LISTENAI

离线语音声学综合效果评估及测试方案参考

文档密级：对外公开

Version 1.1 2021.1.17

声明

本手册由聆思科技版权所有，未经许可，任何单位和个人都不得以电子的、机械的、磁性的、光学的、化学的、手工的等形式复制、传播、转录和保存该出版物，或翻译成其它语言版本。一经发现，将追究其法律责任。

聆思科技保证本手册提供信息的准确性和可靠性。聆思科技保留更改本手册的权利，如有修改，恕不相告。请在订购时联系我们以获得产品最新信息。

对任何用户使用我们产品时侵犯第三方版权或其它权利的行为聆思科技概不负责。另外，在聆思科技未明确表示产品有该项用途时，对于产品使用在极端条件下导致一些失灵或损毁而造成的损失概不负责。

变更记录

版本	变更内容	变更人	审核人	日期
1.0	初稿	宋瑶	李逸卿	2020-11-03
1.1	新增聆思参考指标	李逸卿	宋瑶	2021-01-17

目录

1 概述

1.1文档简介

1.1.1 发布背景

由于目前国家或行业没有制定针对人工智能离线语音相关方案的测试及验收标准，为了保障离线语音方案的质量和用户体验，实现测试结果的相互认可和可重复性，特制定此标准。

文档中相关测试定义、方法可参考 GB/T 36464.2-2018 国家标准；相关普通话及带口音普通话定义参考《普通话水平测试等级标准》。

本标准由安徽聆思智能科技有限公司提出。

1.1.2 适用范围

本方案主要针对离线语音声学效果给出一套综合评估和测试方案。包含：

- 测试项：从语音唤醒率、识别率、串扰率、误唤醒频度、响应时间、主观体验六个方面，提供综合项和细分项的评估维度。
- 测试指导：包括测试方案所需设备、语料制作、场地、场景搭建等要求。为保持多场景复测数据的一致性，测试语料、场景搭建、信噪比、环境混响值等均需客观一致，确保方案设置场景、被测场景语最终用户使用场景能够被复现。
- 测试结果异常排除方法。

1.2 测试术语及定义

1.2.1 基本术语

本文档只针对唤醒词和离线命令词制定的标准

1) 混响：一种声学特性，声波在室内传播时，要被墙壁、天花板、地板等障碍物反射，每反射一次都要被障碍物吸收一些。这样，当声源停止发声后，声波在室内要经过多次反射和吸收，最后才消失。这种现象叫做混响，从声音发出经多次反射后直至听不到声音为止的这段时间叫做混响时间。

2) 分贝：声压级的单位，大约等于人耳通常可觉察响度差别的最小分度值，单位为 dB，。通常使用声压计（也叫声级计、分贝仪）作为检测仪器，声压计通常有 A、C、Z 三种计权，dBA 表示的就

是 A 计权下的声压级，本文中所有分贝均指 dBA。

3) 唤醒词：相当于给设备起的名字，说话人通过说出唤醒词，唤醒语音模块，使其进入指令识别状态。例如：模块的唤醒词为“小飞小飞”，则使用者只要说出“小飞小飞”即可唤醒模块。

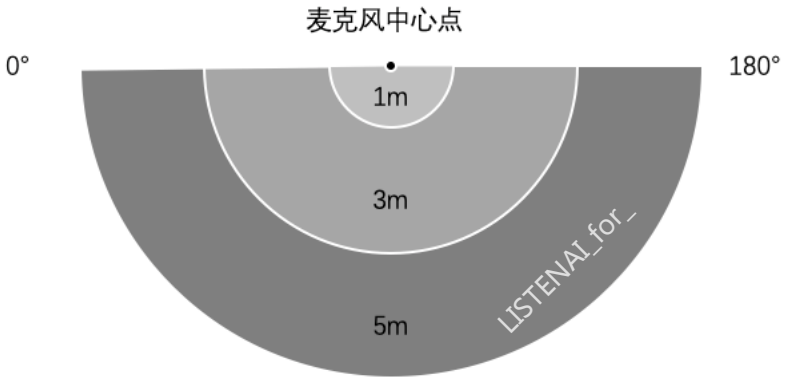
4) 命令词：智能语音产品制定的控制指令，多数是对产品功能控制的文本描述，用户说出命令词，系统识别成功后，配合设备的控制系统，实现功能的触发，如“打开空调”。

5) 语音打断：在设备播音过程中，用户发出第二个命令词（指令），检查第二条语音指令是否执行成功；语音打断会因交互过程中所处阶段的不同，分为唤醒打断和识别打断。

打断成功率跟喇叭播放声音的音量以及噪音环境有关。

6) 拒识：用户说出正确指令（唤醒词/命令词），但设备没有识别结果和反馈。

7) 方位角：双麦阵列呈线性排布，以双麦中间点作为尖角点，阵列线性方向为角边，顺时针方向定义夹角大小（即正对麦克风时，右侧为 0 度，如图示）。



8) 俯仰角：麦克风与声源的高度差（垂直方向），为操作方便，本方案关于存在俯仰角度的交互场景，均用距离高度差（单位：cm）进行标注。

9) 响应时间：从用户完整说完唤醒词/命令词最后一个音节起，到被测模块播出反馈提示音第一帧的时间长度，精确到毫秒（ms）；

10) 信噪比（SNR）：本文统一指人声分贝值减去环境噪声分贝值的差值

11) 自激：设备播放上一句命令词的应答语，被设备又识别成了下一句命令词。例如用户说“调到 16 度”，设备应答“主人，已为您调到 16 度”，这个应答语被设备自身的麦克风收音，又识别成了命令词“调到 16 度”，设备第二次应答“已经是 16 度了”。

12) 噪音误识别：用户被唤醒后，设备将环境噪音识别成了命令词。例如房间里正在播放电视剧，用户喊了唤醒词之后还未说命令词，设备将电视剧噪音识别成命令词“制冷模式“。此项指标中只统计将噪音识别成命令词的情况，识别成其他语句不计入统计。

13) One-shot：每次识别前都需要唤醒，且唤醒后仅支持一次识别

14）全双工：设备被唤醒后支持多次识别，在未超时的前提下，不限制识别次数。超时时间视设备而定，一般是 15s~30s。

15）场景：指目标应用产品所处的噪声、混响环境，可以从空间大小、混响环境、噪音类型进行区分，空间分布主要有客卧、厨房、卫生间、阳台；根据时间段不同，对于设备而言的环境噪音环境也会不同，例如家居环境主要分为白天（安静），白天（常噪），傍晚聚集（高噪），晚间睡眠场景等。

说明：以上人声噪音中，均不包含唤醒词和命令词相关音频，统计过程中，不将其作为误唤醒/误识别进行处理。

16）测试集：测试音频集，根据测试用途分为唤醒测试集、识别测试集、误唤醒测试集。

1.2.2 效果指标相关定义

本文档针对语音交互过程中的体验指标按唤醒、命令词识别、~~语音播报反馈~~进行阐述，每个大项下有对应的评估子维度和指标。

1.2.2.1 唤醒

1） 唤醒率：不同场景下，设备待唤醒状态下获取唤醒词的成功次数/总播放次数*100%。如没有回声消除功能，测试过程中，需要等播报结束再进行唤醒。

2） 综合唤醒率：综合各类场景，角度取的唤醒率，将各场景（安静、常噪、高噪）*（方位角/俯仰角），按照日常出现频率，取权重进行换算。其中频率模拟日常家居环境配比噪音类型，安静（晚间睡眠）：常噪（白天）：高噪（聚集时间）=8:10:6。

3） 唤醒打断成功率：设备在播音过程中，获取唤醒词的成功次数/总播放次数*100%，打断唤醒分为两种情况：① 离线打断唤醒：设备正在播放命令词的应答语，如“已为您调到 16 度”、“已为您调到最大音量”，用户说唤醒词让设备重新唤醒；② 在线打断唤醒：设备播放音乐时，用户说唤醒词让设备唤醒。两种情况中，设备重新唤醒成功，才算成功打断。两种情况的测试数据需要区分统计。

4） 误唤醒频度：模拟用户实际使用场景，将待测设备调至待命状态下，在不含唤醒词的前提下，记录规定时间段内设备被误唤醒的次数。

5） 唤醒响应时间：用户说出唤醒词，到设备成功唤醒并给出唤醒应答语的时间间隔，统计区间为：唤醒词最后一个字说完至设备唤醒应答语第一个字播出前

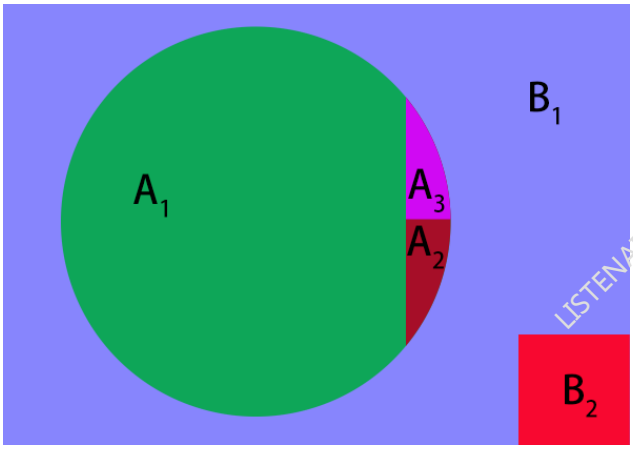
1.2.2.2 识别

1） 识别率：设备被唤醒，进入识别状态后，获取命令词识别成功的次数/总~~播放~~唤醒次数*100%，如没有回声消除功能，测试过程中，需要等播报结束再说命令词。

2) 综合识别率：综合各类场景，角度取的识别率，将各场景（安静、常噪、高噪）*（方位角/俯仰角），按照日常出现频率，取权重进行换算。其中频率模拟日常家居环境配比噪音类型，安静（晚间睡眠）：常噪（白天）：高噪（聚集时间）=8:10:6。

