蜂鸟离线标准化方案
语音性能测试指导书

V1.1.0

2020 年 02 月

版本变更记录

编号	版本号	变更内容
1	V0.0.1	初稿
2	V1.0.0	第一版正式版
3	V1.1.0	1) 章节 2.3，增加噪音集示例； 2) 章节 3.2，分章节说明不同测试需要烧录的版本； 3) 章节 4.2.4，补充日志说明

目录

1 名词解释

编号	名词	说明
[N1]	测试语音集	按照测试方案要求事先录制的唤醒词和离线命令词音频文件集，使用高保真音箱播放语音集来唤醒或者控制设备以测试产品的语音性能。
[N2]	外噪集	家居环境中常见的噪音集，使用电脑播放，以测试产品在噪音环境中的语音性能。
[N3]	待唤醒状态	设备未被唤醒时处于待唤醒状态，在此状态下设备无法识别命令词。
[N4]	识别状态	设备被唤醒后，由待唤醒状态切换为识别状态，在此状态下，设备可以识别命令词，直到超时退出识别状态后，再次切换为待唤醒状态。
[N5]	SNR	人声和噪声的分贝差值。
[N6]	SER	人声和设备自噪的分贝差值。
[N7]	识别超时时间	唤醒设备后，设备进入识别状态，如果在一定时间内未识别到任何命令，则退出到待唤醒状态，这段时间称为识别超时时间。

2 制定测试方案

开始语音性能测试前，首先需要根据产品的定位和项目目标制定测试方案。测试方案通常包括：测试场景、测试语音集和外噪噪音集。

2.1 测试场景

通用语音性能测试方案包含以下测试场景：唤醒率、识别率、误唤醒率和唤醒打断率。测试场景中需要确定如下因素的值，包括：房间混响，房间底噪，设备位置，外噪类型，外噪的角度、距离、高度和分贝，人声的角度、距离、高度和分贝，自噪分贝等。

房间混响（RT60）：室内声源停止发声后仍然存在的声音延续现象叫做混响，一般要求测试间混响不超过 0.4s。

房间底噪：房间中默认背景噪音的分贝值，一般要求测试间底噪不超过 40dB。

设备位置：根据产品可能的实际使用方式，确定设备在性能测试时摆放的位置，比如设备高度、离墙的距离、离地面的距离、角度等。

外噪类型：即外噪噪音集类型，详见 2.3 节。

外噪的角度、距离、高度和分贝：外噪到设备麦克的角度、距离，外噪距离地面的高度，在设备麦克处测量到的外噪分贝值。

人声的角度、高度、距离和分贝：性能测试时播放的测试语音集称为人声。人声到设备麦克的角度、距离，人声距离地面的高度，在设备麦克处测量到的人声分贝值。

自噪分贝：设备麦克处测量到的设备自噪声的分贝值，设备 TTS 播报的分贝值等。需要根据产品实际最大音量来设计测试场景中自噪的分贝值。

其中房间混响（RT60）、房间底噪和设备位置这三个因素是各个测试场景的通用因素，这几个因素确定后，将应用于每个测试场景。

2.1.1 唤醒率测试

唤醒率测试是指当设备处于待唤醒状态[N1]时被唤醒成功的概率。

除通用因素外，通常唤醒率测试还需要确定的因素如表 1 所示。可以根据产品定位设计噪声和人声相对设备同向或者不同向的测试场景，或者多噪声源的测试场景，以及不同的 SNR[5]场景。SNR 推荐选取 10dB。

表 1 唤醒率测试场景示例

测试场景编号	外噪距离	外噪角度	外噪分贝	人声距离	人声角度	人声分贝	SNR
1	/	/	/	1 米	0º	55dB	/
2	1.8m	45º	57dB	3 米	45º	67dB	10dB
3	1.8m	45º	57dB	5 米	135º	67dB	10dB
……	……						

2.1.2 识别测试

识别率测试是指当设备处于识别状态[N2]时成功识别词表里包含的命令词的概率。

除通用因素外，通常识别率测试还需要确定的因素如表 2 所示。同唤醒率测试一样，识别率测试也可以根据产品定位去设计多样的测试场景。SNR 推荐选取 10dB。

表 2 识别率测试场景示例

测试场景编号	外噪距离	外噪角度	外噪分贝	人声距离	人声角度	人声分贝	SNR
1	/	/	/	5 米	90º	55dB	/
2	1.8m	0º	57dB	3 米	45º	67dB	10dB
3	1.8m	45º	57dB	1 米	135º	67dB	10dB
……	……						

