ICS 33.050
M 30

团　体　标　准

T/TAF 043-2019

智能产品语音识别测评方法
第二部分：智能音箱

Evaluation Specification of Speech Recognition for Smart Products

Part 2: Smart Speaker

2019-10-14 发布　　2019-10-14 实施

电信终端产业协会　　发布

目录

前　　言

本标准按照 GB/T 1.1-2009给出的规则编写。

本标准由电信终端产业协会提出并归口。

本标准起草单位：中国信息通信研究院、华为技术有限公司。

本标准主要起草人：李玮、傅蓉蓉、张小雨、董千洲、刘毓炜、郑宗斌、高伟、孙亮、陈会

引　　言

目前智能音箱行业空前火热，互联网巨头纷纷扎堆布局智能音箱。智能音箱被认为是物联网的“新入口”，它集成了 AI 智能引擎，搭载语音交互、有声资源播放、智能家居控制、生活 O2O 服务等功能，使用场景丰富，潜力巨大。但与此同时，用户在使用智能音箱时也面临标准缺失带来的问题，比如唤醒效果差、误唤醒、语音识别错误、答非所问、音质差、抗噪能力不足等，成为阻碍智能音箱发展、无法满足用户体验的重要原因。

因此，有必要针对智能音箱制定面向消费者的语音交互测评方法，从而为智能音箱的质量评估提供必要的参考依据，帮助更好地规范智能音箱市场，保障用户的使用体验。

智能产品语音识别测评方法 第二部分：智能音箱

1 范围

本标准规定了智能音箱语音交互性能测评指标和测试方法。

本标准适用于指导第三方测评机构对智能音箱的性能测评工作。

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，仅所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

SJ/T 11380-2008 自动声纹识别（说话人识别）技术规范

GB/T 21023-2007 中文语音识别系统通用技术规范

DB50/T 489-2013 智能家居监控系统技术要求

DB50/T 488-2013 智能家居监控系统测试规范

ITU-T P.851 基于口语对话系统的电话服务的主观质量评价（Subjective quality evaluation of telephoneservices based on spoken dialogue systems）

3 术语、定义和缩略语

3.1 术语和定义

下列术语和定义适用于本文件

3.1.1 智能音箱

智能音箱是一款具备传统音箱功能，同时在此基础上拥有语音识别、语义理解和语音合成的整套语音交互功能的智能设备。产品可拥有多种互联网服务内容，成为消费者从互联网上获取信息的一种工具，例如点播歌曲、听新闻，或是了解天气预报，它也可以对智能家居设备进行控制，比如打开空调、预约电饭锅煮饭等。

3.1.2 麦克风阵列 Microphone Array

麦克风阵列是应用于语音处理的系统，它由一定数目的麦克风按规则排列组成，用来对声场的空间特性进行采样并处理的拾音技术。

3.1.3 语音唤醒 Voice Wake-up

设备在待机状态下，检测到语音输入中的唤醒词，并做出语音应答回应的操作。

3.1.4 语音唤醒词 Voice Wake-up Word

用来进行语音唤醒时的关键词，一般为短语(2~4 音节)或者用户自定义的一句话。

3.1.5 语音识别 Automatic Speech Recognition

就是让智能音箱通过识别和理解过程把语音信号转变为相应的文本技术。语音识别技术主要包括特征提取技术、模式匹配准则及模型训练技术三个方面。

3.1.6 语音合成 Text To Speech

语音合成是通过机械的、电子的方法产生人造语音的技术。TTS技术（又称文语转换技术）隶属于语音合成，它是将计算机自己产生的、或外部输入的文字信息转变为可以听得懂的、流利的汉语口语输出的技术。

