

T/CRSS

重庆市机器人学会团体标准

T/CRSS XXXX—XXXX

服务机器人 中文语音交互系统 性能测评方法

Service robots-Chinese speech interactive system-Testing assessment approaches

（征求意见稿）

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。

— XX — XX 发布

XXXX — XX — XX 实施

重庆市机器人学会 发布

目 次

前言 II

1 范围 3

2 规范性引用文件 3

3 术语和定义 3

4 创建测评语音材料数据库 4

 4.1 测评语料设计 4

 4.2 语音录制要求 4

5 创建背景噪声数据库 5

 5.1 噪声种类 5

 5.2 噪声采集环境 5

6 测试环境搭建 6

 6.1 生活场景 6

 6.2 消音室内场景 6

7 测试方法 7

 7.1 概述 7

 7.2 基于语音材料数据库的测试方法说明 7

 7.3 基于现场口呼的测试方法说明 7

 7.4 一般要求 7

 7.5 试验项目 7

8 评价标准 10

 8.1 任务驱动的多轮对话 10

 8.2 开放域聊天 10

 8.3 语音识别准确率 10

 8.4 语音合成 10

附录 A（规范性） 语义解析评分标准表..... 11

附录 B（资料性） 五种常见音频的采样率..... 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。

本文件由重庆市机器人学会提出并归口。

本文件起草单位：

本文件主要起草人：

服务机器人 中文语音交互系统 性能测评方法

1 范围

本文件规范了服务机器人中文语音交互系统性能测评方法的术语定义、技术要求、试验条件、试验方法、等级评价指标。

本文件适用于服务机器人中文语音交互系统。

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 3096-2008 声环境质量标准
- GB/T 17696-1999 声学测听方法第3部分：语言测听
- 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 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 36464.4-2018 信息技术 智能语音交互系统第4部分：移动终端术语和定义

3 术语和定义

下列术语和定义适用于本标准。

3.1

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, 3.1]

3.2

语音交互系统 speech interaction system

由功能单元（或其组合）、数据资源等组成的能够实现与人类之间进行语音交互的系统。

[来源：GB/T 36464.2-2018, 3.2]

3.3

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, 3.1]

3.4

在线 online

处于跟互联网相连接的状态。

[来源：GB/T 36464.2-2018, 3.3]

3.5

脉冲编码调制 PCM pulse code modulation

对连续变化的模拟信号进行抽样、量化和编码产生的数字信号。

3.6

语音识别准确率 speech recognition accuracy

将人类的语音信号转化为正确文本的概率。

3.7

任务驱动的多轮对话 task-oriented spoken dialogue systems

满足人类特定目的的持续对话。

3.8

问答 Q&A

直接根据问题给出答案。

3.9

开放域聊天 open-domain chat

不局限话题的聊天。

3.10

响应时间 response time

从语音例句播放完成至系统做出语音反馈开始而所用的时间。

3.11

语音唤醒 speech wakeup; 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，切换到命令字识别、连续语音识别等其他处理状态的过程。

[来源：GB/T 36464.2-2018, 3.13]

3.12

误唤醒 false wakeup

语音唤醒过程中出现的，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。

3.13

稳定性 stability

语音交互系统在正常网络条件下稳定工作的能力。

4 创建测评语音材料数据库

4.1 测评语料设计

测评语料应从句型覆盖、词汇量覆盖、语义覆盖以及词域覆盖等加以设计。评测集文本分为若干组，每组可以由若干人发音组成。设计要求如下：

a) 应针对被测评的服务机器人产品的自身特性进行设计；

注：如针对儿童陪护的服务机器人产品应加入相关语料，如儿歌、故事等对话语料内容。

b) 应从不同领域、不同环境的语料中随机选择，同时考虑发音现象的覆盖性；

c) 应充分考虑句型、词汇、语义等的覆盖性；

d) 应根据不同测试项目的试验目的进行测试句的内容设计；

e) 测评语料除了核心主句，还应配备核心主句的相似变形句进行测评。

注：语料基础话题类别可参考：新闻、天气、时间日程、股票、体育、科技、日历、美食、旅行、音乐点播、电影、购物、花艺、家居、服装、美妆、养老、医疗、游戏等多种类别方向。