2.1.3 误唤醒测试

误唤醒率测试是指设备在产品定义的应用场景下被非唤醒词成功唤醒的概率。需要根据产品定义的应用场景中设备可能处于的环境来设计误唤醒的测试场景，比如在家居应用场景中，设备可能处于安静、外噪、设备自噪等环境。

除通用因素外，通常误唤醒率测试还需要确定的因素如表 3 所示。误唤醒率一般采用的衡量单位为次/小时。

表 3 误唤醒率测试场景示例

测试场景编号	噪声类型	噪声距离	噪声角度	噪声分贝	测试时长
1	安静	/	/	/	168 小时
2	外噪	2 米	45º	65-70dB	120 小时
4	设备自噪	/	/	65-70dB	120 小时
……	……				

2.1.4 唤醒打断率测试

对于有 AEC 功能的产品，通常还需要测试唤醒打断率。唤醒打断率是指设备有自噪时，即有 TTS 播报时，被唤醒成功的概率。

除通用因素外，通常唤醒打断率测试还需要确定的因素如表 4 所示。SER[6]取值推荐选取-10dB 和-15dB。

表 4 唤醒打断率测试场景示例

测试场景编号	设备自噪类型	设备自噪分贝	外噪距离	外噪角度	外噪分贝	人声距离	人声角度	人声分贝	SER	SNR
1	设备自噪	70dB	/	/	/	3 米	0º	67dB	-10dB	/
2	设备自噪+外噪	70dB	2 米	45º	57dB	3 米	0º	67dB	-15dB	10dB
……	……									

2.2 测试语音集

语音性能测试采用播放预先录制好的语音集进行测试，在测试方案里需要给出语音集的具体要求。语音集包括唤醒词语音集和命令词语音集。

2.2.1 语料要求

语音性能测试方案需要给出测试语音集的语料分布和大小，以及语音格式的要求。

语料分布包括：性别、年龄、地域分布比例。

语料大小是指语音集中包含的命令词条数、人数。

语音格式推荐采用 48K 采样率，16bit，单通道，wav 或者 pcm 格式。

按照这些要求，在房间混响小于 0.4 的安静环境（底噪小于 40dB）中，使用标准麦克风进行语音集的录制。录制的音频文件要求不能有破音、削波和失真。

2.2.2 语音集标准化处理

用于语音性能测试的语音集要求做标准化处理，即能量归一化。经过标准化处理后，语音集里每条命令词的平均 RMS 振幅在近似水平。推荐使用 SoX 工具处理语音文件。

语音集根据测试执行的方式，可以选择不同的语音集合方式。常用的有两种集合方式：长音频集和短音频集。

长音频集是指所有命令词在一个音频文件里，每个命令词之间用一段静音间隔，可以根据测试目的来设置静音间隔段的时长。

短音频集是指每条命令词单独一个音频文件，测试使用音频播放工具逐条播放音频文件。

2.3 外噪集

家居环境下设备的外噪类型有两种：稳态噪音和非稳态噪音。

稳态噪音是指常用家电噪音，如冰箱、空调、微波炉、抽烟机、吸尘器等电器发出声音），以及马路交通噪音等。

非稳态噪音一般值含有人声的噪音，如电视剧、新闻、音乐，交谈声等。

由此将性能测试使用的外噪集也分为两类：稳态噪音集和非稳态噪音集。

稳态噪音集是指将事先录制的多种稳态噪音短音频拼接成一个长音频文件，并进行能量归一化处理后的噪音音频文件。

非稳态噪音集是指将多种非稳态噪音短音频拼接成一个长音频文件，并进行能量归一化处理后的噪音音频文件。

外噪集示例获取地址：https://pan.baidu.com/s/1eVZtcsJvpNU3Gq2WMPiH4w，提取码：38o5

3 测试准备

3.1 测试软硬件工具

语音性能测试中需要使用到的硬件工具包括：至少两台笔记本电脑（一台连接高保真音箱播放测试语音集，一台播放噪音集）、串口线、USB 转 TTL 模块、高保真音箱、卷尺、量角器、分贝仪和音频延长线

