚拟数字人的技术要求和测量方法。
本文件适用于支持语音和视觉交互的虚拟数字人及其系统的研发、设计和测试。

2 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是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 3096-2008 声环境质量标准
GB/T 5271.29-2006 信息技术 词汇 第29部分：人工智能 语音识别与合成
GB/T 12060.5-2011 声系统设备 第5部分：扬声器主要性能测试方法
GB/T 12060.16-2017 声系统设备 第16部分：通过语音传输指数客观评价言语可懂度
GB/T 14277-2013 音频组合设备通用规范
GB/T 21023-2007 中文语音识别系统通用技术规范
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 34145-2017 中文语音合成互联网服务接口规范
GB/T 35273-2020 信息安全技术 个人信息安全规范
GB/T 35312-2017 中文语音识别终端服务接口规范
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
GB/T 36464.3-2018 信息技术 智能语音交互系统 第3部分：智能客服
GB/T 36464.4-2018 信息技术 智能语音交互系统 第4部分：移动终端
GB/T 36464.5-2018 信息技术 智能语音交互系统 第5部分：车载终端
SJ/T 11380-2008 自动声纹识别（说话人识别）技术规范
SJ/T 11540-2015 有源扬声器通用规范
GB/T 38665.1-2020 信息技术 手势交互系统 第1部分：通用技术要求
GB/T 38665.2-2020 信息技术 手势交互系统 第 2 部分：系统外部接口
SJ/T 11348-2016 平板电视显示性能测量方法
GB/T 35273—2020《信息安全技术个人信息安全规范》

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 virtual digital human

是基于计算机视觉和语音合成等技术，进行形象、声音、动作等的模型训练后，可以生成虚拟人像并与使用者交互的设备。

3.2

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。

[GB/T 36464.2-2018，定义3.1]

3.3

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。
[GB/T 21023—2007，定义3.1]

3.4

语音合成 speech synthesis

将给定的文本转换成与之对应的语音的过程。
[GB/T 34145—2017，定义3.1]

3.5

自然语言理解 natural language understanding

让计算机能够理解自然语言文本中蕴含的含义及意图的过程。

3.6

语音唤醒 speech wake-up;voice trigger

处于音频流监听状态的语音交互系统，在检测到特定的特征或事件出现后，切换到命令词识别、连续语音识别等其他处理状态的过程。
[GB/T 36464.2-2018，定义3.13]

3.7

误唤醒 false wake-up

语音交互系统处于音频流监听状态，无音频流或者音频流中没有出现唤醒所需的特征或事件时，语音唤醒系统被唤醒的现象。

[改写GB/T 36464.2-2018，定义3.14]

3.8

噪声 noise

语音采集过程中，采集到的能干扰对目标语音信号的识别、理解或处理的信号。

3.9

声纹 voiceprint

指语音中所蕴含的、能表征和标识特定说话人的独有的特性或特征。
[SJ/T 11380—2008，定义3.1.1]

3.10

声纹识别 voiceprint recognition

根据待识别语音的声纹特征识别该段语音所对应的说话人的过程。
[SJ/T 11380—2008，定义3.1.6]

3.11

麦克风阵列 microphone array

由具有确定空间拓扑结构的多个麦克风组成的，对信号的空间特性进行采样并处理的系统。

3.12

语音打断 speech interruption

语音交互系统在播放声音的过程中，当语音采集设备检测到有效语音输入时，终端播放声音，转到语音识别等其他处理过程。

[GB/T 36464.2-2018，定义3.18]

3.13

隐私标签 privacy label

由厂商或者开放平台应用定义的涉及使用者私密信息的数据，对该类型数据加以标识的标签。

3.14

手势 gesture

用户利用上肢（包括手部和手臂）表达交互意图时，所执行的具体姿态或动作。

3.15

手势识别 gesture recognition

从输入的手势图像/视频数据确定用户手势状态。

3.16

人体姿态估计 pose estimation

从2D图像中，预测人体的13个关节点和5个头部关键点的图像坐标。13个人体关节点的定义为：1 脖子（neck）、2 右肩（right shoudler）、3 右肘（right_elbow）、4 右腕（right_wrist）、5 左肩（left_shoudler）、6 左肘（left_elbow）、7 左腕（left_wrist）、8 右髋（right_hip）、9 右膝（right_knee）、10 右踝（right_ankle）、11 左髋（left_hip）、12 左膝（left_knee）、13 左踝（left_ankle）；5个头部关键点的定义为：14 鼻子（nose）、15 右眼（right_eye）、16 左眼（left_eye）、17 右耳（right_ear）、18 左耳（left_ear）。

4 系统逻辑结构

支持语音和视觉交互的虚拟数字人应（或者可）由人物生成、人物表达、合成显示、识别感知、分析决策等模块构成：

表 1 支持语音和视觉交互的虚拟数字人逻辑结构

数字虚拟人	2D 数字虚拟人	3D 数字虚拟人
人物生成	人物图像分布建模	人物建模等
人物表达	语音生成、动画生成(驱动、渲染)等	
合成显示	终端显示技术	
识别感知	语音语义识别、人脸识别、动作识别等	
分析决策	知识库、对话管理等	