3.1.7 自然语言处理 Natural Language Processing

自然语言处理（NLP）是计算机科学，人工智能，语言学关注计算机和人类（自然）语言之间的相互作用的领域。它包括知识图谱，语言理解，语言生成等多种领域。

3.1.8 自然语言理解 Natural Language Understanding

是自然语言处理(NLP)的子集。包括分词，词法分析，句法分析，文本分类等多种技术涵盖多种技术领域的应用。

3.1.9 背景噪声 Background Noise

模拟智能音箱所在环境中可能发出的扩散噪声信号。

3.1.10 信噪比 Signal-to-noise Ratio（SNR）

智能音箱发出的语音信号或用户的输入语音信号与噪音信号或干扰源信号的能量比。

3.1.11 房间混响 Room Reverberation

室内声源发出的声波能量，在传播的过程中不断经墙面来回反射以致能量逐渐衰减的过程。

3.1.12 测试语料 Test Corpus

用于测试智能音箱语音交互功能的音频集合。

3.1.13 唤醒率 Voice Wake-up Precision

假设智能音箱在唤醒时会出现两种状态，其中：
1)总共有P次类别为1的样本，假设类别1为成功唤醒。
2)总共有N次类别为0的样本，假设类别0为唤醒失败。
唤醒率的定义：

$$F = \frac{P}{P+N}$$

唤醒率反映了智能音箱的灵敏度，同时在一定程度上和误唤醒率正相关。

3.1.14 误唤醒 False Voice Wake-up

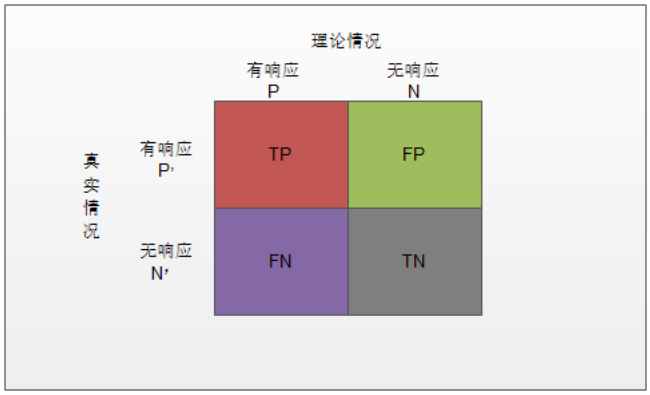

图1

假设智能音箱在使用过程中会出现两种状态，其中：

1)总共有P次类别为1的样本，假设类别1为进行了唤醒行为。

2)总共有N次类别为0的样本，假设类别0为无唤醒行为。

误唤醒理论上定义为 FP 即：

在一定连续使用音箱的时间范围内，没有进行唤醒行为却发生响应的次数。

3.2 缩略语

下面缩略语适用于本文件

ASR	Automatic Speech Recognition	语音识别
TTS	Text To Speech	语音合成
NLP	Natural Language Processing	自然语言处理
NLU	Natural Language Understanding	自然语言理解

4 功能要求

4.1 智能音箱系统框架

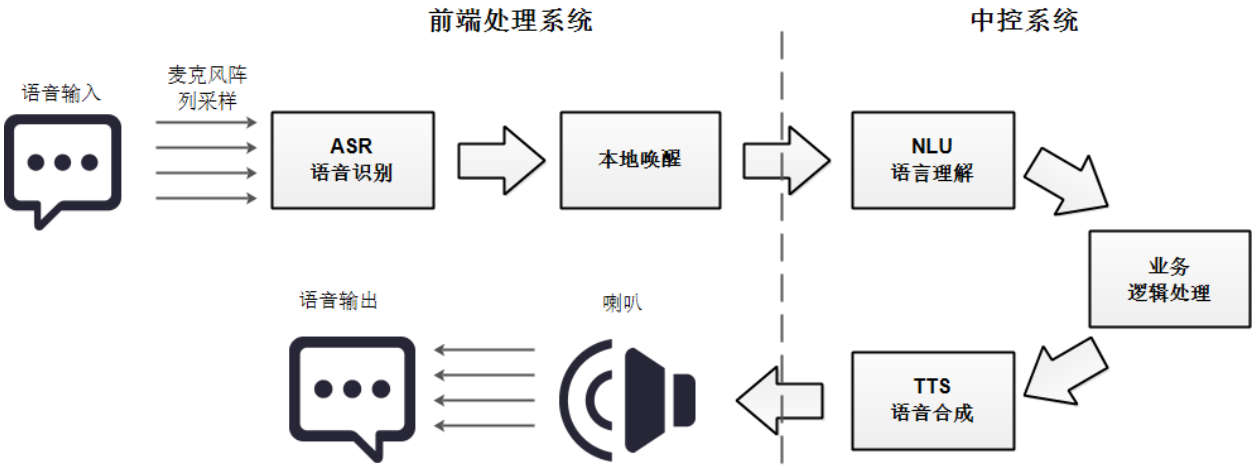

图2 智能音箱基本框架图

智能音箱系统框架主要包括语音前端交互系统，由麦克风阵列、语音识别模块和喇叭组成；中控处理单元，由语义理解模块、业务逻辑处理和语音合成模块组成，以上为最基本的智能音箱框架图。