软件工具包括：音频文件播放工具、录音工具，或者 Python2.7、 arpt 工具，串口工具。

3.2 烧录设备

语音性能测试有两种方式：手动测试和 arpt 工具自动化测试。选择不同的测试方式，烧录到设备的软件构建包也略有不同。

在做唤醒率测试时，为了更高效的进行测试，一般会将识别超时时间[7]改短，推荐改成 1 秒，即将配置文件 config.bin 中 lasr_asr 结构体里的 timeout 字段值改为 1，再进行烧录。

在做其它场景的测试时，识别超时时间可以保持为原始值。

3.2.1 手动测试

如果选择手动测试，可以根据测试需要修改识别超时时间后，再烧录到设备。

3.2.2 arpt 工具自动化测试

如果选择 arpt 工具自动化测试，需先将软件构建包里配置文件 config.bin 中 log 结构体里 arpt_enable 字段的值改为 1，再进行烧录。该字段可以控制设备在识别到命令词时是否输出指定的 arpt 格式日志（arpt 日志格式见附录 A）到串口，供 arpt 工具使用。

3.2.3 唤醒打断测试

唤醒打断测试需要在设备有自噪（即 TTS 播报）时唤醒设备，为了满足测试要求，可以采取使用较长的唤醒应答语，在设备播报唤醒应答时播放唤醒集进行唤醒打断测试。推荐使用播报时长 5 秒以上的唤醒应答语，构建一个烧录包，然后再选择是手动测试还是 arpt 工具自动化测试。