3) 识别打断成功率：设备在播音过程中，获取命令词的成功次数/总播放次数*100%。只有支持全双工功能的设备，才有识别打断成功率这个测试指标。识别打断成功率有两种情况：① 离线识别打断率：设备正在播放上一句命令词的应答语，用户喊另一句命令词，看是否成功打断设备的播音；② 在线识别打断率：设备处在识别状态，且正在播放音乐，用户喊另一句命令词，看能否进入识别模式。第一种情况，打断成功指用户第二次说命令词成功的中止了设备的播音行为，并进入识别模式。第二种情况，不要求音乐中止播放，但需要进入识别模式。两种情况的测试数据需要分开统计。

4) 串扰：设备接收到相关语音内容，有识别结果但识别成其他集内词。串扰主要分为集内串扰和集外串扰。关于串扰及相关指标的关系，如下图示意，A（A1∪A2∪A3）为集内词测试集，B（B1∪B2）为集外词测试集，A∪B 构成测试全集。**本文档主要讨论集内串扰率。**



A1——正确识别

A2——有识别结果但识别成其他集内词

A3——无识别结果

B1——无识别结果

B2——有识别结果

集内正确率计算方式：A1 / A × 100%

集内串扰率计算方式：A2 / A × 100%

5) 识别响应时间：用户说出命令词，到设备成功识别给出应答语的时间间隔，统计区间为：命令词最后一个字说完至设备应答语第一个字播出前；此项测试指标仅针对识别正确的情况

2 效果评估

2.1 确定应用目标

- **目标交互场景**：根据产品属性、安装位置和所在场景的类型、空间大小，判断人机交互的常规活动场景，确定人机在空间内的所处位置、距离、角度以及俯仰角。
- **目标交互人群**：根据产品定位，判断目标交互人群的交互效果是否满足。确保覆盖目标发音人人群的口音、音色、年龄层次。人群标签选择方法请参考 3.1.1/3.1.2 小节。

综合交互场景和交付人群标签这两点，确认各项效果指标的预期范围。

2.2 场景评估

2.2.1 环境混响

混响值对语音效果的影响较大，混响越大，唤醒率和识别率会出现一定程度的下降。可根据产品交互场景，确定测试场地（声学实验室）的混响在何种范围内。如被测设备是一台音箱，常规交互场景多数为客厅、办公室，实际测试时，测试场地可选择中混响环境。

另有一种情况，由于设备使用场景丰富，测试场景相应地可选择多种。例如被测设备是一台手机，则使用场景非常广阔，小混响、中混响、大混响都需要在测试中考虑到。实际测试时，测试场地需要覆盖三种混响类型。

混响时间	环境底噪	场景
小混响，0.1~0.3s	≤40dB	卧室（木质地板）
中混响，0.3-0.7s	≤45dB	客厅，阳台，家电卖场环境、办公室场景、包厢场景、电视场景、聚会场景
大混响，≥0.7s	≤45dB	厨房，浴室，大仓库场景

环境混响测试方法，可参见 3.4.1 小节。

2.2.2 环境噪音

以下为环境噪音类型的描述，测试过程中，因噪音环境的不确定性，容易影响效果测试的结果，在进行语音方案效果的验证测试时，需分析产品实际应用场景中环境噪音的主要内容。通常环境噪音不会只包含一种，因此需要确认各种类型的噪音在整体噪音中的时长占比。

通常来说，噪音分为人声噪音和非人声噪音两种类型。其中人声噪音对唤醒/识别影响大于非人声噪音。另外，噪音音量起伏不定，导致测试环境信噪比不够稳定，经常变化也会导致唤醒/识别效果下降。以上两点是评估噪音的核心因素。

在实际测试时，为了评估各种噪音下的效果，通常选择两种做法：

① 将各种噪音拼接成一条长音频，进而覆盖各种类型的噪音；

② 将各种噪音混杂在一起，同样达到覆盖各种噪音的目的。

以上两种方法都需要考虑到各种类型噪音的时长占比。例如被测设备的使用场景是浴室，噪音

内容包含：水流声、人说话声、窗外马路噪、洗衣机运行噪音。其中水流声占比 70%、人说话声 20%、洗衣机运行噪音 5%、窗外马路噪 5%。确认好时长占比后，可选择将这四种噪音拼接在一起在测试中使用；也可以在满足时长占比的前提下，选择任意几种混杂在一起，在测试中播放。

噪音类型	
人声噪音	闲聊、音乐、电视、新闻、广告
非人声噪音	窗外马路噪、自然界白噪音：雨声、风声、打雷声、开关门声音、脚步声、风扇噪音、水流声、家电设备运行噪音、洗衣机运行噪音

以上噪音测试集的具体制作方法，可参见 3.1.3 小节。

2.2.3 环境搭建

离线语音测试环境需能模拟语音识别设备常规应用时所处的真实环境及工作情况。

以下是参考《GB50096-2011 住宅设计规范》中“5 套内空间” 对厨房、卫生间、阳台、起居室（厅）以及卧室套内空间面积的规范，给出的参考。在准备测试方案之前，可根据被测物的实际使用场景，进行相关环境搭建。

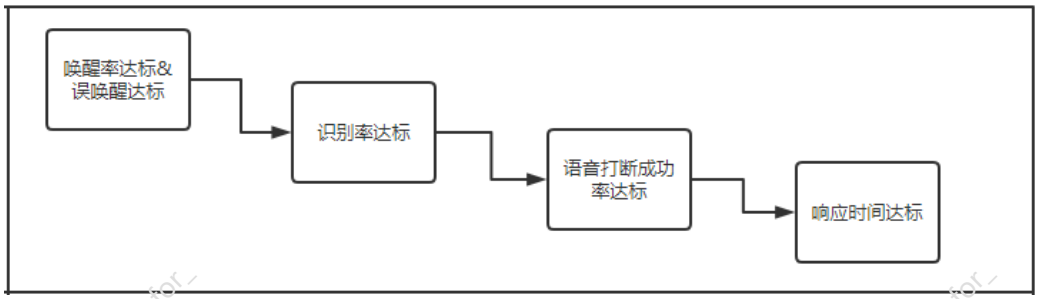
应用场景	环境噪声（dB）	混响（s）	最小距离（m）	最大距离（m）	应用场景参考面积（㎡）	适用语音设备
厨房	50-60	0.65-0.75	1	2	5-10	厨电设备，如：微波炉，抽油烟机洗碗机等
卫生间	50-60	0.65-0.75	1	2	5-10	卫浴设备，如：浴霸，马桶等
阳台	50-60	NA	1	2	5-10	阳台设备，如：洗衣机，晾衣机等
客厅	50-60	0.45-0.55	1	5	15-35	客厅设备，如:空调,茶具，电视等
卧室	50-60	0.45-0.55	1	5	10-20	卧室设备,如:空调，台灯，电视等

2.3 效果评估

本方案只针对离线语音进行效果评估

2.3.1 评估维度

评估维度主要包括唤醒率、误唤醒频度，识别率、语音打断成功率、响应时间。按照各项指标对于日常交互体验的影响，各项指标（含子指标）的重要优先等级排序如下（高→低）：



2.3.2 参考指标

客户可自行定义各场景下，各评估维度需要达到的指标。以下为聆思建议的参考指标：

唤醒率/识别率

环境	设备工作状态	人声(dB)	环境噪声（dB）	指标	场景适用性说明
安静环境	非运转/非播报	60-70	NA	最小距离：≥97% 最大距离：≥95%	适用于所有设备
	播报	60-70	NA	最小距离：≥95% 最大距离：≥88%	适用于音箱等长播报及音频播放的语音识别设备
	运转	60-70	设备运转噪声 50-60	最小距离：≥95% 最大距离：≥88%	适用于能产生机械噪声的语音识别设备
	运转（强噪声）	60-70	设备运转噪声 65-75	最小距离：≥93% 最大距离：≥85%	适用于能产生高强度机械噪声的设备，如：抽油烟机
常噪环境	非运转/非播报	65-75	环境噪声 50-60	最小距离：≥92% 最大距离：≥88%	适用于所有设备
	播报	65-75	环境噪声 50-60	最小距离：≥92% 最大距离：≥85%	适用于音箱等长播报及音频播放的语音识别设备
	运转	65-75	环境噪声 50-60 设备运转噪声 50-60	最小距离：≥92% 最大距离：≥85%	适用于能产生机械噪声的语音识别设备
	运转(强噪声）	65-75	环境噪声 50-60 设备运转噪声 65-75	最小距离：≥90% 最大距离：≥75%	适用于能产生高强度机械噪声的设备，如：抽油烟机