4.2 语音录制要求

4.2.1 录制设备及环境

为了清晰地录制说话者的声音，应考虑如下事项：

a) 录音场所要选择隔音效果好的，且能够隔绝反射音及外部噪声的地方，如消音室；

b) 需要选择中音带域和高音带域的频率响应特性好的测量麦克风；

c) 调整说话者的嘴和麦克风至合适的距离，如图 1 所示，建议为 15~20cm，不要超过声音等级的最大许容值，以免形成不正确的发音；

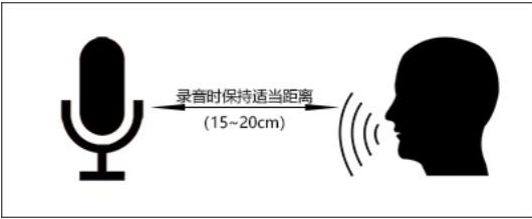


图1 录音距离示意图

- d) 会受到事先录制的声音音压影响的说话者，其噪音、讲话语速和理想声谱宜依照 ISO/TR 4870-1991 的 2.24~2.26 中的规定；
- e) 错误的发声、波形截断、无法理解的声音、舌短音、感叹词等都属于错误的，需要让发声者再次发声；
- f) 录制音频时，使音频信号等级的平均最大振幅保持在 16bit 左右（动态范围 96dB），音频信号的采样频率可设置为 44.1kHz，参见附录 B；
- g) 音频信号以 44.1kHz 的 16bit 线型 PCM 数据进行存储，再将音频文件压缩为 MP3 的格式进行使用；
- h) 录音过程至少包括录音、标注和确认三个步骤，保证测评数据库的正确性。

注1：男：低音82~392 Hz，基准音区64~523 Hz；中音123~493 Hz，高音164~698 Hz。

注2：女：低音82~392 Hz，基准音区160~1200 Hz；中音123~493 Hz，高音220~1.1 kHz 。

4.2.2 发音人员

测评语音录制时对发音人员的要求如下：

- a) 发音人员应持有国家普通话水平测试等级二级乙等（含）以上证书；
- b) 发音人员的选择应在符合系统对发音人员限制的条件下，尽可能选择具有代表性和统计分布规律的发音人员，特别是考虑不同口音、不同年龄、不同语速、不同教育背景、不同说话韵律等因素；
- c) 无论是特定人还是非特定人的中文语音交互系统，进行测试的发音人员至少为 20 个人以上；每人需对测试语料中的一组或多组发音；
- d) 不同发音人尽量采用不同语料组；
- e) 对于非特定人的中文语音交互系统，特别应强调对有一定口音的中文适应能力和中文语调问题。

5 创建背景噪声数据库

5.1 噪声种类

考虑到在机器人的服务环境中可能存在的主要噪声，应准备能够模拟此种环境的噪声。环境噪声选定为如下两个范畴：

- a) 白噪声；
- b) 不同场所的生活噪声。
 - 家庭：TV、电话铃声、吸尘器、冰箱、空调、洗碗的声音；
 - 教育场所：教室中存在持续吵闹和讨论的声音、老师讲课的声音；
 - 公共场所：饭店、医院、电影院、银行、机场、火车站、商场、超市的声音。

注：在测试过程中，以上所列举的噪声需持续发声5s以上。

5.2 噪声采集环境

为了录制噪声，应根据实际情况采取以下建议进行噪声采集：

- a) 需要在实际生活的噪声环境中进行录音；
- b) 需要选择中音带域和高音带域的频率响应特性好的测量麦克风；
- c) 将噪声源和麦克风调整到合适的距离，不要超过噪声等级的最大许容值；
- d) 在一段噪音区间内的停顿时间要短于 0.3s；