3.3 测试场景布局

按照测试场景的具体要求，将被测设备、高保真音箱和外部噪音摆放到适当的位置。

声源与设备麦克风间角度的定义：以法线为 0 度，法线与 mic1 夹角为负，如图 1 所

示。

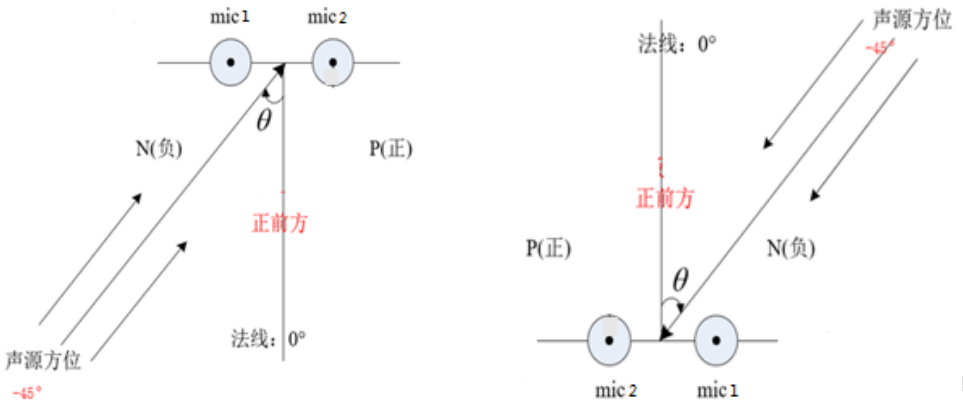

图 1 声源与麦克风间角度测量的示意图

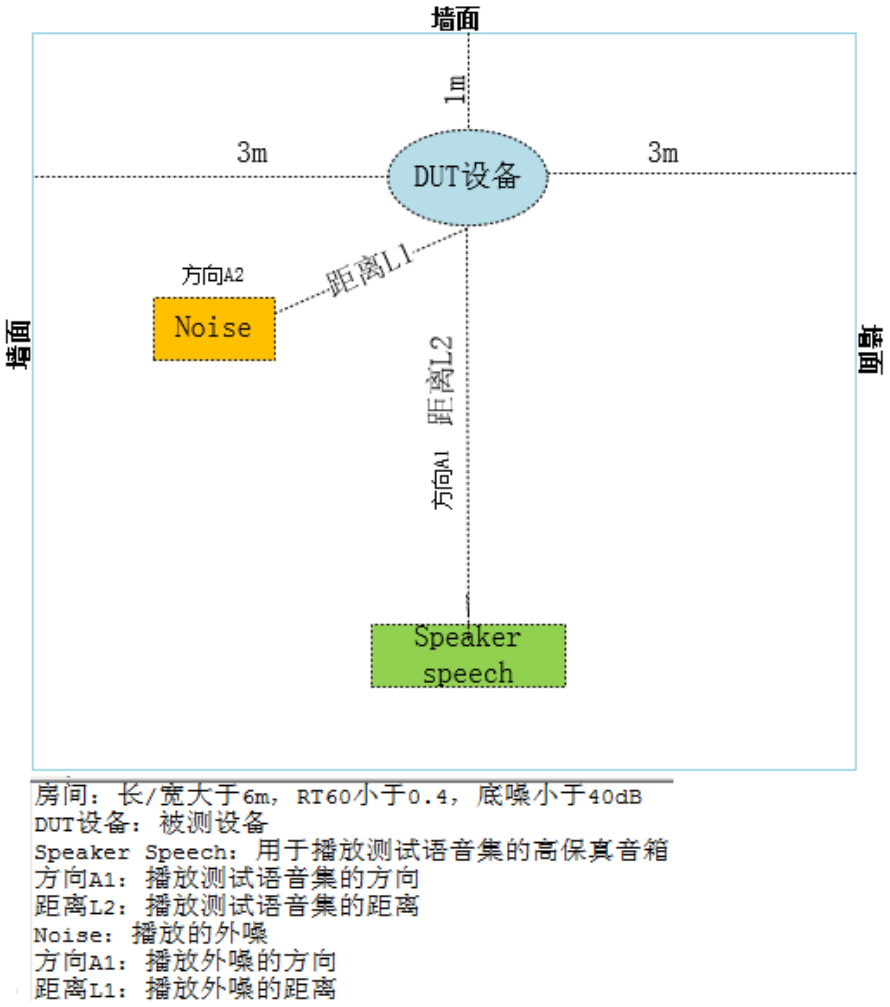

图 2 测试场景布局示意图

测试场景布局完成后，按照测试场景要求测量人声分贝和噪音分贝。

外噪分贝测量方法：用笔记本电脑播放指定噪音，用分贝仪在设备麦克风处测量多段噪音分贝值，取最高分贝作为外噪分贝。

自噪分贝测量方法：让设备播放指定噪音，用分贝仪在设备麦克风处测量多段噪音分贝值，取最高分贝作为自噪分贝。

人声分贝测量方法：从待测试语音集里每人至少挑选一条语音，用高保真音箱播放，并用分贝仪在设备麦克风处测量人声分贝值，取最低分贝作为人声分贝。

4 测试执行

4.1 手动测试

4.1.1 唤醒率、识别率、打断唤醒率测试

在连接高保真音箱的笔记本电脑上，使用音频播放工具播放语音集，并手动记录每条命令词是否唤醒或者识别成功。测试过程中，需要确保高保真音箱播放命令词时设备处于正确状态，比如测唤醒率时，需要保证播放命令词时设备是处于待唤醒状态的，否则需要重新测试该条命令词。

4.1.2 误唤醒率测试

通过 USB 转 TTL 模块和串口线将被测设备和笔记本电脑连接，打开电脑上的串口工具，连上设备后保存设备的串口日志到文件。测试结束后，查看串口日志文件，统计设备被唤醒的次数

4.2 arpt 工具自动化测试

arpt 工具是一个语音性能自动化测试工具，可以自动播放语音集、判断测试结果、存储测试日志以及统计测试数据。

使用 arpt 工具测试，测试语音集必须是短音频集（见 2.1.2 节）。

4.2.1 编写 arpt 工具的 annote 文件

使用 arpt 工具进行语音性能测试前，需要为每个测试语音集编写对应的 annote 类型文件（UTF-8 格式），放到测试语音集音频文件同级目录下，路径不要带中文，一起做为 arpt 工具的输入，arpt 工具将按照 annote 文件里列出的音频文件顺序播放。

annote 文件的第一行以*##*开头，后面再跟 6 个字段，都以 Tab 键分隔，分别为：语音集所在文件夹名、测试模式、音频格式和唤醒词文件名。从 annote 文件的第二行开始，列出待测语音集中所有音频文件名及其对应的命令词（注：命令词中包含数字的，需要写成汉字，如温度调到二十六度），以 Tab 键分隔。如图 1 所示。