5 技术要求

5.1 概述

支持语音和视觉交互的虚拟数字人应具备语音交互和视觉交互功能；支持语音和视觉交互的虚拟数字人应提供多种基于网络的应用和服务；具备视觉交互功能，包括手势、表情、步态、人体姿态等交互方式；具备语音交互功能。

5.2 声学性能要求

产品的声学性能参数及要求应满足表2要求。

表 2　声学性能要求

序号	项目		单位	性能要求
1	额定声频率响应范围		Hz	由产品标准规定
2	幅频响应差（L&R 或 FL&FR）		dB	≤3
3	声压总谐波失真	250 Hz～6300Hz	%	≤7
		对于超过允许值，但峰宽小于或等于 1/3oct 的独立的失真峰，允许不超过 3 个；但不允许有大于 1/3oct 的失真峰		
4	噪声声级		dB（A）	≤30

5.3　显示性能要求

产品的显示性能参数及要求应满足表3要求。

表 3　显示性能要求

序号	项目		单位	性能要求
1	亮度		cd/m^2	150
2	对比度		倍	≥200:1
3	亮度均匀性		%	≥70
4	相关色温		K	由产品标准规定
5	色域覆盖率		%	NTSC% ≥70%
6	白平衡误差	$\Delta u'$	—	不劣于±0.020
		$\Delta v'$		不劣于±0.020
7	亮度可视角	水平	（°）	≥80
		垂直		≥60
8	色度可视角		（°）	由产品标准规定
9	固有分辨力		像素数	由产品标准规定
10	漏光		cd/m^2	≤1

5.4　语音交互性能要求

产品的语音交互性能参数及要求应满足表4要求。

表 4　语音交互性能要求

序号	项目		单位	性能要求
1	语音识别	字准确率	%	≥85
2	声纹识别		%	≥80
3	语音打断	语音打断成功率	%	≥85
4	语音合成	音色		由产品标准规定
5		语种	—	支持中文、英文； 支持中英文混读。
6		MOS 分	—	≥3.5
7	语音唤醒	误唤醒频度	次/天	≤1
8		唤醒正确率	%	≥85

9	语音交互	交互拒识率	%	≤15
10		平均响应时间	s	≤1.7
11		最大响应时间	s	≤2
12	休眠要求		—	应具备休眠键，并且明确提示用户音箱是否处于休眠状态。 在休眠状态下，音箱应停止拾音。

5.5 视觉交互识别技术要求

厂商无特殊说明应符合以下指标要求：
识别距离范围：　　1.5m ～ 4m。
识别角度范围：　　左右角度均不小于30°。
识别率：　　大于或等于90%。
识别响应时间：　　不大于2s。
体感识别应满足表5所示的要求。

表 5 体感识别性能要求

序号	项目	单位	性能需求
1	摄像设备	像素	RGB不低于640x480，D不低于640x480。
2	处理速度	帧/每秒	不低于30
3	手势检测	%	以IoU1=0.5为阈值，手部区域检测的召回率不低于95
4	手势识别	%	单类召回率不低于90
5	人体姿态估计	%	PCKh@0.53不低于90

注[1]IoU 指预测标记框与真实标记框的交并比（Intersection Over Union），用于定义图像中物体的预测框是否为预测正确。IoU>0.5时定义预测的手部标记框为正确。

注[2]NM指Normal Case下的步态识别；BG指Subjects Carrying Bags下的步态识别；CL指Subjects Wearing Coats or Jackets 状态下的步态识别。

注[3]PCK是Percentage of Correct Keypoints的缩写，h@0.5指以图像中用户头部标记框的对角线长度（像素单位）为阈值，判断预测的人体关键点是否准确。具体计算方式为：定义头部标记框的对角线长度（像素单位）为h，若预测点与真实点的欧几里得距离（像素单位）<0.5x(0.6xh)，则定义为该关键点预测正确，否则预测错误。

5.6 隐私安全技术要求

被测终端具备用户隐私保护相关能力，声明收集和使用用户数据的范围并征求用户的授权同意，符合GB/T 35273—2020相关规定。

6 测试方法

6.1 语音交互测试要求

6.1.1 测试语料要求

测试语料应覆盖被测系统的核心词汇，并从被测系统词汇量覆盖、业务覆盖、音节覆盖，以及常用性角度进行设计，具体要求应按GB/T 21023-2007执行。

6.1.2 语音测试集要求

语音测试集应符合以下要求：

a）语音识别准确率测试应至少由男女老少各 25 名发音人进行录制，语音唤醒功能测试应至少由 100 名发音人录制，具体要求应按 GB/T 21023-2007 执行；
b）声纹识别测试应至少由 50 名发音人录制验证，具体要求应按 GB/T 21023-2007 执行。

6.1.3 环境噪声要求

表 6　语音识别测试环境要求

家居环境	房间门窗	电视（可选）	抽油烟机（可选）	空调（可选）	被测产品位置处的环境混响要求 s	信噪比 dB	被测产品位置处的环境噪声声压级 dB(A)
低噪	关	关	关	关	混响时间 0.2～0.3	≥20	≤45