说明：

- 最小距离，根据“环境”，“应用场景”参考“表 1”确定具体距离。
- 最大距离，根据“环境”，“应用场景”参考”表 1”确定具体距离。

误唤醒次数≤1 次/24h

串扰率<5%

响应时间<1000ms

3 测试方案

整个测试流程包含测试集准备、测试设备准备、测试场景搭建及设备摆放参考等测试前期准备工作。

3.1 测试集

3.1.1 唤醒/识别测试集统一要求

1) 录音人选择

性别参考比例：男：女 = 1:1

年龄参考比例：

[18,30)	[50,70)	[6,12)
80%	10%	10%

口音参考比例：

普通话	北方口音	上海口音	长沙口音	广州口音	客家口音	闽南口音
20%	64%	6%	3%	3%	2%	2%

注：以上所有语料均指普通话，三甲普通话不带口音，各地口音占比参考依据为 2019 年中国方言分布数据。

2) 音频质量要求

- 口音音频不具备代表性，如要求录制湖南口音，但录音人员说的是标准普通话，音频就不满足需求，需要重新录。口音的测试价值本就在于口语发音上，音频必须符合当地口音发音习惯。
- 音频语量过低、喷麦，发音含糊不清。音量太低、喷麦或发音不清直接影响声学效果，不符合正常用户习惯，不具备测试意义，建议质检时剔除此类数据。
- 慢语速音频过慢、快语速音频过快。快语速或慢语速的音频，在质检时以句子朗读清晰为底线。
- 必须在小混响的环境中录音，例如声学实验室、消音室、录音室等，不可在中大混响环境录制，如会议室、楼梯间、卫生间、空旷的方便。参考上文中 2.2.1 小节描述。

3) 音频格式要求

- 每个人每种语速喊一句命令词是一条单独的音频
- 需要有文本标记每个音频的发音人姓名（或代号）、语速、年龄、性别
- 每条音频中命令词前后静音端分别控制在 1s 左右

- 每条音频的幅值都在 8000~15000 左右，不要低于 5000，高于 20000
- 要求 wav 格式、16bit、单声道、48k（最低 16k）
- 每条音频的底噪幅值不高于 500

3.1.2 唤醒测试集选择

1) 录音人选择

测试集要求，20 个人，每个词每人至少录制 20 句

语速参考比例：

快语速	正常语速	慢语速
20%	60%	20%

以四个字的唤醒词为例，不同语速在音频中人声波形长度做如下区分：

快语速：800ms~1000ms

正常语速：1000ms~1800ms

慢语速：1800ms~2000ms

2) 音频质量要求：

- 音频录制之后需要检查，① 音频内容是否错误，如小美小美喊成小飞小飞；② 除唤醒词之外是否有咳嗽声，或口气词，如“小美小美啊”。语料错误直接影响测试结果，必须严格质检。
- 音频发音不准。这种情况多存在英文唤醒词，如 hi，常见发音有“嗨、害、嘿”。有些唤醒词存在多种发音方式，出现这种情况时，需要提前告知录制人员哪一种是正确发音，同时在质检时进行检查。

3) 测试集制作方法：

- 唤醒测试集最少要求有 200 句，必须按上文要求覆盖各种语速
- 为了保证始终在待唤醒状态下测试唤醒率，唤醒测试集中每个唤醒词之间的时间间隔，需要大于被测物的识别超时时间。制作测试集时需要在每个唤醒词音频之间增加对应时长的静音段
- 为了保证测试中人声音量在一定的范围内，制作测试集时要对音频做能量规整，参考做法是将各个音频的幅值调整到统一的范围内。

3.1.2 识别测试集选择

1) 录音人选择

测试集要求，20 个人，每句命令词每个人至少读 5 遍

语速参考比例：

快语速	正常语速	慢语速
20%	60%	20%

命令词的语速要求，无法同唤醒词一样做细化要求，以发音人日常习惯为准，无需刻意的过快或过慢读命令词。

2) 音频质量要求：

- 音频录制之后需要检查，① 音频内容是否错误，是否断错句子、漏字、多字、错字；② 是否含有语气词，如“打开空调”读为“打开空调啊”。语料错误直接影响测试结果，必须严格质检。
- 4、音频发音不准。需要提前告知录制人员哪一种是正确发音，同时在质检时进行检查。

3) 测试集制作方法：

- 识别测试集必须覆盖被测设备的所有命令词，每句最少要求有 20 句语料音频，必须按上文要求覆盖各种语速
- 不管待测设备是否支持全双工，均采用 one-shot 模式测试，以保证每个命令词均在识别状态测试
- 制作识别测试集时，每句唤醒音量可大于命令词音量约 8db 左右，保证在各测试场景中，唤醒音量足够大，一定能够被唤醒。唤醒词采用一句语速正常、发音清晰的音频，可循环使用。
- 测试结果统计时需要根据唤醒记录，过滤掉因为未被唤醒而识别为空的情况；若唤醒后识别为空，视为识别错误，需要计算到结果中。
- 保证唤醒后有恰当的时间间隔，再播放命令词，防止设备应答语没有被命令词打断，亦要防止识别超时
- 保证命令词播放后有足够的时间间隔，再播放下一句唤醒，避免设备应答语没有被打断，而导致没有唤醒
- 为了保证测试中人声音量在一定的范围内，制作测试集时要对音频做能量规整，参考做法是将各个音频的幅值调整到统一的范围内。

3.1.3 噪音语料选择

基于 2.2.2 小节中对环境噪音的区分，在人声噪音和非人声噪音的大类下，再进行以下细分，其中电视噪、音乐噪、闲聊噪属于人声噪音；厨房噪、自然环境噪音、设备自噪属于非人声噪音。

1) 噪声种类说明：

- 电视噪音：电视剧、新闻、综艺、电影；
- 音乐噪音：国语歌、粤语歌、英文歌、韩文歌、轻音乐/古典；
- 闲聊噪音：正常人声对话；
- 自然环境噪音：窗外马路声、风声、雨声、施工装修声、开关门声、走路声、水流声；
- 厨房噪音：包含炒菜、洗碗、油烟机、微波炉等噪声；
- 设备自噪：设备本身运行的噪声，如冰箱、空调工作状态下电机、风扇等发出的声音；

以上是各种常见的噪音分类说明，在测试时根据实际需求，可选三个方案：① 单种噪音，例如只测电视噪；② 循环噪，即多种噪音拼接到一起，例如电视噪、音乐噪、自然环境噪、闲聊噪四种噪音拼接形成一个长音频；③ 混合噪，即多种噪音同时发生或相继发生在一个音频中，类似于我们在家庭中体验的真实环境，有人说话，有人在看电视，远处又有厨房做饭的声音传来。混合噪需要在实际环境中录制。

根据声学效果评测标准，建议选择循环噪或混合噪。

2) 混合噪说明：

- 家居噪：
  - 昼间家居噪：包含电视噪、闲聊噪、音乐噪、厨房噪、设备自噪、自然环境噪
  - 夜间家居噪：包含设备自噪、鼾声、窗外风声、车流声（视居所而定）
- 卖场噪：店铺轻音乐、闲聊噪
- 办公噪：闲聊噪、设备噪（电脑键盘声）

3) 音频质量要求：

- 检查音频中是否含有唤醒词和命令词，若包含，该部分的音频需要剔除。必须严格质检。

- 录制音频时保持噪音一直持续播放，若噪音源停止较长时间（＞5min），可先停止录制等噪音重新响起再录。如客厅里暂时没有人，去睡午觉了，可以等午睡后再继续录制。
- 录制自然环境噪、厨房噪或混合噪时，噪音内容比较复杂，尽量覆盖上文中列举到的所有种类，不要求所有种类的声音同时发生，可分段录制。
- 电视噪、音乐噪可直接下载音源，要求声音清晰即可。

4) 音频格式要求

- 每种类型的噪音需要有文本标记噪音内容。
- 每条音频的幅值不要低于 2000，高于 20000
- 要求 wav 格式、16bit、最低 16k
- 每条音频静音段（幅值小于 800）不超过 1min

5) 测试集制作方法：

- 单噪测试集：保证噪音在播放时分贝在±5db 内起伏，不符合要求的音频片段，可剔除，或调整幅值。
- 循环噪测试集：拼接音频时需要做能量规整，以保证整个音频在播放时分贝在±5db 内起伏。
- 混合噪测试集：由于是在实际场景中录制的，因此没有音量控制的要求，但尽量截取有代表性的音频片段作为测试集，例如一段 2h 的真实环境噪音，只有半小时有丰富的噪音，其他时间都非常安静，可选取这半小时作为测试集。

3.1.4 误唤醒测试集选择

1) 语料要求

睡眠场景下误唤醒语料：建议时长 24h，最低不得少于 12h

环境描述	**参考比例**
设备噪音	50%
自然环境噪音	50%

日常场景下误唤醒语料：建议时长 72h，最低不得少于 24h

环境描述	参考比例	总计
新闻	12%	82%
综艺	12%	
相声	6%	
影视剧	40%	
记录片	6%	

广告	6%	
国语歌	6%	12%
英语歌	3%	
韩语歌	3%	
闲聊	6%	6%

2) 音频质量要求

- 误唤醒测试集中使用的音频，例如电视剧、综艺等可从真实环境如客厅中录制，也可以直接从视频网站上下载音源。
- 各分类的音源需要定期更新，优先选择最新最热的电视剧、综艺、音乐等。