语音集所在文件夹名：不要有中文；

测试模式：arpt 工具支持多种测试模式，对应不同的测试场景，分别用数字表示，如下所示：

1-唤醒率测试

2-识别率测试

3-误唤醒率测试

4-唤醒打断率测试

音频格式：包括 3 个字段，分别是语音集音频文件的采样率、分辨率和通道数；

唤醒词文件名：在识别率和误识别率测试时需要播唤醒词，此时该字段填写唤醒词文件名（唤醒词文件需与语音集的音频文件在同级目录）。 其它测试场景可以填写 N 表示不使用该字段；

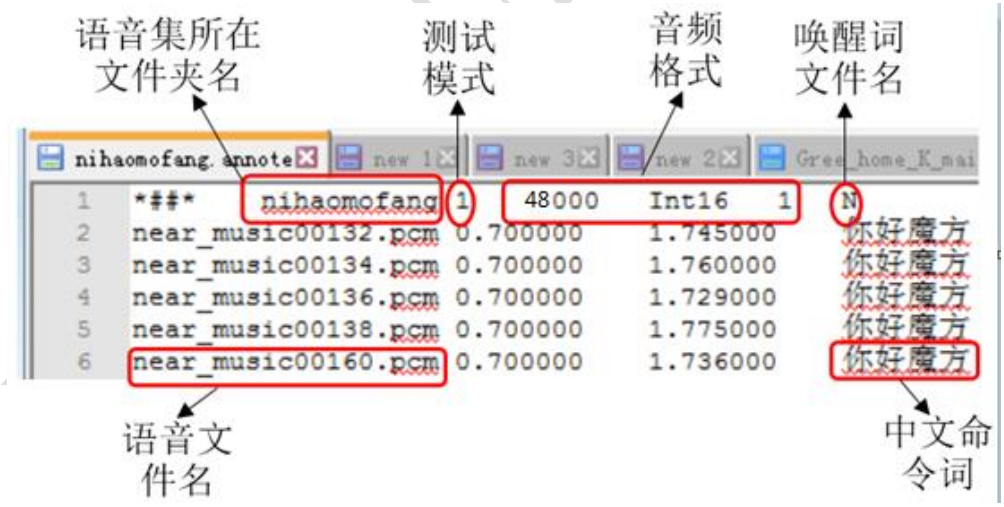

图 3 唤醒词语音集的 annote 文件

```
*###    kongtiao_comm   2   48000   Int16   1   nihaomofang.wav
chenguoliang_male_001_guiyihua_0292.wav 打开左右摆风
chenguoliang_male_001_guiyihua_0296.wav 打开左右摆风
chenguoliang_male_001_guiyihua_0298.wav 打开左右摆风
chenguoliang_male_001_guiyihua_0302.wav 打开左右摆风
chenguoliang_male_001_guiyihua_0304.wav 打开左右摆风
chenguoliang_male_001_guiyihua_0362.wav 二十八度
chenguoliang_male_001_guiyihua_0364.wav 二十八度
chenguoliang_male_001_guiyihua_0368.wav 二十八度
chenguoliang_male_001_guiyihua_0370.wav 二十八度
```