6.1.4　测试设备要求

测试设备要求如下：

a)　声音重放设备：由信号发生器、功率放大器和扬声器组成。应满足以下条件：

——功率放大器和扬声器产生的声源幅度非线性影响值应足够小；

——声音重放设备产生的本底噪声应足够小。

b)　声压测试设备：声级计。

c)　识别时间测试设备：宜采用示波器或高速相机测试识别时间，或者开发自动化软件进行测试。

6.1.5　拾音距离要求

测试所描述的拾音距离为通常为1 m、3 m和5 m，使用的测试距离应在测试报告中说明。

6.1.6　语音交互测试布置

推荐按照图1布置测试，推荐在A=90° 且B=150° 、A=60° 且B=150° 下测试，使用的空间布置应在报告中说明。

当声源与待测样品的空间布置（包括但不限于角度、高度、摆放位置等）对测试结果有影响时，应改变空间布置重复测试，并提供不同布置下的测试结果，摆放高度建议1.5米。

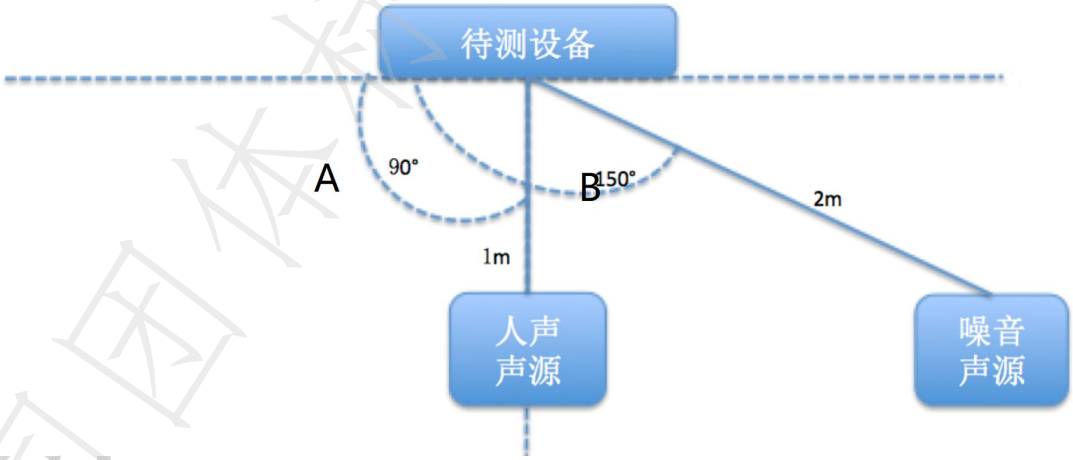

图 1　测试布置图

6.1.7　视觉交互测试布置

推荐按照图1布置测试，推荐在A=90° 且B=150° 、A=60° 且B=150° 下测试，使用的空间布置应在报告中说明。

6.2　视觉交互测试要求

6.2.1　工作条件

支持语音和视觉交互的虚拟数字人应处于工作状态，并处于手势可操控界面。应确保支持语音和视觉交互的虚拟数字人软、硬件系统工作正常。摄像头、手势识别功能应处于工作状态。

6.2.2 环境条件

环境光强为200 lx～250 lx，环境光分布均匀，没有强烈的背光光源照射。

6.2.3 测试场地

测试应在室内进行，要求室内空间长度超过5 m、宽度超过4 m，背景色为50%灰。

6.2.4 测试位置

被测支持语音和视觉交互的虚拟数字人屏幕中心距离地面高度1.2 m。摄像头应朝向虚拟数字人正前方（外置摄像头放在虚拟数字人中心线上部），测试中可以调节摄像头的拍摄角度保证用户躯干及手部在摄像头中可见。

6.2.5 测试辅助仪器

卷尺、角度仪、数字秒表、视频记录仪。

6.3 声学性能测试

支持语音和视觉交互的虚拟数字人的声学性能参照SJ/T 11540-2015、GB/T 12060.5-2011、GB/T 14277-2013测试。

6.4 显示性能测试

支持语音和视觉交互的虚拟数字人的显示性能参照SJ/T 11348-2016测试。

6.5 语音交互测试

6.5.1 语音识别

在低噪测试场景下，如下设置播测音源的音量：距离待测设备1 m处，播放唤醒语料或识别语料，在待测设备麦克风处测的平均声压级为65 dB(A)，以此为基准音量。

将虚拟数字人被测系统调至待命状态，在拾音距离内使用回放设备播放语音识别测试语料，记录低噪环境下虚拟数字人被测系统的识别结果，并与预期结果进行比对。

有必要时可以在其他信噪比下执行测试，并在报告中说明具体的测试安排。

6.5.2 语音合成

选取100个体验人员，男女各50人，通过对虚拟数字人被测系统语音唤醒或语音识别命令的反馈，测听合成语音同真人语音在音质、可懂度和自然度等方面的差异，并以平均意见得分（MOS分）量化进行主观测评，记录平均结果。

6.5.3 交互拒识率

交互拒识率测试方法如下：

a) 将虚拟数字人被测系统调至待命状态，使用回放设备在拾音距离内播放语音识别测试语料，记录当次语音交互会话是否成功和有效；
b) 在低噪条件下按上述步骤完成测试，计算各测试场景下的语音交互拒识率。