- e) 发出并收集 5.1 中所定义的噪音；
- f) 噪声录制格式为 16bit, 48kHz 采样的 WAV 格式，录制的噪声幅度最大值不能超过 30000(单位：样本值)，以免噪声幅度销顶；
- g) 要区分扩散场的背景噪声和点声源干扰噪声录制的区别；
- h) 扩散场的背景噪声录制应在不同位置设置麦克风进行录制，便于还原真实噪声场景；
- i) 点源干扰噪声的录制只需要一个麦克风在距离发声源较近的地方录制。

6 测试环境搭建

6.1 生活场景

可用一些实物搭建被测设备实际工作时的使用场景或将被测设备置于实际工作时的使用场景中进行测试。

注1：常见家用服务机器人使用场景为客厅，模拟家居客厅环境参考：正常标准白墙房间（无异形，最好不超过30平方米）内布置大小方桌各一个，常见规整尺寸的沙发、电视柜、电视、落地空调、落地灯、装饰隔板、窗帘等等家具，家具大小尺寸样式无具体要求，尽量贴合真实客厅场景，房间所处位置的实际生活噪声不超过 7.3.5 的要求即可，无需再次叠加新的背景噪声。当被测设备带有智能家居控制功能时，应为其配置可被控制的家电产品。

注2：常见商用服务机器人使用场景为办事大厅、展览厅、会场等，可将待测机器人及测试设备带入实际场景中进行试验展开。

6.2 消音室内场景

6.2.1 背景噪声还原

为了模拟真实的待测设备使用场景，推荐以下噪声播放要求：

- a) 一个点声源干扰只能用一个音箱设备播放，要考虑点声源干扰的距离和角度，以及点声源的个数；
- b) 扩散场的背景噪声一定要用多个音箱设备同时播放，比如房间的四个墙角，分别放置音箱，并且音箱的出声口对着墙，让声音经过墙面反射；
- c) 户外场景的噪声和室内场景的噪声应考虑混响，比如要在消音室模拟户外的噪声场景，且要在不同混响时间的房间模拟室内的噪声场景；
- d) 要设置不同的背景噪声和点源干扰等级，比如绝对安静，较安静(40dB-50dB)，嘈杂(60-70dB)，很嘈杂(80dB 以上)，此声压为机器人麦克风阵列处测量的声压；
- e) 回声场景（机器人本身扬声器播放而又被麦克风接收到的回声）要考虑机器人本身播放语音声压大小等级（比如 50dB 以下，50dB-70dB，80dB-最大声），此声压为机器人麦克风阵列处测量的声压，且需考虑机器人所处环境的混响、是否同时有点源干扰或者背景噪声等变量。

6.2.2 语料声源位置

- a) 按照实际具体使用场景下，使用设备与声源的相对距离、相对高度进行声源的摆放；

注1：部分被测产品可参考如下，教育机器人多为桌面级产品，可设定语料发声口距离机器人中心点30cm±5cm，设定其他产品的距离时最好不超过40cm；语料声源设备模拟使用者（成人）状态站姿高度或者坐姿高度时时，根据生活常见社交距离及机器人拾音情况，可设定语料发声口距离机器人中心点50cm~100cm。

注2：GB/T36464系列标准中规定，小于1m为近场距离，大于1m（含）为远场距离。

- b) 语料声源设备数量定为 1 个。

6.2.3 待测设备的位置

考虑到实际情况中，待测设备的使用情况，设备的位置会明显影响收音的质量，应该对待测设备的位置做以下考虑：

- a) 是否靠墙；
- b) 是否在墙角；
- c) 靠墙的距离；
- d) 按照实际使用场景下，使用者和待测设备常见的距离设置发声设备与待测设备的距离。

7 测试方法

7.1 概述

为保证中文语音交互系统评测的再现性，评测应尽量采用基于语音材料数据库的测试方法。无法采用基于语音材料数据库测试的，可采用基于现场口呼的测试方法。评测语料的设计与测评语音材料数据库的录制应保证与实际使用场景的一致性，测试的结果以满足规范的评测报告形式给出。