3) 音频格式要求

- 每种类型的音频需要有文本标记音频内容和时长。
- 每条音频的幅值不要低于 2000，高于 20000
- 要求 wav 格式、16bit、最低 16k
- 每条音频静音段（幅值小于 800）不超过 5min

4) 测试集制作方法：

- 将各种类型的音频拼接在一起，或在测试时顺序播放。
- 拼接音频时需要做能量规整，以保证整个音频在播放时分贝在 55±5db 内起伏。
- 记录每种音频的播放时间和结束时间，以便知晓误唤醒发生在哪一段音频中。

3.2 测试设备

3.2.1 设备列表

（1） 高保真音箱/人工嘴 *n：

播放唤醒/识别语料，推荐使用人工嘴，如无可用高保真音箱代替；绝对不可使用非高保真音箱播放。

播放噪声或误唤醒语料，推荐使用高保真音箱，若无可用普通蓝牙音箱代替，但需要保证音质清晰。

播音需要使用频率响应失真度较小的高保真音箱，可参考的监听音箱参数指标见下表：

频率响应精准度	（±2.5dB）74Hz-18KHz
失真度(THD)	≤0.5%
灵敏度	≥93dB

（2） 音响支架 *2：

1 个用于放置播放唤醒词的高保真音箱，1 个用于放置播放噪音的高保真音箱；

（3） 个人笔记本电脑 *2：

1 个用于播放唤醒语料，1 个用于播放噪声语料；

备选方案：若有声卡设备，可使用一台笔记本电脑同时控制两台音箱

（4） 声级计 *1：

用于噪声、混响、声源信号强度测量；

（5） 量角器 *1：

用于测量方位角，放置播放唤醒词/噪声的高保真音箱和麦克风阵列设备等；

（6） 卷尺 *1：

用于测量设备之间的距离；

（7） 音频线 *若干（至少 2 条）：

用于将笔记本电脑连接到高保真音箱的音频线。

3.2.2 推荐设备

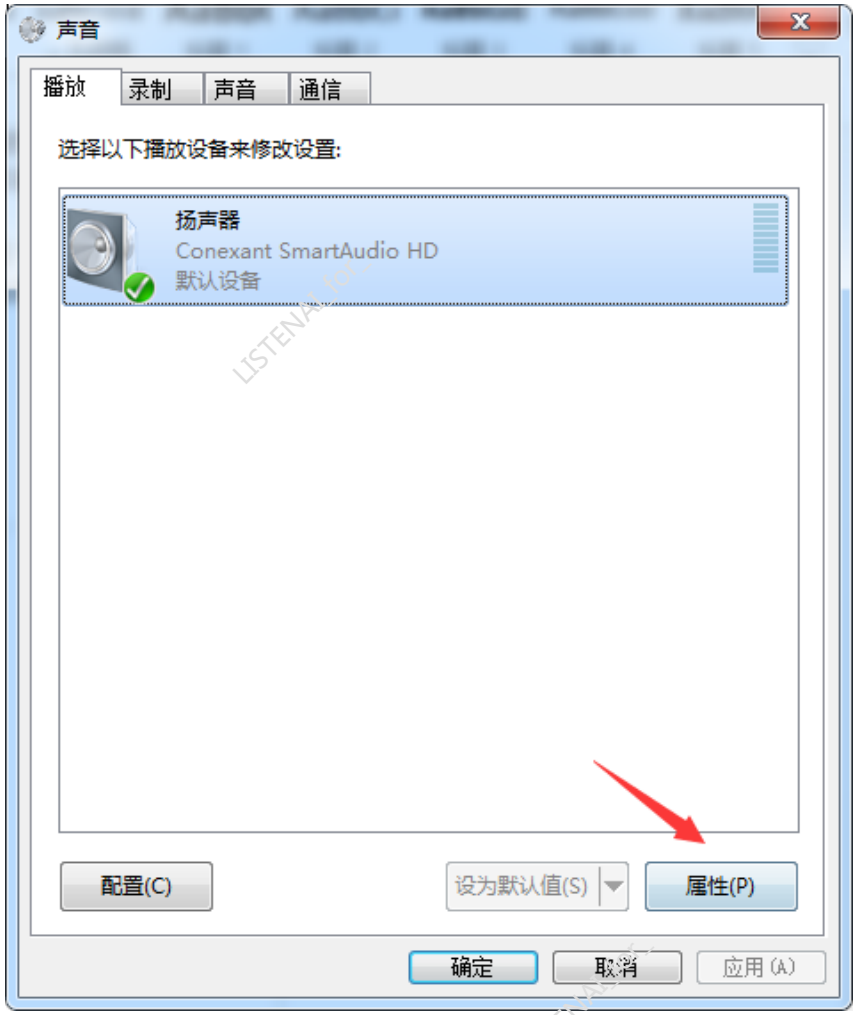
设备	功能描述	型号	厂家
人工嘴	实现播放识别与唤醒语料	AM3000	瑞森新谱
人工嘴支架	支撑人嘴并实现高度调节	\	瑞森新谱
人工嘴功放	实现驱动人工嘴播放声音	PA4000	瑞森新谱
高保真音箱	实现播放识别与唤醒语料	8010A	真力
声卡	实现背景噪声喇叭播放通道配置	fireface UC	RME 或 Gras
电动转台	实现控制产品绕自身中心转动	电动转盘 NA450	转盘王
声级计	实现系统语音/噪声的声压级测量，控制 SNR	AWA5636-1	杭州爱华

3.2.3 设备使用注意事项

1、 高保真音箱的使用：

（1） 当使用 widows 或者 Mac 电脑上的音频播放应用软件播放音频时，把应用软件的播放音量设置到最大，在音量的调整上，统一采用仅调整 Windows 或者 Mac 操作系统的音量。

（2） 使用高保真音箱播音时，为确保播放音量的一致性，需要做以下设置：在 Windows 电脑上设置“播放设备”，如下图，播放标签页的喇叭/耳机，点击“属性”，设置“增强页签，勾选“禁用所有声音效果”，应用确认即可。Mac 电脑没有此项设置，无需操作。



2、声压计的使用与测量

（1）声压计需要定期送往有国家质检总局认证的地方计量质量检测研究院中校准，至少每年校准一次。

（2） 测试分贝时，需将声压计的收音口置于被测设备的麦克风处中间位置进行测量，切勿触碰到被测设备，防止碰撞带来的分贝差异。

（3） 声压计统一采用 dBA，时间计权选择 slow 模式测量。

（4） 测试分贝时人声不要连续无间隙的播放，因为声音的消减需要时间，尽量保证人声间隔大于 1s

（5） 人声/噪音需要分开测量，且要保证房间无其他声音干扰。例如测试唤醒词音量时，必须关闭噪音，不要有其他人讲话。测试噪音音量时，关闭人声播放。

（6） 测试人声/噪音时，需要检查整条音频播放音量，做法是：取整条音频的前中后各 30s 作为定标音频，分别测量三段定标音频的分贝值。若测试时要求分贝是 65±5db，即要求整条音频的分贝值大部分在 65db 左右（可在 64~67db 内浮动），最高不超过 70db，最低不低于 60db。若大部分时间内，音频的分贝值都集中在 68~70 意味着音频整体音量偏高，若集中在 60~63，则意味着音频整体偏低于要求。

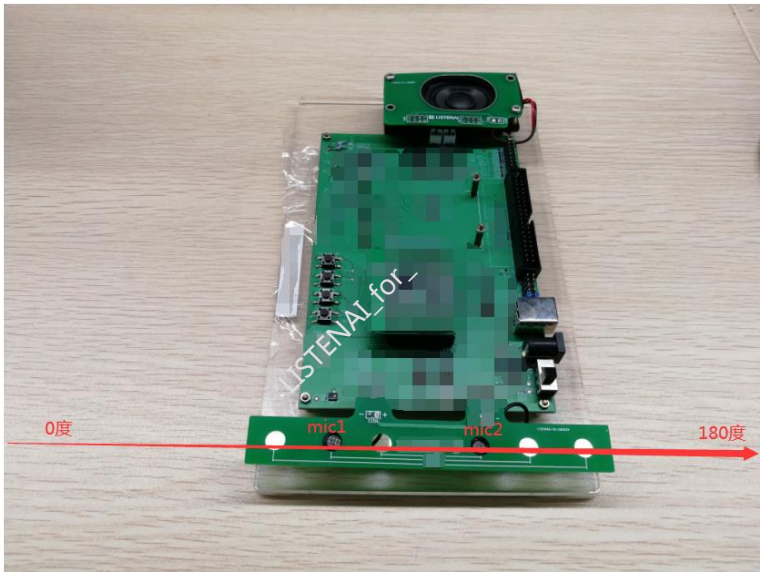
3、 距离测量

所有距离均指高保真音箱距离设备 mic 连线的中心位置的水平距离。

4、 角度测量

若设备是整机，设备的商标正面朝人摆放，设备麦克风左右侧分别为 0 度和 180 度，在此基础上测量播放人声/噪音高保真音箱的角度；

若设备是模组，设备麦克风正面朝人摆放，以设备麦克风的连线画一条线，左右侧分别是 0 度和 180 度。



3.3 测试准入条件

3.3.1 声学结构检测

产品需经过声学结构测试后，符合声学结构标准，方可进行声学效果测试。声学结构检测项如下：

1) 麦克风阵列频率响应一致性

在安静实验室(背景噪声 35dBA),音箱在 0.8m 处播放白噪声(保障白噪声到 MIC 处的音量>=70db),使用被测设备进行录音，使用 Adobe Audition 分析音频。打开 Adobe Audition 上的振幅统计查看各个声道的平均 RMS 振幅，若相差在±3dB 范围内， 可认为麦克风阵列频率响应是一致。