4.2 接口

智能音箱应至少具有并行接口、通用串行总线接口、以太网接口或无线接口中的一种接口，并应符合国家相关标准的规定。

4.3 网络

智能音箱应可通过无线网络或有线网络连接到以太网，播放网络上的音频资源；另外智能音箱应该具有蓝牙(BlueTooth)功能，可以通过连接手机等BlueTooth设备播放手机里的音频。

4.4 语音交互

4.4.1 语音识别

智能音箱应能在非极限环境下，将用户输入的语音信号转为文字信号，并且在语音信号转文字的过程中要有一定的容错能力。容错能力是指发音含糊或者音近的的情况下可进行矫正的能力。产品宜应识别多种人群，如发音不准的小孩和老人，宜适应多地域人群，例如：西南、广东、广西等地域发音差异显著的人群。

4.4.2 语音唤醒

产品应能在安静场景、外部噪声和自噪声场景下通过语音识别检测到语音输入中的唤醒词，并且做出相应的应答(软硬件形式不限)。依照唤醒距离可分为远场唤醒和近场唤醒。远场唤醒指声源与音箱距离在3米及以上范围唤醒的操作。近场唤醒指声源与音箱距离在3米以内范围唤醒的操作。

产品的语音识别功能在语音唤醒阶段应支持本地识别，确保消费者的隐私不外泄，不应连接以太网。

4.4.3 语义理解

自然语言语义的表示主要有三种：分布语义，框架语义，模型论语义。现阶段智能对话平台通常采用模型语义的一个变形：领域（domain）、意图（intent）、词槽（slot）来表示语义。在实际使用时，由于智能音箱是物联网的入口，因此人与音箱的对话会涉及很多场景。音箱应对非规范的口语有一定的容错能力，根据用户查询(俗称 Query),判断 Query 所属的领域、具体意图，并解析出用户 Query 所对应的词槽，进而为后续意图的正确响应做准备。

4.4.4 语音合成

语音合成的实现涉及语言学、语音学的诸多复杂知识，因实现细节的不同，语音合成系统合成的语音在准确性、自然度、清晰度、连贯性等方面也有着不一样的表现，产品应能将文字转化为语音信号。目前典型的TTS系统可分为前端和后端两部分。前端只要处理文本的分词，韵律结构等的预处理，后端则通过声学建模，训练调参，再由声码器合成人所听到的声音。如图所示。语音合成的声音应可以正确分词，识别数字，儿化音，轻音等。由于后端的评价指标过于主观，暂时不做硬性要求。