7.2 基于语音材料数据库的测试方法说明

基于语音交互标准库的测评过程采用录制的语音数据对被测系统进行间接测试。方法说明如下：

- 间接测试指利用高保真播放设备（或者人工嘴）把语音交互标准库中语音输出到被测系统；
- 在测试过程中，适时调整高保真播放设备与被测系统之间的距离（详见6.2.2）及其摆放高度，以满足日常使用环境的实际需求；
- 设备播放方向正对被测样品正面。

7.3 基于现场口呼的测试方法说明

现场口呼评测基本情况的要求与对测试语料的录制要求类同，如下：

- 需要有两个以上识别结果记录者，记录被测系统对当前发音的输出结果；
- 记录表应包括发音人、记录人、操作人、监督人、发音内容、语音交互结果等内容；
- 全部发音者测试结束后，统一按照性能标准进行指标评估，评估至少有两个人以上参与；
- 对于交互结果能以文件形式给出的，被测系统对发音人的响应还应给出文件形式的输出结果，以便测评工作参考内容充分。

7.4 一般要求

7.4.1 人员要求

- 试验人员应具备熟练操作被试品、测试设备并具有相应的技术和能力；
- 口呼测试人员应额外符合条款 4.2.2 的要求。

7.4.2 试验要求

除本标准或详细规范另有规定外，所有试验应在下列条件下进行：

- 温度：室内 $0^{\circ}\text{C}\sim 40^{\circ}\text{C}$ ；
- 相对湿度：10%RH \sim 90%RH；
- 背景音/环境噪声音量：50dB \sim 70dB；
- 测试语音播放音量应保持在 75（ ± 5 ）dB；
- 被测设备电量充足，可正常使用，并且应确保被测设备具有语音拾音功能，可通过对话方式对其进行控制和交互；
- 测试所需网络满足上行带宽不低于 100kbit/s、下行带宽不低于 50kbit/s，并保持稳定的连通状态；
- 混响时间：用于播放语料的扬声器处混响时间不超过 0.65s。

7.4.3 数据记录及处理

除个别项目有单独的规定外，一般用扬声器或人工嘴播放测试语料3次，在此过程中，测试人员可依据情况灵活选择口呼形式，测试次数也应为3次，最后以最优数据进行原始记录。

7.5 试验项目

7.5.1 问答

7.5.1.1 试验目的

测试语音交互系统在执行一问一答任务时的“答案准确度”。

7.5.1.2 试验方法

语音交互系统在正常工作的状态下，根据被测系统语音交互类别特性，有选择性地通过测评语音材料数据库中的问答句进行提问，当得到答案后测试人员记录该反馈答案（如果答案过长，测试员可简洁汇总记录），并依照系统反馈结果是否能解答对应问题为标准，以打分方式进行记录，打分规则应符合本标准附录A表A.1语义解析评分标准中的规定进行。

注：机器人用本地或者云端固定答复用户时，如你这样问还能愉快地玩耍吗等等，可简要记录为固定不相关。

7.5.2 任务驱动的多轮对话

7.5.2.1 试验目的

对被语音交互系统进行连续、有关联且伴有明确任务目的的对话，测试其对语义的处理深度。

注：该对话分为弱相关性多轮对话和强相关性多轮对话，在设计相应语料时应注意。

7.5.2.2 试验方法

语音交互系统在正常工作的状态下，根据被测系统的语音交互类别特性，有选择性地播放测评语音材料数据库中具有针对性的内容类型（如天气、股市、时间日程等）来进行测试，每一类每一组下的多轮对话例句均应进行测试。测试人员通过系统反馈结果，以是否完成相关任务为标准，记录成功完成的“有效对话轮数”，并计算该项目的任务完成率，详见条款8.1。

注：建议每组多轮对话设置为五轮。

7.5.3 开放域聊天

7.5.3.1 试验目的

测试语音交互系统在随机的语音对话中稳定保持话题延续性的能力。

7.5.3.2 试验方法

语音交互系统在正常工作的状态下，根据被测系统语音交互类别特性，有选择性地通过测评语音材料数据库中的开放域测试例句作为开始，进行聊天，之后根据被测系统的反馈结果，以测试人员口呼为主，口呼语料与被测结果相关且随机。测试人员记录被测样品的话题延续轮数，并对该项目进行打分，打分规则应符合条款8.2中的规定进行。