2) 麦克风增益

麦克风增益需要满足在设备最大音量播放歌曲等音频时,麦克风原始录音各个通道都不截幅(各 MIC 通道、回采信号通道)，在此条件上尽可能的加大增益。

3) 整机气密性

在安静实验室（背景噪声 35dBA)，使用音箱播放白噪声（保障白噪声到 MIC 处的音量>=70db)，统计用橡皮泥堵住麦克风进声孔之后,麦克风拾音的下降量。在同一环境下进行录音,使用 Adobe Audition 统计正常录音(未使用橡皮泥堵住麦克风进声孔)与用橡皮泥堵住麦克风录音的振幅，计算出拾音下降量，拾音下降量在 15dB 以上说明麦克风的气密性是良好的（拾音下降量大，说明密封良好；拾音下降量小，说明有漏音或整机中出现谐振等现象，需要进行硬件结构排查)。

4) 整机喇叭

整机设备喇叭需通过硬件测试，保证喇叭在各频段表现无明显谐振，输出功率需要满足最大播放音量，建议不超过 100dBA。

3.3.1 软件功能检测

产品需具备基础软件功能，才能完成声学效果测试。软件功能检测项如下：

测试项	要求
音频检验	• 可录制原始音频 • 可录制降噪后音频 • 支持长时间录制以上两组音频
日志	• 唤醒日志打印正确，包含唤醒时间、唤醒次数、唤醒词 • 唤醒日志中角度打印正确 • 识别日志打印正确，包含识别结果、识别时间
回声消除	• 设备本机可播音，例如 push 音频文件播放。 • 设备可调节音量。 • 本机播音的同时，进行唤醒日志的打印及降噪后音频的录制。

3.4 测试场地搭建

3.4.1 环境混响时间测试

混响时间使用声级计采用房间冲击响应进行反向积分法测试得到，测量方法参考《GBT 50076-2013 室内混响时间测量规范》，文档另外单独提供。

如果想减小测试场地的混响，主要思路是复杂化声音反射路径。可悬挂大面积厚窗帘、铺地毯；在房间各个角落摆放家居，同时家居摆放位置错开。

如果想增大测试场地的混响，可在墙面大面积贴玻璃；减少房间内家居的数量，保持墙壁和地面的裸露。

3.4.2 房间底噪测量

测量时保持房间安静，打开声压计，待声压计中分贝值趋于稳定，该数值即为房间底噪。测试时需要注意早-中-晚，房间底噪可能会不同。若底噪变化超过±5db，会对测试结果产生影响，为了保证测试结果可复现，需要在底噪相同的环境下测试。

高端声学实验室底噪通常在 20db 以下，通用型声学实验室的底噪为 30db~40db，家居环境昼间噪音上限为 55db~60db，夜间噪音上限为 45db~50db。办公区办公期间噪音上限约为 60db~65db，午休期间噪音上限约为 45db~55db。

若想减小房间底噪，可尝试以下做法：

（1） 密封窗户，如果必须保留窗户，窗户采用隔音窗材料。

（2） 墙壁上大面积贴吸音棉，若墙壁较薄，可做一道木方（轻钢）加聚酯纤维吸音板内充吸音棉的夹层，在墙壁内部再加一层框架。

（3） 门内增加吸音棉

（4） 房间内悬挂吸音效果好的窗帘，地面铺地毯。

若想增大房间底噪，参考以上相反的做法。

3.4.3 信噪比测试

在保证实验室安静的前提下，分别单独测量唤醒/命令词测试集，与噪音语料的分贝值，测试方法参考 3.1 小节。唤醒/命令词分贝值减去噪音分贝值之差即为信噪比。在实际测试时，信噪比受多种因素影响，测试时遇到以下几项变化，都需要重新测量分贝并计算信噪比：

（1） 实验室房间底噪出现变化

（2） 播放语料的电脑、音箱的音量设置发生了变化，或硬件本身发生了变化（如更换音箱）

（3） 语料本身发生了变化

（4） 被测设备、播音音箱在房间内的摆放位置、相对距离、相对角度、相对高度差发生了变化

（5） 测试场地发生变化（如家居改装）、或更换测试场地

3.5 测试场景设计

3.5.1 测试场地选择

测试场地根据设备类型和目标应用场景进行配置，主要分为客卧、厨房、浴室、阳台四大类。本文档当前仅考虑前两种场景，浴室、阳台可结合具体应用进行参考设计。

1） 客卧测试场地

DUT（被测物）的位置有三种情况，分别时室内居中放置、室内靠墙放置、室内角落放置，具体选择哪种位置因被测物实际使用情况而定，比如风扇，可选择室内居中位置、立式空调可选择室内角落或室内靠墙。

如下图 1 所示，为 DUT 室内居中放置噪声朝里布局示意图，图 2 为 DUT 室内角落放置噪声朝里布局示意图，图 3 为 DUT 室内靠墙放置噪声朝里布局示意图。客户可按照产品实际情况选择图 1、图 2、图 3 的任意两种或者一种布局方式进行测试。在实际测试中，考虑到测试投入，建议选择单一噪声源（噪声源 1 或噪声源 2）测试即可。