6.5.4 响应时间

响应时间测试方法如下：

a) 准备虚拟数字人测试设备及其网络环境，开启被测系统拾音功能，用回放设备在拾音距离内播放语音识别测试语料，记录当次成功的语音交互会话测试录音输入完成的时刻 t_e 和返回服务结果的时刻 t_r，计算当次语音交互会话的响应时间；
b) 分别在表 2 所示网络环境下，按上述步骤完成测试，然后计算平均识别时间、平均实时系数和最大响应时间。

实时系数测试遵照GB/T 21023-2007中5.3要求。

最大响应时间测试遵照厂商产品质量要求。

6.5.5 语音唤醒

语音唤醒测试包括唤醒正确率测试和误唤醒频度测试，方法如下：

a) 唤醒测试：在低噪测试场景下，将虚拟数字人被测系统调至待命状态，使用回放设备播放唤醒测试语料，记录被测系统是否给出正确响应，统计各场景下的唤醒正确率，其计算方法见公式（1）；

$$P_r = \frac{N_{sw}}{N_w} \times 100\% \quad \text{.....(1)}$$

式中：

P_r——唤醒正确率；

N_{sw}——正确唤醒次数；

N_w——总唤醒次数。

b) 误唤醒测试：将虚拟数字人被测系统调至待命状态，测试 24 h，记录被测系统被误唤醒次数，统计误唤醒频度。

6.5.6 语音打断

设定人在距离待测设备1m处发声，统一在待测设备的麦克风处测量得到人声音量70 dB，待测设备的内噪声85 dB，信回比-15 dB，并以此为音量基准在不同距离重复进行语音打断测试。

使用定制的语音打断的唤醒词或命令词集，在待测设备播放状态或者语音交互状态中，使用统一的内噪声素材，进行语音打断，按公式（2）计算语音打断成功率。

$$P_i = \frac{N_i}{N} \times 100\% \quad \text{.....（2）}$$

式中：

P_i——语音打断成功率；

N——交互内容中需要执行打断操作的次数；

N_i——被语音交互系统正确响应的次数。

在声源距离待测样品1m距离处，语音打断成功率应不小于85%（B级）或90%(A级）。

6.5.7 声纹识别

根据产品使用说明设置声纹并验证其能正常使用。

6.6 视觉交互测试

6.6.1 识别距离范围测试

6.6.1.1 测试步骤

测试步骤如下：

a) 被测支持语音和视觉交互的虚拟数字人放在测试工装车上， 将和摄像头调整到测试所需的一般状态。

b) 测试人员按照图 1 所示，在地面标出起始位置、最小识别测试位置和最大识别测试位置。

c） 测试人员站到最小识别位置，测试厂商说明书中规定的所有动作各一次，测试期间手势的要求动作全部被正确识别则按照步骤 d）进行测试，如果测试手势出现无法识别的情况，则按照步骤 e）进行测试。

d） 以不大于 10 cm/s 的速度向前移动测试工装车，同时进行测试厂商说明书中规定的某一简单手势动作，直至手势无法被正常识别时，停止测试工装车的移动，测量支持语音和视觉交互的虚拟数字人屏幕与测试人员之间的距离为测试的最小识别距离 L_{min}。

e） 以不大于 10 cm/s 的速度向后移动测试工装车，同时进行测试不被识别的手势动作，直至手势被正常识别时，停止测试工装车的移动，测量支持语音和视觉交互的虚拟数字人屏幕与测试人员之间的距离为测试的最小识别距离 L_{min}。

f） 将测试工装车恢复至起始位置，测试人员站到最大识别位置，测试厂商说明书中规定的所有手势各一次，测试期间手势的要求动作全部被正确识别则按照步骤 g）进行测试，如果测试手势出现无法识别的情况，则按照步骤 h）进行测试。

g） 以不大于 10 cm/s 的速度向后移动测试工装车，同时进行测试厂商说明书中规定的某一简单手势动作，直至手势无法被正常识别时，停止测试工装车的移动，测量支持语音和视觉交互的虚拟数字人屏幕与测试人员之间的距离为测试的最大识别距离 L_{max}。

h） 以不大于 10 cm/s 的速度向前移动测试工装车，同时进行测试不被识别的手势动作，直至手势被正常识别时，停止测试工装车的移动，测量支持语音和视觉交互的虚拟数字人屏幕与测试人员之间的距离为测试的最大识别距离 L_{max}。