注：建议记录截止轮数为5轮，并以5轮判定其聊天时长为满分参考值。

7.5.4 在线平均响应时间

7.5.4.1 试验目的

从用户角度出发，测试语音交互系统在执行对话任务时，提供反馈所需要的时间。

7.5.4.2 试验方法

自语音例句播放完成的瞬间起测试人员开始计时，至被测产品做出语音反馈的开始瞬间完成计时。记录时间，每一个响应的时间总和除以总共测试次数即为在线平均响应时间，并将实际测试结果记录在原始记录中。

7.5.5 语音识别准确率

7.5.5.1 试验目的

测试语音交互系统对接收到的语音，正确识别语音中文字的能力。

7.5.5.2 试验方法

语音交互系统在正常工作的状态下，根据被测系统语音交互类别特性，有选择性地通过测评语音材料数据库中的每一类下的每一句语料进行测试，测试人员对照相应例句的识别结果记录每一句测试例句准确识别的字数，计算语音交互系统对声源例句中文字的“识别准确度”，详见7.5.5.3，将计算结果及相应得分填写在原始记录表中，打分规则应符合条款8.3中的规定进行。

注：测试例句的识别结果大多数时候可通过从有显示屏（若有）的被测系统观察实时获得，或从系统辅助APP云端获得，或从系统提供商后台调出被测系统语音识别的文字数据记录等等。

7.5.5.3 识别准确率计算方法

参见式（1）和式（2）。

$$WER = \frac{M_e}{N} \times 100\% \dots\dots\dots (1)$$

$$WCR = \frac{M_c}{N} \times 100\% = 100\% - WER \dots\dots\dots (2)$$

式（1）~（2）中：
N —— 语音文本总字数；
M_c —— 识别结果中正确识别字数；
M_e —— 错误识别字数；
WER —— 语音识别的字错误率；
WCR —— 语音识别的字正确率。

7.5.6 语音唤醒正确率

7.5.6.1 试验目的

测试语音交互系统在使用过程中被唤醒词正确唤醒的能力。

7.5.6.2 试验方法

统计其他项目所有测试语料进行前，按照下式（3）计算语音唤醒正确率，记录在原始记录中。
式中：

$$B = \frac{C}{D} \times 100\% \dots\dots\dots (3)$$

B —— 语音唤醒正确率；
C —— 唤醒词正确唤醒语音交互系统的次数；
D —— 总呼唤次数。

7.5.7 语音误唤醒频度

7.5.7.1 试验目的

测试语音交互系统在不使用待机时或者使用过程中被误唤醒的情况。

7.5.7.2 试验方法

将被测产品放于试验室环境中最可能受影响的一处并调至待命状态，在不主动唤醒产品的情况下，循环播放预录制的环境背景噪音，并通过摄像机记录产品状态，测试时间持续几个小时后，测试人员通过视频统计被试品自动唤醒的次数，并记录在原始记录中，计算频度方法见下式（4）。

$$\theta = \frac{E}{h} \dots\dots\dots (4)$$

式中：
θ —— 语音误唤醒频度；
E —— 被误唤醒的次数；
h —— 测试时间，小时。

7.5.8 使用稳定性

7.5.8.1 试验目的

综合考察语音交互系统在使用过程中的性能稳定性，如非法操作、死机频率、语音反馈突然中断等各类情况。

7.5.8.2 试验方法

- a) 此项目测试人员应全程参与过被测语音交互系统的其他所有项目测试，并具备服务机器人测试经验，方可参与主观评价；
- b) 在其他所有项目测试完毕后，多名测试人员进行综合判定，以文字描述的方式体现在原始记录中。