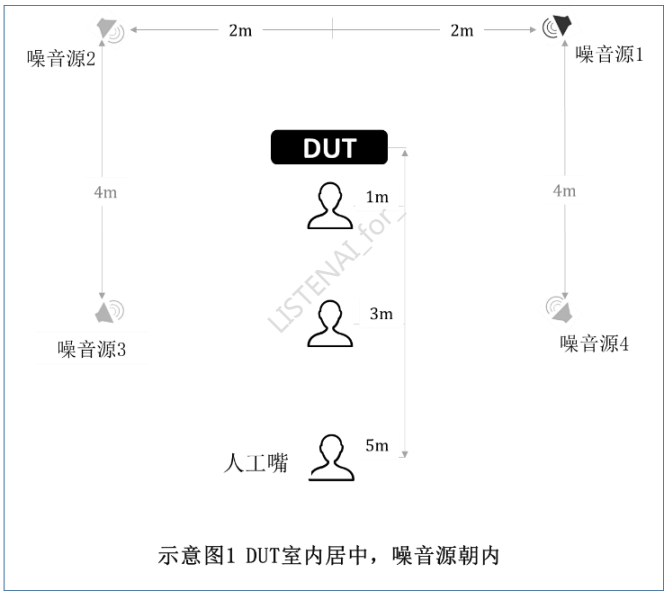


示意图1 DUT室内居中，噪音源朝内

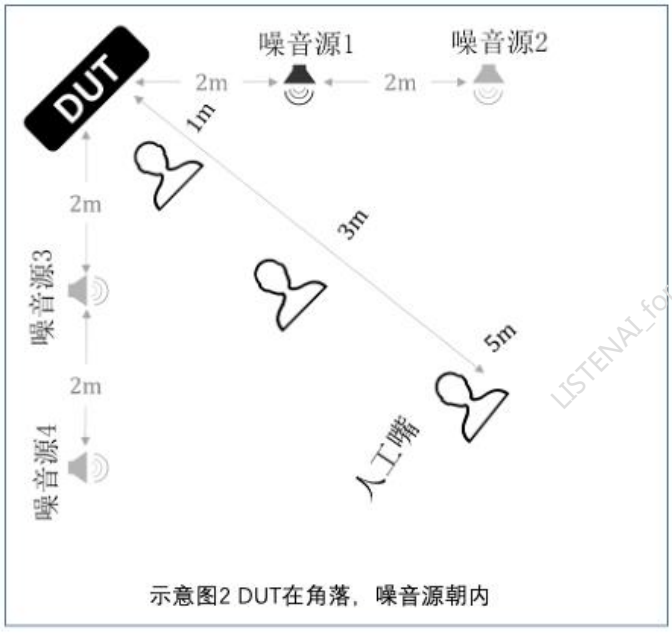


示意图2 DUT在角落，噪音源朝内

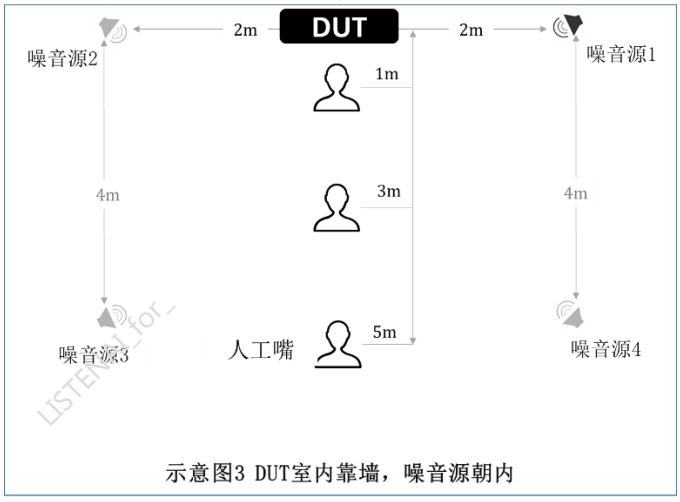


示意图3 DUT室内靠墙，噪音源朝内

2） 厨房测试场地准备

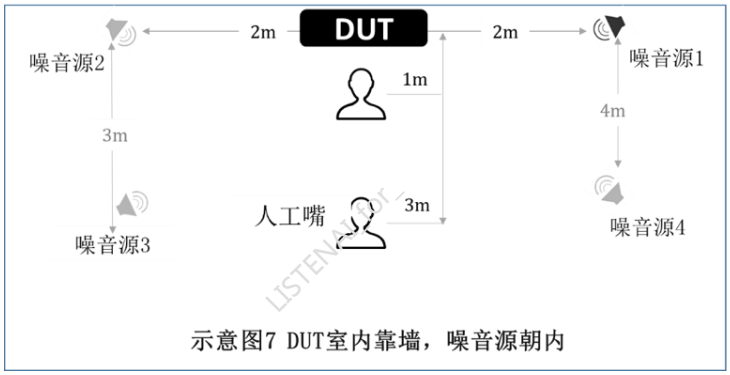


示意图7 DUT室内靠墙，噪音源朝内

3.5.2 设备摆放思路示例

以下通过三种常见的设备（空调、冰箱、音箱），解析测试设备摆放思路。

1） 空调类

空调分为挂式空调和立式空调两大类。立式空调有两种摆放位置，① 摆放在墙壁角落 ② 背靠墙但距离角落还有一段距离。

以地面为起点

	设备高度	人声高度	高度差
挂式空调	2.2m~2.4m	1.6~1.7m	0.5~0.8m
立式空调	1.7~1.8m	1.6~1.7m	0.1~0.2m

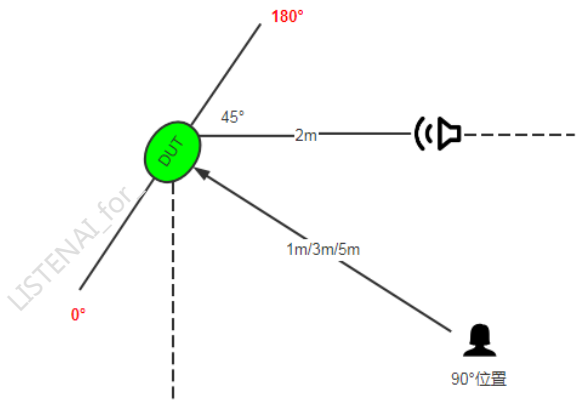
以设备为起点

	设备位置	人声位置	相对距离
挂式空调	0m	1m/3m/5m	1m/3m/5m
立式空调	0m	1m/3m/5m	1m/3m/5m

以空调麦克左侧为 0° （参考 1.3.1 小节中方位角术语释义）

	设备位置	人声位置	相对角度
挂式空调	0°	30° /90° /150°	30° /90° /150°
立式空调-靠墙	0°	30° /90° /150°	30° /90° /150°
立式空调-角落	0°	90°	90°

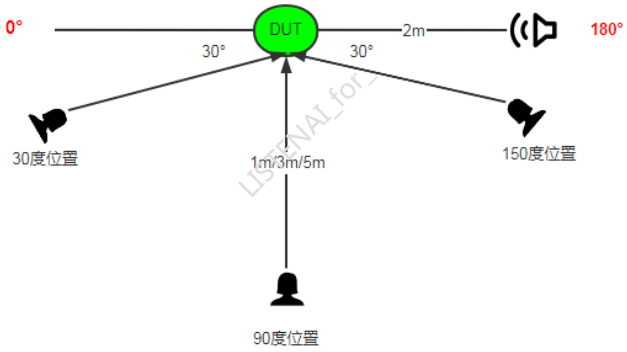
立式空调放在房间角落时，与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：



虚线代表两侧墙壁

立式空调背靠墙壁时，与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：

墙壁与 0 度与 180 度之间的线重合。



挂式空调亦可采用上图场景。

2） 冰箱类

冰箱一般摆放在客厅或厨房中，通常靠墙摆放。在厨房摆放时，有两种场景，一种为冰箱与灶台同侧，另一种为 L 型，即冰箱与灶台分别在相互垂直的墙壁两侧。

注：由于冰箱的麦克风一般在柜门显示屏附近，打开柜门后不考虑语音交互功能。

以地面为起点

	设备高度	人声高度	高度差
冰箱	1.5m~1.7m	1.6~1.7m	0.1~0.2m

以设备为起点

	设备位置	人声位置	相对距离
客厅冰箱	0m	50cm/1m/3m	50cm/1m/3m
厨房冰箱-同侧	0m	50cm/1m/2m	50cm/1m/2m
厨房冰箱-L 型	0m	50cm/1m/3m	50cm/1m/3m

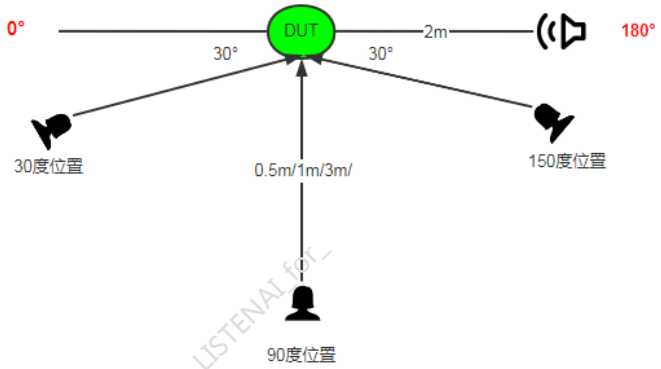
以设备为起点

	设备位置	噪音位置	相对距离	相对角度
客厅冰箱	0m	2m	2m	180°
厨房冰箱-同侧	0m	2m	2m	0°
厨房冰箱-L 型	0m	3m	3m	60°

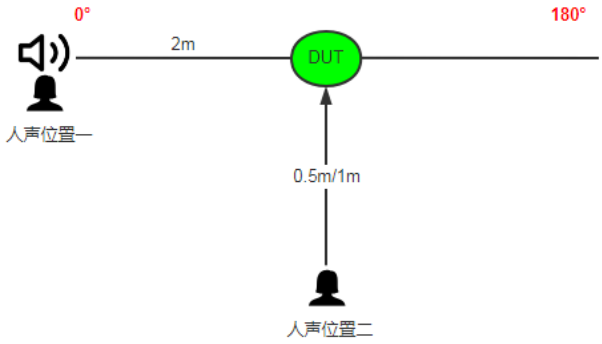
以麦克左侧为 0°（参考 1.3.1 小节中方位角术语释义）

	设备位置	人声角度	相对角度
客厅冰箱	0m	30°/90°/150°	30°/90°/150°
厨房冰箱-同侧	0m	0°/90°	0°/90°
厨房冰箱-L 型	0m	30°/90°/150°	30°/90°/150°

客厅冰箱靠墙放置，与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：

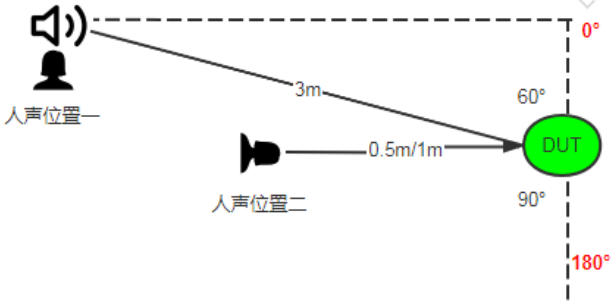


厨房冰箱-同侧，与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：



厨房冰箱-L 型，与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：

虚线为两侧墙壁



3） 音箱类

音箱一般有两种摆放方式，一种是放在桌面或床头柜上，另一种是放在橱柜上。音箱大部分语音交互场景，人多为坐着。

以地面为起点

	设备高度	人声高度	高度差
桌面音箱	0.8m~1m	1.2~1.4m	0.2~0.6m
橱柜音箱	1.6~1.7m	1.2~1.4m	0.3~0.5m

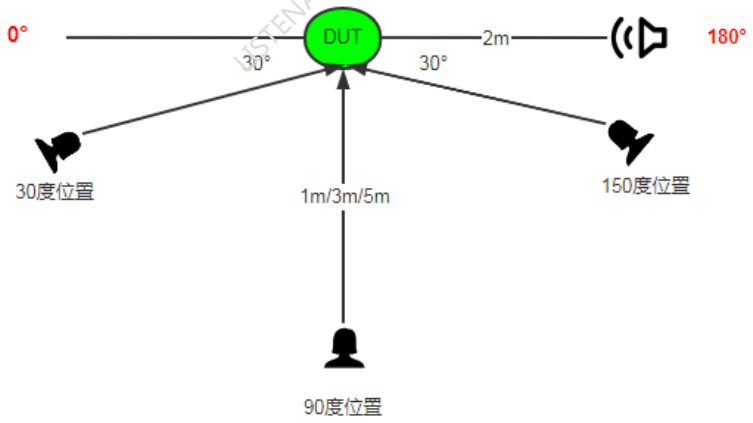
以设备为起点

	设备位置	人声位置	相对距离
音箱	0m	1m/3m/5m	1m/3m/5m

以麦克左侧为 0°（参考 1.3.1 小节中方位角术语释义）

	设备位置	人声位置	相对角度
音箱	0°	30°/90°/150°	30°/90°/150°

音箱与噪音（图中喇叭所示）、人声的距离&角度示意图如下所示：



3.5.6 测试场景说明

在明确了测试场地和设备摆放后，再结合信噪比和噪音的选择，即可完成整体测试场景设计。以上文中空调类、冰箱类、音箱类分别举例：

1） 空调类

常见的应用场景为客厅，属于中混响环境，噪音内容及信噪比参考如下：

应用场景	噪音内容	噪音分贝	人声分贝	信噪比
日间客厅	电视噪、闲聊噪、音乐噪、设备自噪、自然环境噪音	60±3db	65±3db	5db
夜间客厅	设备自噪、自然环境噪音	45±3db	55±3db	10db