图 4 离线命令词语音集的 annote 文件

误唤醒的 annote 文件只有第一行，其中描述音频格式的三个字段含义与前面的不同，如下图所示：

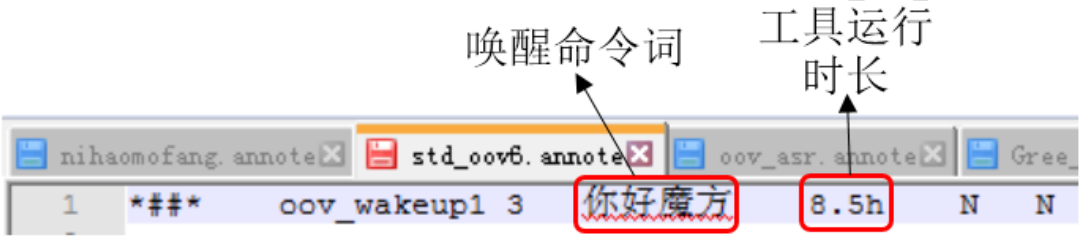

图 5 误唤醒率的 annote 文件

4.2.2 设置 arpt 工具运行参数

通过 USB 转 TTL 模块和串口线将被测设备和笔记本电脑（与高保真音箱连接的笔记本电脑）连接，如果电脑上有串口工具连接着设备，需要断开连接。

打开命令行窗口，切换到 arpt 工具所在目录（路径不包含中文），执行 python main.exe。在弹出的窗口中修改 arpt 工具需要用到的参数：

COM：包含 5 个参数，分别是 port、Baud Rate、Data Bits、Parity 和 Stop Bits。

Port: 端口，在笔记本电脑的设备管理器里查看；

Baud Rate：固定值 115200；

Data Bits：固定值 8；

Parity：固定值 N；

Stop Bits：固定值 1。

Audio Path：语音集路径，语音文件所在目录的上一级文件夹路径，不要含中文。

Version Tag：测试结果日志存放路径。

Wait TTS End Time： 播放唤醒词后等待所填时间后再去判断是否进入识别状态，根据真实情况设置，以避免工具陷入循环播唤醒词的情况。

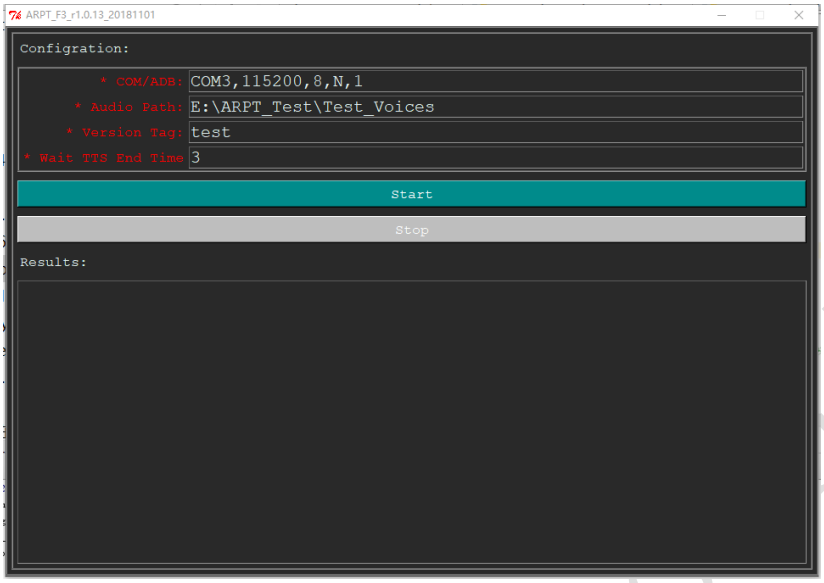

图 6 arpt 工具默认运行窗口

4.2.3 启动测试