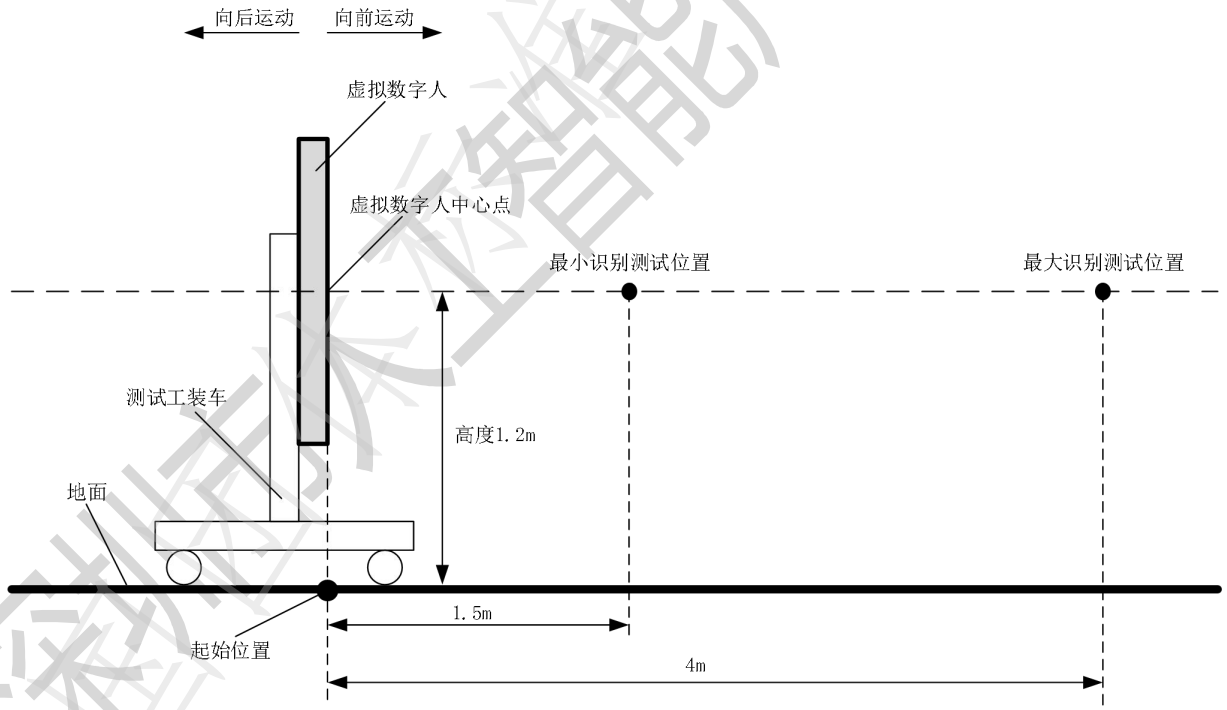

图 1　识别距离范围测试图示

6.6.1.1　测试结果

识别距离合格范围：L_{min}≤1.5 m 且 L_{max}≥4 m。

6.6.2　识别角度范围

6.6.2.1　测试步骤

测试步骤如下：

a) 被测支持语音和视觉交互的虚拟数字人放在转动台上， 将虚拟数字人和摄像头调整到测试所需的一般状态。
b) 按照图 3 所示，测试人员站在测试位置，起始角度应使虚拟数字人正面面对测试人员。
c) 以不大于 1°/s 的角速度逆时针转动被测支持语音和视觉交互的虚拟数字人，同时进行测试厂商说明书中规定的某一简单手势动作，直至手势无法被正常识别时，停止测试工装车的移动，测量转动的角度 θ_1。
d) 将被测支持语音和视觉交互的虚拟数字人恢复到起始角度，然后以不大于 1°/s 的角速度顺时针转动被测支持语音和视觉交互的虚拟数字人，同时进行测试厂商说明书中规定的某一简单手势动作，直至手势无法被正常识别时，停止测试工装车的移动，测量转动的角度 θ_2。