7.5.9 语音合成

7.5.9.1 测试目的

综合主观评价被测语音交互系统所反馈的语音质量和用户接受度。

7.5.9.2 试验方法

男女测试人员各2人，通过其他项目的语音反馈结果，测听合成语音同真人语音在音质、可懂度、自然度和流畅度等方面的差异，并以平均意见得分对主观测评进行量化，并将文字描述结果及相应得分填写在原始记录表中，打分规则应符合条款8.4中的规定进行。

8 评价标准

8.1 任务驱动的多轮对话

- a) “有效对话轮数”为测试五轮全部语料后，采用递进方式判断记录，按照其前面的最优结果进行有效轮数的记录，如单轮对话记1，能连续进行到相邻二轮对话记2，能连续进行到相邻三轮对话记3；
- b) 计算该测试项目的任务完成率并记录，总有效对话轮数(N)与总预设对话轮数(M)的比值， $N/M \times 100\%$ 为任务完成率(C)。

8.2 开放域聊天

- a) 5分：优（聊天内容种类丰富且持续轮数多），4分：良（聊天持续轮数多），2~3分中（聊天有一定内容但持续轮数短），1分：及格（能进行聊天但持续轮数极短），0分：差（不能进行响应聊天）；
- b) 该项目以聊天话题的延续性作为评判标准，如被试品能成功延续话题或完成用户的相关需求，则视为测试成功，记录话题延续轮数。

8.3 语音识别准确率

识别准确率90%以上5分，识别准确率85%-90%之间4分，识别准确率70%-85%之间3分，识别准确率60%-70%之间2分，识别准确率60%以下1分。

8.4 语音合成

该项仅针对被测产品做出的回复中所包含的语音合成部分（直接播放歌曲、录音内容，及无应答均不在打分范围内）进行打分，5分：优（接近人类自然语音状态）；4分：良（机械合成感不明显）；3分：中（机械合成感较明显），0~2分：差（有明显机械合成感）。记录平均结果，语音合成平均意见分为总分/打分人数。

附 录 A
(规范性)
语义解析评分标准表

表A.1 语义解析评分标准表及示例

评分标准	要求	解释	举例	
5 分	回答正确+内容精准。	全部理解语义，并给出了精准的回答。	Q: 泰山多高?	A: 1545 米。
4 分	回答正确+有多余内容。	正确理解语义，并用有效的语料进行回答，但句式上有多余内容。	Q: 泰山多高?	A: 泰山坐落在山东省中部,为中国五岳之首,主峰玉皇顶,海拔 1545 米,高度居五岳第三位。
3 分	基本回答正确+多余数据+句式内容有错误+答案啰嗦（长度超过 60 字）。	基本理解语义，但句子内容错误包括：错别字、缺漏字、内容有误，答案长度过于啰嗦。	Q: 泰山多高?	A: 泰山又名岱山，坐落在山东省中部,为中国五岳之首,古称“主峰玉 x 顶,海拔 1545 米,世界自然与文化遗产,世界地质公园,国家 5A 级景区……
2 分	基本理解语义+无数据支持而无法回答。	系统基本理解语义，但数据库不支持回答，或者对用户进行提示性回复。	Q: 泰山多高?	A: 我不知道泰山多高，但我知道泰山在山东泰安。
1 分	回答错误+弱关联性。	没有全部理解语义，数据不够有效，语料属于问题范围但关联性小。	Q: 泰山多高?	A: 也许是 500 米，去一趟还要买门票。
0 分	回答错误+无关联性。	系统没有理解语义，并且回答错误，无关联性。	Q: 泰山多高?	A1: 旅游高峰期特别多，还是不要去了。 A2: 你在说什么，我听不懂。

附 录 B
(资料性)
五种常见音频的采样率

表B. 1 五种常见音频的采样率

采样率	质量级别	用途
48kHz	演播质量	数字媒体上的声音或音乐
44. 1kHz	CD 质量	高保真声音或音乐
32kHz	接近 CD 质量	数字摄像机音频
22. 05kHz	FM 收音质量	短的高质量音乐片断
11kHz	可接受的音乐	长音乐片断
5kHz	可接受的话音	简单的声音，电话