设置好后，点击【Start】按钮启动测试，此时【Start】按钮将变成【Pause】按钮，【stop】按钮将高亮可用，如图 7 所示。

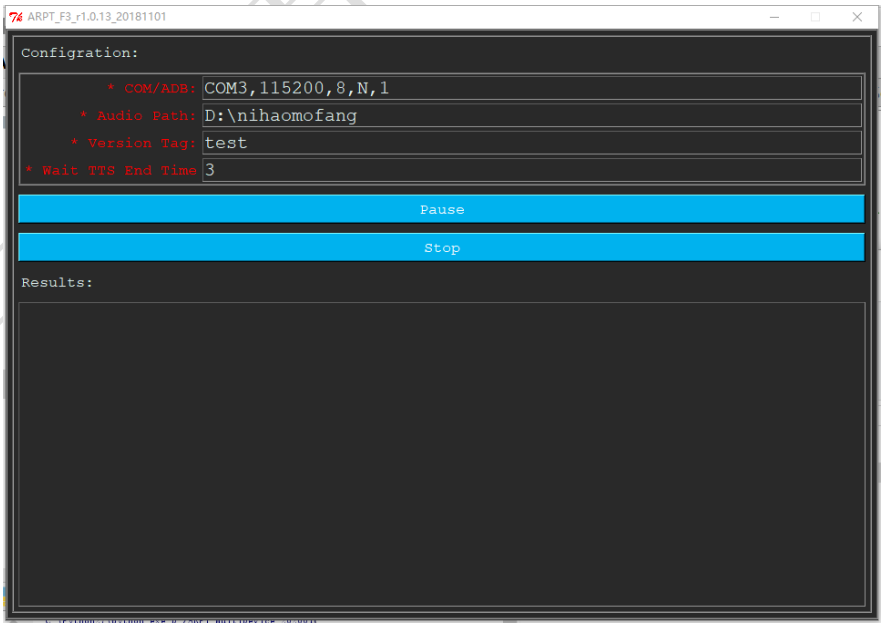

图 7 启动测试运行窗口示意图

测试过程中，可以点击【Pause】按钮暂停测试，此时【Pause】按钮将变成【Start】按钮，工具暂停播放语音文件。再次点击【Start】按钮，测试继续。

测试过程中，点击【Stop】按钮或者语音集播放完毕后，测试停止，此时【Stop】按钮将变灰不可用，【Pause】按钮将变成【Start】按钮，在运行窗口中 Result 下展示测试结果，见图 8 所示。

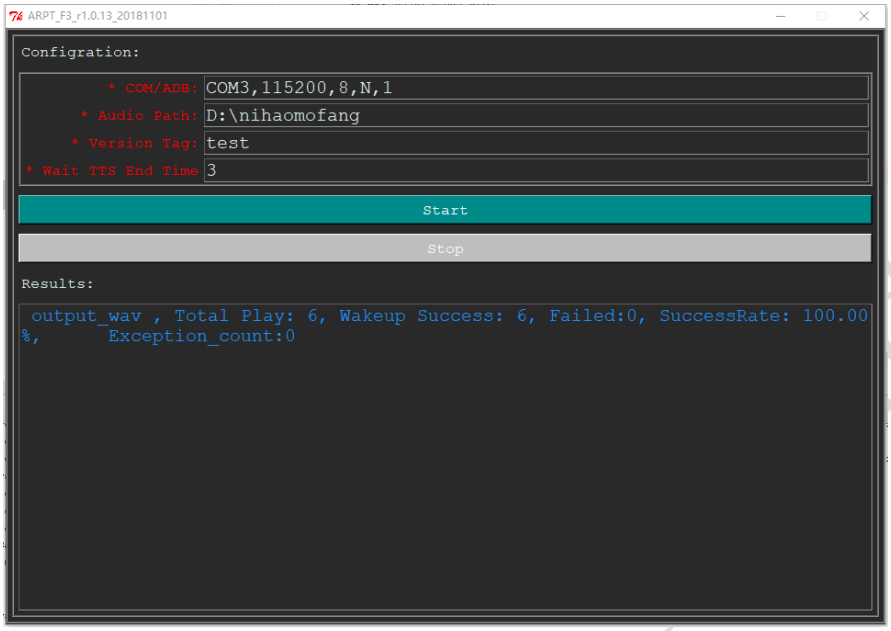

图 8 测试结束运行窗口示意图

在图 8 所示窗口里点击【Start】按钮可以直接重新测试，或者修改窗口里的配置项后点击【Start】开始新一轮测试。

4.2.4 arpt 工具的日志

arpt 工具会生成两个日志文件./log/ arpt.output 和./result/…/detailed.log，detailed.log 文件见 4.3.2 节设置的参数 Version Tag 的值就是其父目录，其下 wakeup、asr_mix、oov_wakeup 和 wakeup_interrupt 目录下分别存放唤醒率、识别率、误唤醒率和唤醒打断率的 detailed.log 测试日志。

arpt.output 文件的内容包括：

1) 播放的每个音频文件名;