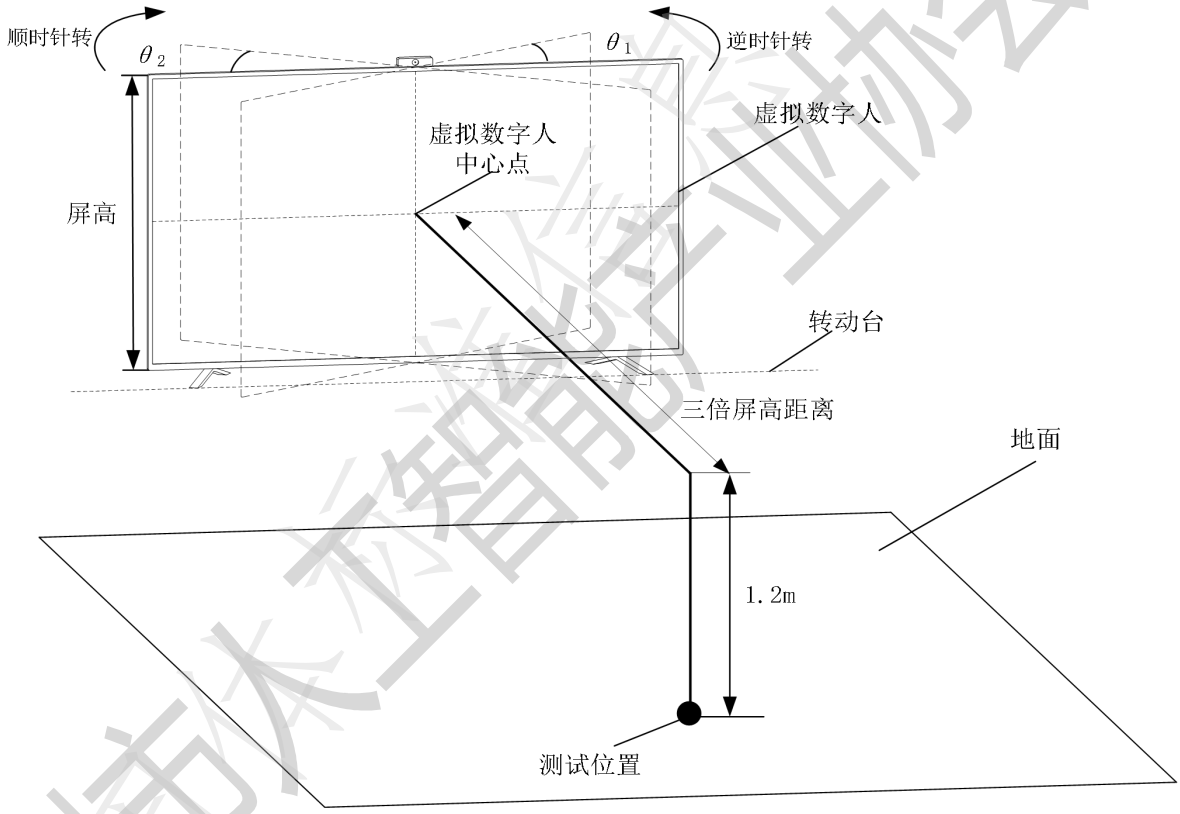

图 2 识别角度范围测试图示

6.6.2.2 测试结果

识别角度合格范围： $\theta_1 \geqslant 30°$ 且 $\theta_2 \geqslant 30°$ 。

6.6.3 识别率测试

6.6.3.1 测试步骤

测试步骤如下：
a) 被测支持语音和视觉交互的虚拟数字人放在平台上，将虚拟数字人和摄像头调整到测试所需的一般状态。
b) 按照图 4 所示， 测试人员站在测试位置，使电视屏幕正面面对测试人员。
c) 分别测试所有的手势，每组手势不少于 10 次，总的测试次数不少于 100 次，并记录识别正确的次数 X 和总测试次数 Y 。

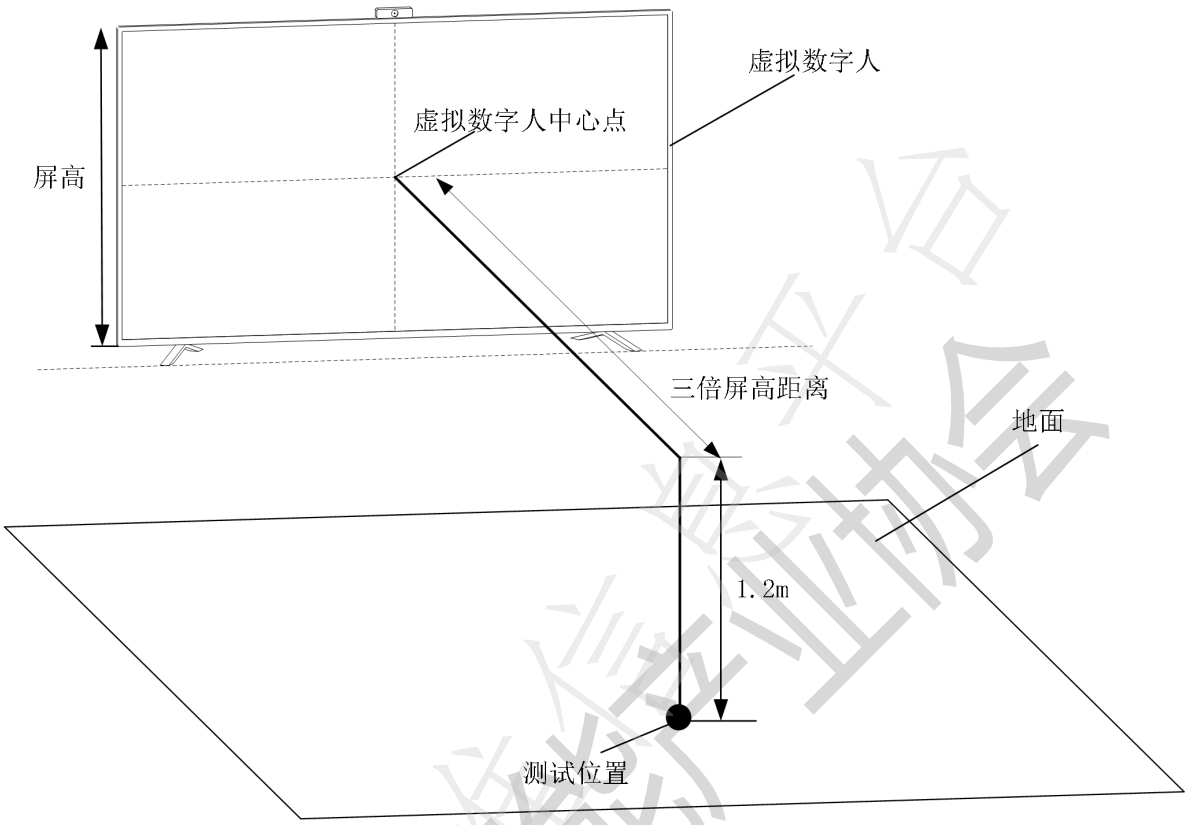

图 3　识别率测试图示

6.6.3.2 测试结果

$$识别率 = \frac{X}{Y} \times 100\%$$

识别率≥90%

6.6.4　识别响应时间

6.6.4.1　测试步骤

a)　被测支持语音和视觉交互的虚拟数字人和数字秒表放在平台上，将虚拟数字人和摄像头调整到测试所需的一般状态。

b)　按照图 5 所示，测试人员站在测试位置，使电视屏幕正面面对测试人员，在测试人员后方调整好视频记录仪的位置，使视频记录仪能同时记录测试人员的手部运动、电视屏幕的响应动作和秒表的时间记录，然后开启视记录仪和秒表。