结合 3.5.2 小节中对设备摆放的要求，空调类测试场景如下所示：

挂式空调

待测设备	人声与设备高度差	人声距离	人声方位	噪音距离	噪声方位	噪音类型	噪音分贝	人声分贝	信噪比
挂式空调	0.6m	1m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
挂式空调	0.6m	3m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						

			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
挂式空调	0.6m	5m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						

柜式空调

待测设备	人声与设备高度差	人声距离	人声方位	噪音距离	噪声方位	噪音类型	噪音分贝	人声分贝	信噪比
立式空调-靠墙摆放	0.2m	1m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
立式空调-靠墙摆放	0.2m	3m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
立式空调-靠墙摆放	0.2m	5m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
立式空调-角落	0.2m	1m	90°	2m	135°	日间客厅噪	60±3db	65±3db	5db
		3m	90°						
		5m	90°						
立式空调-角落	0.2m	1m	90°	2m	135°	夜间客厅噪	45±3db	55±3db	10db
		3m	90°						
		5m	90°						

2） 冰箱类

场景的应用场景是客厅和厨房，其中客厅属于中混响环境，厨房处于大混响环境。噪音内容及信噪比参考如下：

应用场景	噪音内容	噪音分贝	人声分贝	信噪比
日间客厅	电视噪、闲聊噪、音乐噪、设备自噪、自然环境噪音	60±3db	65±3db	5db
夜间客厅	设备自噪、自然环境噪音	45±3db	55±3db	10db
厨房	厨房噪、闲聊噪、设备自噪、自然环境噪音	65±3db	65±3db	0db

结合 3.5.2 小节中对设备摆放的要求，冰箱类测试场景如下所示：

客厅冰箱

待测设备	人声与设备高度差	人声距离	人声方位	噪音距离	噪声方位	噪音类型	噪音分贝	人声分贝	信噪比
客厅冰箱	0.2m	0.5m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
客厅冰箱	0.6m	1m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
客厅冰箱	0.6m	3m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						

厨房冰箱

待测设备	人声与设备高度差	人声距离	人声方位	噪音距离	噪声方位	噪音类型	噪音分贝	人声分贝	信噪比
厨房冰箱－同侧	0.2m	0.5m	90°	2m	0°	厨房噪	65±3db	65±3db	0db
		1m	90°						
		2m	0°						
厨房冰箱-L侧	0.2m	0.5m	90°	3m	60°	厨房噪	65±3db	65±3db	0db
		1m	90°						
		3m	60°						

3） 音箱类

常见的应用场景为客厅，属于中混响环境，噪音内容及信噪比参考如下：

应用场景	噪音内容	噪音分贝	人声分贝	信噪比
日间客厅	电视噪、闲聊噪、音乐噪、设备自噪、自然环境噪音	60±3db	65±3db	5db
夜间客厅	设备自噪、自然环境噪音	45±3db	55±3db	10db

结合 3.5.2 小节中对设备摆放的要求，空调类测试场景如下所示：

待测设备	人声与设备高度差	人声距离	人声方位	噪音距离	噪声方位	噪音类型	噪音分贝	人声分贝	信噪比
音箱	0.4m	1m	30°	2m	180°	日间客厅噪	60±3db	65±3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45±3db	55±3db	10db
			90°						
			150°						
音箱	0.4m	3m	30°	2m	180°	日间客	60±	65±	5db

			90°			厅噪	3db	3db	
			150°						
			30°			夜间客厅噪	45 ± 3db	55 ± 3db	10db
			90°						
			150°						
音箱	0.4m	5m	30°	2m	180°	日间客厅噪	60 ± 3db	65 ± 3db	5db
			90°						
			150°						
			30°			夜间客厅噪	45 ± 3db	55 ± 3db	10db
			90°						
			150°						

4 测试结果统计和异常排除

4.1 唤醒率

对每组测试项目，统计唤醒次数，最后统计唤醒率；

唤醒率：模块被唤醒次数/总唤醒次数；其中唤醒率又分为综合唤醒率和分场景唤醒率。

例如:某个场景下对模块进行唤醒一共100次,其中模块被唤醒99次,则唤醒率为99/100=99%;

测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）唤醒率低于 80%

2、 安静场景唤醒率每次测试结果不一致，相差超过 5%

3、 唤醒率大于 100%

排除思路：

1、 环境是否发生变化（周围有人讲话、外界有噪音），检查重新测量底噪和混响是否满足要求

2、 设备本身不稳定，如结构未达到声学评估标准，若是，需要重新检测声学结构

3、 唤醒测试集快语速或慢语速占比过高，若是，需要重新制作测试集

4、 唤醒测试集中每句唤醒词间隔太短，导致设备没有反应过来，或未成功打断设备应答语，若是，需要重新制作测试集

5、 唤醒日志打印不准确，重复打或者少打，若是，需要修改设备日志打印方式

6、 设备出现了误唤醒，也会导致唤醒率大于 100%，若是，需更换噪音内容；若是在安静状态也容易发生误唤醒，需要先调整唤醒门限再进行测试

4.2 误唤醒频度

在对应误唤醒测试场景下，将被测设备调整为对应模式，统计规定时间内误唤醒的次数；

环境描述	噪音分贝	测试时长
新闻	60±5db	≥72h
综艺	60±5db	
相声	60±5db	
影视剧	60±5db	
记录片	60±5db	
广告	60±5db	
国语歌	60±5db	
英语歌	60±5db	

韩语歌	60±5db	
闲聊	60±5db	
自然环境噪音	45±5db	
安静场景(房间底噪)	40db	≥24h

测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）测试 24h，误唤醒大于 3 次

2、 加噪场景测试 24h,误唤醒大于 10 次

3、 加噪场景测试 72*10 次，误唤醒等于 0 次

排除思路：

1、 安静场景下出现误唤醒次数较多，需要重新调整唤醒门限再进行测试

2、 加噪场景下误唤醒较多，可能是误唤醒音频中含有与唤醒词相似的词汇，或含有唤醒词

3、 测试超长时间未出现误唤醒，需检查唤醒功能是否还处于正常状态

4.3 识别率

对每组测试项目，统计识别正确句数，最后统计识别句准确率；

识别率：模块识别正确句数/总识别句数；其中识别率又分为综合识别率和分场景识别率。

例如：在唤醒的前提下，对模块进行识别交互一共 100 次，其中模块识别正确句数 99 句，则识别句准确率为 99/100=99%；

测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）识别率（句正确率）小于 50%

2、 安静场景唤醒率每次测试结果不一致，相差超过 5%

3、 识别结果重复

4、 出现大量识别为空

排除思路：

1、 环境是否发生变化（周围有人讲话、外界有噪音），检查重新测量底噪和混响是否满足要求

2、 设备本身不稳定，如结构未达到声学评估标准，若是，需要重新检测声学结构

3、 识别测试集快语速或慢语速占比过高，若是，需要重新制作测试集

4、 识别结果重复一般是日志打印问题，需要修改日志打印方式

5、 出现大量识别为空，一般有两种情况，① 唤醒率低，大量语料未唤醒；② 设备唤醒后未识别；需要提高唤醒词的音量，并检查是否命令词未对设备应答语成功打断。另外有一种极端情况是，识别测试集中的命令词，设备本身不支持。

4.4 串扰率

对每组测试项目，统计命令词指令与命令词动作意图不一致的集外串扰词句数，最后统计集外串扰率；

集外串扰率：集外串扰词句数/总识别句数。例如：对模块进行识别交互一共 100 次，其中模块识别将其中 1 句命令词动作识别成其他命令词动作，则识别集外串扰率为 1/100=1%；