2) arpt 格式打印语句；

3) 当检测到设备超时后未退出识别状态，会打印” !!!wait @enter wakeup_normal@ timeout”

detailed.log 文件的内容包括：

1） 每条命令词音频文件的播放时间，文件名，命令词内容，设备识别到的结果，分数，以及测试结果，如图 9 所示；

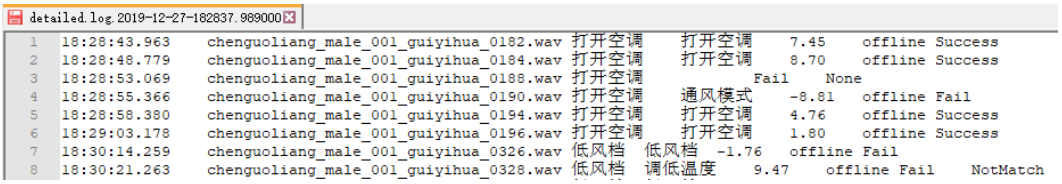

图 9 detailed.log 文件内容示意图

2） 如果某条命令词音频播放后，对应的 arpt 格式日志被冲乱导致解析测试结果失败，则将本次播放做无效处理，记录为无效次数 Exception_count，并打印异常信息“!!!Exception”到 detailed.log 文件。

3） 测试统计结果，包括播放总数、成功次数、失败次数、成功率和无效次数。

如果无效次数 Exception_count 过高，需分析原因，尽量减少日志被冲乱的现象，以提高自动化测试效率。如果 Exception_count 个数较少，可以在 detailed.log 中搜索“!!!Exception”，找到出错的命令词文件名，然后根据文件名去 arpt.output 中搜索，查看该命令词对应的日志，人工判断该条命令词的测试结果是否成功。

每次重新运行 main.exe 后，arpt.output 文件的内容会被清空，所以建议每测试完一个场景及时保存日志。每次重新运行 main.exe 后，会生成一个新的 detailed.log，文件名中带时间。

5 常见问题

1. 点击 arpt 工具 Start 按钮后，报串口启动失败。

   解决方法：查看是否有串口工具连接着设备，断开串口工具的连接后再次点击 Start 按钮。

2. 使用 arpt 工具开始测试后，没有按照预期播放语音集。

   解决方法：查看 arpt 运行窗口里 **Audio Path** 参数的值是否正确。

3. 使用 arpt 工具开始测试后，没有按照预期的顺序或者时间间隔播放命令词。

   解决方法：查看烧录到设备的软件版本，是否已将配置文件 config.bin 中 log 结构体里 arpt_enable 字段的值改为 1。

附录 A

设备状态	arpt 格式日志表示设备状态
待唤醒状态标志	enter wakeup_normal

识别状态标志	enter asr_normal
开始语音播报状态标志	TTS START
语音播报结束状态标志	TTS END
唤醒关键标识： offline_result:[wakeup_normal] 唤醒成功： KWS\taboffline_result:[wakeup_normal]\tabcommand[你好魔方]\tabscore[6.050000]\tab{离线语义解析} 唤醒失败： offline_result:[wakeup_normal]\tabcommand[你好魔方]\tabscore[6.050000]	唤醒成功： KWS offline_result:[wakeup_normal] command[你好魔方] score[6.050000] {"asr_recongize":"你好魔方","text":"你好魔方","service":"cn.yunzhisheng.setting","semantic":{"intent":{"operations":[{"operator":"ENTER_REGCON","operands":"ATTR_ASR"}]}},"general":{"type":"T","text":"你好请吩咐","pcm":"hello.pcm"}} 唤醒失败： offline_result:[wakeup_normal] command[你好魔方] score[-19.570000]
关键标识： online_json:[asr_normal] 识别成功： KWS\taboffline_result:[asr_normal]\tabcommand[打开空调]\tabscore[0.760000]\tab{离线语义解析} 识别失败： offline_result:[asr_normal]\tabcommand[打开空调]\tabscore[-30.799999]	识别成功： KWS offline_result:[asr_normal] command[打开空调] score[0.760000] {"asr_recongize":"打开空调","service":"cn.yunzhisheng.setting","semantic":{"intent":{"operations":[{"operator":"ACT_OPEN","deviceType":"OBJ_AC","deviceExpr":"空调"}]}},"general":{"type":"T","text":"空调已开机","pcm":"63.pcm","pcm_cool":"3.pcm","pcm_heat":"4.pcm","pcm_dry":"5.pcm","pcm_prefan":"6.pcm","pcm_auto":"7.pcm"}} 识别失败： offline_result:[asr_normal] command[打开空调] score[-30.799999]