c)　测试人员依次测试厂家规定的基本手势动作，测试组数不能小于 5 组，每组动作不少于 3 次，同时由视频记录仪记录整个过程。

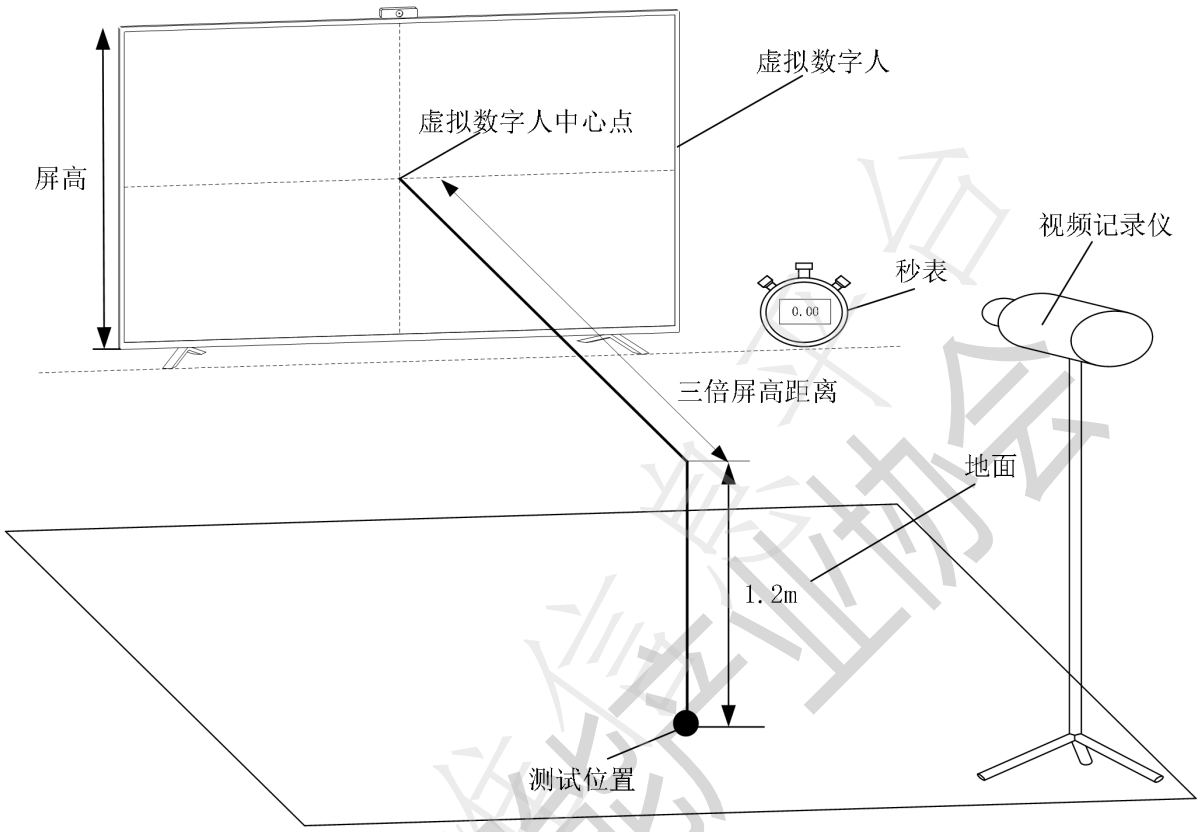

图 4　识别响应时间测试图示

6.6.4.1　测试结果

根据视频记录仪的画面，由秒表确认各单次手势的响应时间（动作起始状态为起点，响应完成为结束，之间的时间差为单次相应时间 t），按照表1记录单次时间并计算单组的平均时间。

表 5 测试结果记录

组数	第1次	第2次	第3次	单组平均值
第1组	t_{1-1}	t_{1-2}	t_{1-3}	$T_1=\Sigma t/3$
第2组	t_{2-1}	t_{2-2}	t_{2-3}	$T_2=\Sigma t/3$
第3组	t_{3-1}	t_{3-2}	t_{3-3}	$T_3=\Sigma t/3$
第4组	t_{4-1}	t_{4-2}	t_{4-3}	$T_4=\Sigma t/3$
第5组	t_{5-1}	t_{5-2}	t_{5-3}	$T_5=\Sigma t/3$