测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）串扰率大于 5%

2、 单个命令词的串扰率大于 10%

排除思路：

1、 设备打印日志错误，识别结果本身是正确的，但打印成了另一句命令词

2、 若串扰率＞10%，命令词门限需要重新调整后再进行测试

4.5 打断成功率

对每组测试项目，在模块语音播报过程中进行唤醒词唤醒，统计唤醒次数，统计唤醒率。若设备不支持打断唤醒，该项可不测试。

打断唤醒成功率：播音过程中唤醒成功次数/总唤醒实验次数

分别统计两组，① 离线打断唤醒成功率：设备正在播放命令词的应答语，如“已为您调到 16 度”，用户说唤醒词让设备重新唤醒；② 在线打断唤醒成功率：设备播放音乐时，用户说唤醒词让设备唤醒。

对每组测试项目，在模块语音播报过程中进行命令词播报，统计识别正确句数，统计识别句准确率；（仅针对支持全双工模式的设备）

打断识别成功率：播音过程中模块识别正确句数/总识别实验句数；

分别统计两组，① 离线识别打断率：设备正在播放上一句命令词的应答语，用户喊另一句命令词，看是否成功打断设备的播音；② 在线识别打断率：设备处在识别状态，且正在播放音乐，用户喊另一句命令词，看能否进入识别模式。

测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）打断唤醒成功率＜50%

2、 安静场景（不加外噪，仅实验室底噪）打断识别成功率＜50%

排除思路：

1、 设备本身不支持打断，即没有回声消除功能，或功能未完善

2、 设备回声信号截幅

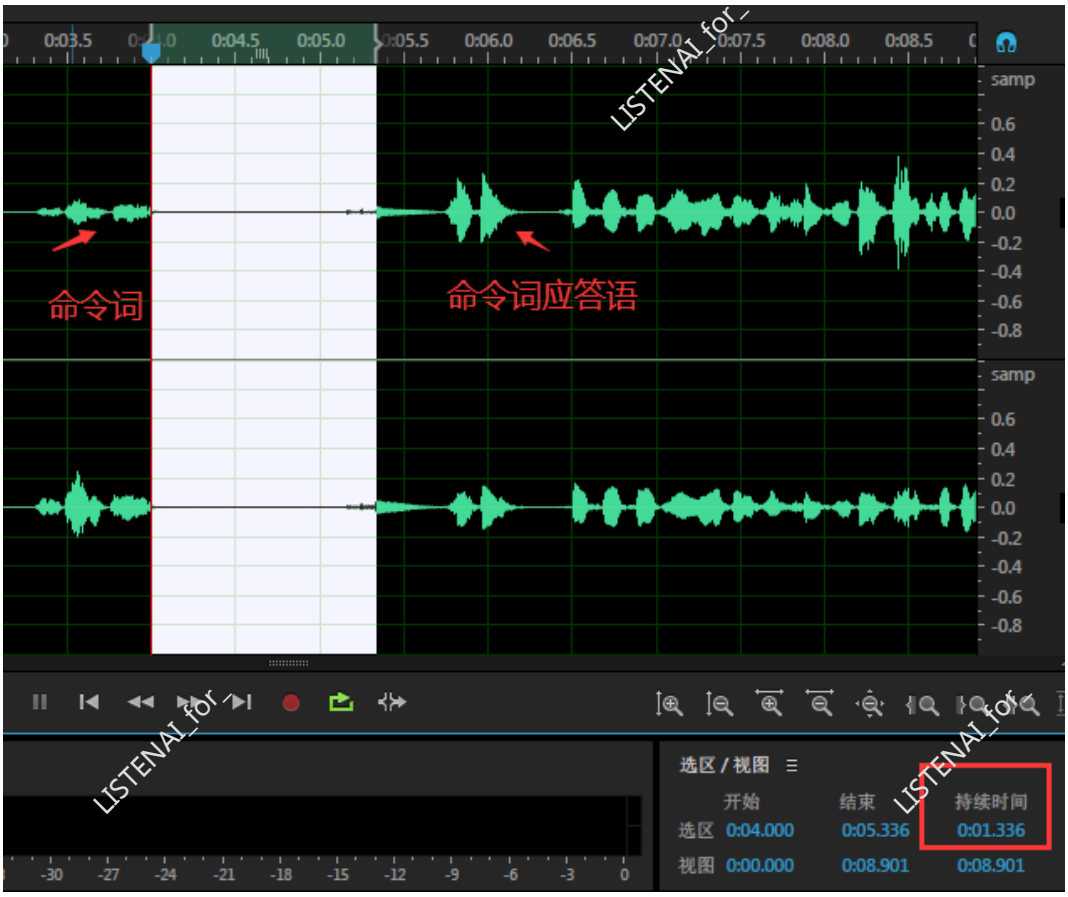
3、 设备增益设置错误，导致录入设备的命令词或唤醒词音量太小

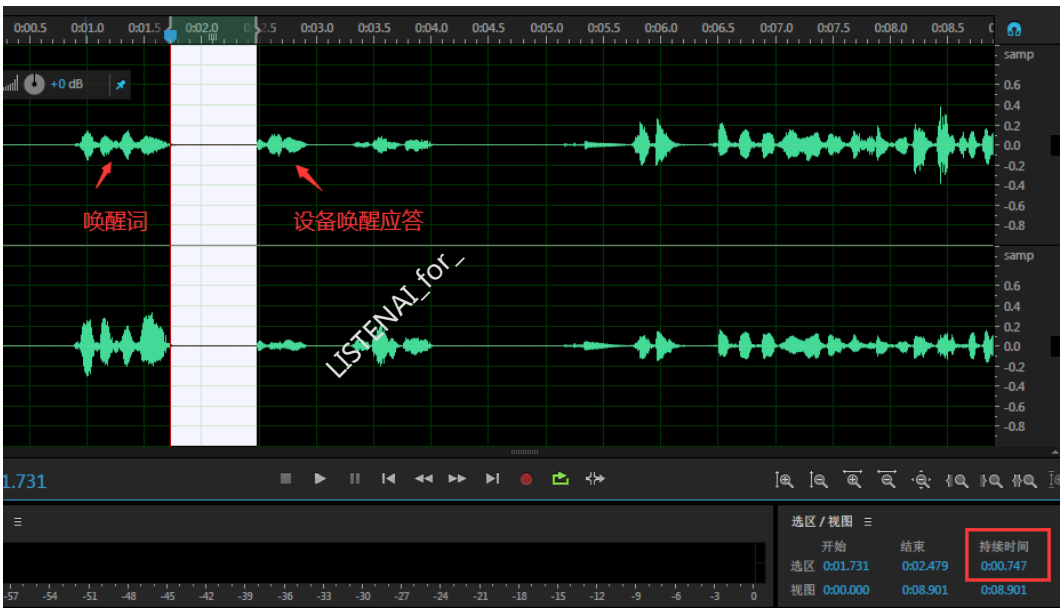
4.6 唤醒&识别响应时间

人工嘴近距离（<50cm）播放完语音指令开始到语音识别模组将识别到的指令推送到设备控制或通信端口的时间间隔。；

响应时间：语音输入结束的时刻—给出结果时刻；

参考测试方法：在测试过程中，使用手机录音，最终会录制到人工嘴/高保真音箱播放唤醒词/命令词之后，设备播报应答语。录制完成后，使用 Adobe Audition CC 或 Cooledit 软件打开音频，截取对应时间选区，获得响应时间。分别统计 50 组数据，取平均值，即得到平均响应时间。具体操作见下图：





测试结果异常辨别：

1、 安静场景（不加外噪，仅实验室底噪）唤醒响应时间＞1s

2、 安静场景（不加外噪，仅实验室底噪）识别响应时间＞2.5s

排除思路：

1、 检查设备运行是否异常，例如 CPU 接近满载，内存溢出等

2、 命令词频繁识别错误

4.7 稳定性测试

对每个被测物进行稳定性测试，分为唤醒状态下的和识别状态下的稳定性测试。统计最长稳定运行时间。

- 唤醒状态下的稳定性测试：每隔 N 秒播放一次唤醒词，运行 72 小时，无死机无重启现象，能正常识别。N 等于唤醒后到退出唤醒状态的时间加 1 秒。
- 识别状态下的稳定性测试：每隔 1 秒播放一次唤醒词，运行 72 小时，无死机无重启现象，能正常唤醒。

4.8 主观效果统计

在完整的产品上进行主观感受体验测试，尽量覆盖不同用户群，建议 20 人以上（男女平均分布），年龄比例以中青年为主，人群选取与前面所述的测试集保持一致（18~30（岁）：40%；30-50（岁）：

40%；50+（岁）：20%。）

其中体验者需按照在测试场景下通过在不同的距离、角度进行唤醒词、命令词进行设备控制，通过效果问卷进行主观打分，满分为 5 分。

环境噪声	样品	体验要点	体验得分 要求
正常家居场景	真实设备	通过简单培训，体验人员选择在测试房间内不同距离、角度进行产品的唤醒、命令词交互操作，给出主观效果体验得分；	满分 5 分；体验平均得分 3.5 分以上为合格；

评分维度	非常满意（5 分）	满意（4 分）	比较满意（3 分）	不太满意（2 分）	不能接受（0 分）
唤醒					
识别					
唤醒过程中的语音打断					
命令词识别过程中的语音打断					
对误唤醒的接受度					
对误识别的接受度					
响应速度					
播报提示音的自然度					

说明:播报提示音的的评价取体验的综合感受，具体从几个维度进行体验;可懂度、清晰度、自然度、表现力、节奏/停顿、语速、语调、音质、音色、理解费力程度。