识别响应时间 T = Max（T_1, T_2, T_3, T_4, T_5），单位为 s。

识别响应时间：≤2s。

隐私安全

6.6.4　隐私安全测试

查看被测产品是否声明收集和使用用户数据的范围并征求用户的授权同意。

附录 A
（资料性附录）
语音测试集构建

A.1 输入输出要求

虚拟数字人在语音交互过程中的输入应满足以下要求：

a) 应支持中文普通话输入，宜支持英语；
b) 可处理语音输入为（180～300）字/分的语速，单次语音输入时长不应超过 30 s，特殊情况下不应超过 60 s；
c) 发音单元的持续时间应不小于 0.2 s，发音单元间的间隔不超过 2 s；若停顿时间超过 2 s，则认为一次语音输入结束。

A.2 测试集构建方法

从噪声、回声、人声，空间、待测设备这几个维度组合构建语音唤醒测试集和语音误唤醒测试集，尽量覆盖各种声学场景，模拟用户真实使用环境。

语音唤醒测试集通过专业录音麦克风在安静环境下组织录制人员录制待测设备的唤醒词。参与录制的人员，需考虑性别、口音、年龄等维度。

误唤醒测试集的构成，主要考虑实际应用场景中引起待测设备误唤醒的噪声来源。例如，家居环境下音箱的误唤醒主要来源于电视、人声谈话等，所以此时选择的误唤醒语料，每24小时包含6小时电视节目，6小时新闻节目，6小时人声对话（可选择谈话节目模拟），6小时音乐播放。典型的测试集构建方法示例见表A.1。

表 A.1 测试集构建方法示例

噪声	噪声来源	平稳噪声（家居环境噪声等）
		非平稳噪声（电视噪声等）
		交通工具
		自然声音
		其他
	噪声类型	点声源干扰
		散射噪声
	到待测设备距离	0.3 m
		1 m
		3 m
		5 m
	与待测设备角度	0°、45°、90°、180°、其他
	信噪比	原始
		(-5～15) dB，步长5 dB
回声	内容类型	音乐、有声节目、听声音Skill、TTS等
	信回比	原始
		(-35 ～ 0) dB，步长5 dB

表 A.1　测试集构建方法示例（续）

待测空间	空间类型	马路
		家居
		办公
	待测空间混响（500Hz）	T60 =（300±30）ms
		T60 =（500±30）ms
		T60 =（800±30）ms
待测设备	设备类型	表明被测设备类型，如小米AI音箱
	位置	一面靠墙 ＜ 0.1 m，三面开阔 ＞ 1 m
		两面靠墙均 ＜ 0.1 m，两面开阔 ＞ 1 m
		一面离墙 0.4 m，三面开阔 ＞1 m
		两面离墙 0.4 m，两面开阔 ＞ 1 m
		四面离墙均 ＞1 m
	高低	离地 0.7 m
		离地 1.5 m
	设备音量	例如，AI音箱0 dB，30 dB，50 dB，90 dB，100 dB
	设备编号	从1-10
目标声源	性别	男
		女
		儿童
	口音	普通话
		地区性方言
	语速	正常（0.85～1.5）s
		较快（0.65～0.85）s
	与待测设备距离	1 m
		3 m
		5 m
	与待测设备角度	0°、45、°90°、180°、其他
	发声位置	站姿：嘴离地面约 （1.5～1.62） m
		坐姿：嘴离地面约 0.8 m
		躺姿：嘴离地面约 0.4 m
	语料内容	唤醒词
		提问句
注：以上主要考虑家居和办公场景		

在线测试时，信噪比/信回比通过改变噪声源和纯语音段音量以及待测设备音量和纯语音段音量来获得。

离线测试语料中，按信噪比/信回比合成测试语料的方法：采用段信噪比计算方法，即纯语音段能量与混合时间段内的噪声/回声能量对比；实际语料合成时，整段噪声/回声语料设置同一增益来获得目标信噪比/信回比，但一段噪声/回声语料中混合多段唤醒词的时候，由于噪声、回声的能量实时在变化，每段唤醒词的信噪比/信回比不可能完全相同，应允许±1dB的误差。

音量设置需根据被测设备的音量范围和实现机制做定制化设计。

A.3 测试场景设置

测试所描述的场景应满足以下条件：

——环境：温度（23～26） ℃，相对湿度（25～75）%，大气压（95～101.3） kPa；

——高度：（1.1 ± 0.01） m；

——半径：距离被测中心（1.5 ± 0.02） m；

——角度：45° 。

典型的测试场景设置见表A.2。

表 A.2 测试的音量、距离、角度、噪音类型设置

测试环境	播放语料的音箱		播放噪音的音箱		音量	
	角度	距离	角度	距离	人声（1m 基准）	噪音
安静	60°	1 m	—	—	70 dB(A)	—
	90°	1 m	—	—	70 dB(A)	—
	90°	3 m	—	—	70 dB(A)	—
	90°	5 m	—	—	70 dB(A)	—
电视噪音	60°	1 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	1 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	3 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	5 m	150°	2 m	70 dB(A)	60 dB(A)
家庭聊天噪音	60°	1 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	1 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	3 m	150°	2 m	70 dB(A)	60 dB(A)
	90°	5 m	150°	2 m	70 dB(A)	60 dB(A)