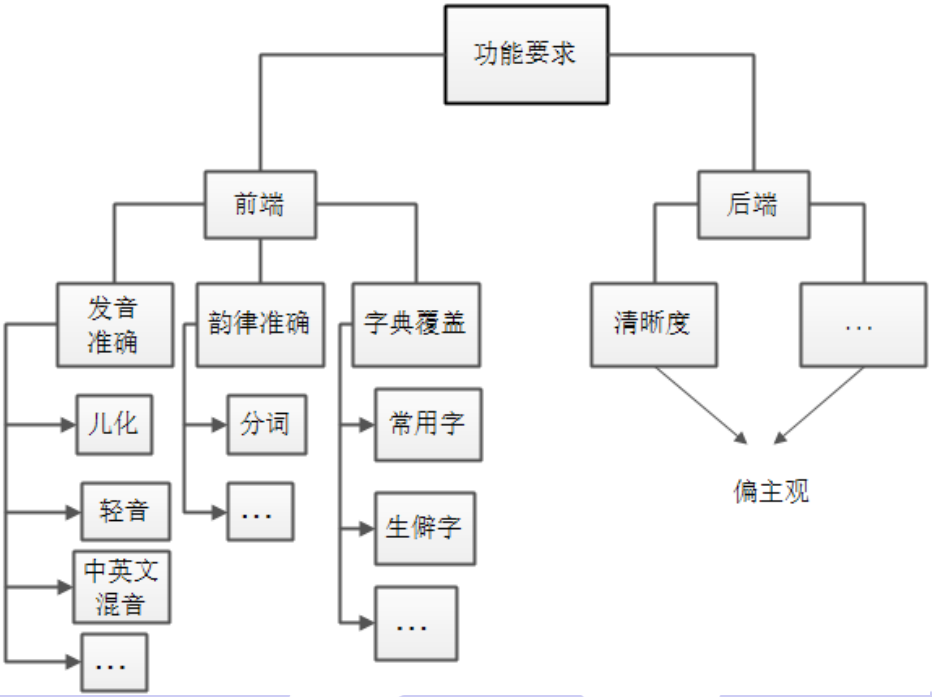

图3 语音合成要求

4.5 资源内容

智能音箱应能为用户播放丰富的音频资源，以满足作为音箱的功能。播放宜不只局限于蓝牙方式连接的本地播放模式，而宜提供丰富的网络资源音频。例如：新闻、音乐、故事，相声等。产品拥有的内容服务，宜满足其定位客户的需求。

4.5.1 基本新闻信息类

宜有基本的音乐、新闻的有声资源，可根据不同的合作方为用户提供更多丰富的新闻信息内容。

4.5.2 日常资源类

查询天气，路况，百科、航班车次、简单咨询(可为物流信息亦或物品询价等)。产品可根据自身定位提供服务的内容。

4.5.3 生活助理类

可实现语音备忘、提醒，闹钟等手机语音助手能够实现的功能。产品可进行扩展，实现语音购物，语音支付等服务。

4.5.4 控制类功能

产品应可控制音频播放，比如暂停／播放、增大音量／减少音量等。智能音箱宜可以控制家居，可对连入家庭局域网内家电进行简单的控制，比如台灯的开关、风扇的开关、电视空调的开关等。

4.6 账号和升级管理

智能音箱应具备账号和设备的绑定功能，一台设备可绑定有限个帐号，可以管理/修改账号信息，设置个人喜好等其他扩展功能(完成基本用户画像)。产品后端应有云平台或者 AI 服务平台做支撑，前端 app 上应列出合作的第三方服务资源，供用户自由选择，不可有捆绑消费设计产品。智能音箱应在

wifi 联网下支持自动升级系统；产品升级时，应具备静默升级功能，不打扰用户的正常生活。

5 语料设计和环境搭建

5.1 语料库设计

为保证智能音箱语音识别系统测试的一致性、可重复性，应采用基于智能音箱语音识别系统测试标准库的测试方法。语音识别测试标准库的建立应按照 GB/T 21023 的要求进行，通过专业录音麦克风在消音室环境下组织录制人员录制。

唤醒测试集的构成，主要是根据待测音箱声明的默认唤醒词录制，发音为中文普通话。录制人员需考虑性别、年龄分布，地域分布，单个设备的唤醒词录制数量不低于 500 句。详细要求参阅下表：

男女比例	1：1
地域要求	满足覆盖七大方言区，单一省份人数不超过总数的 30%且单一省份不少于 1%； 录音人说普通话，不可以是方言，允许不同程度的口音；
年龄段占比要求：	7~12 岁：5%
	13~17 岁：10%
	18~30 岁：60%
	31~50 岁：20%
	50+岁：5%
采集要求	录音参数为 48000hz，单声道，16bit，wav 格式，非压缩性语音文件；

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。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。

语音识别测试集的构成，需覆盖待测设备支持的所有领域，包括但不限于播放音乐/有声类播放（按歌名、歌手、流派、心情、场景、语种等方式）、广播电台、新闻、播放控制、音量控制、蓝牙、收藏功能、天气查询、闹钟、备忘录提醒、家居控制、生活辅助、娱乐对话等，测试语料的规模不少于 1000 句。（测试集涉及的技能领域可参考附录 B）

5.2 测试环境

测试环境应为混响测试环境，具体要求参见 ETSI EG 202 396-1（第 6 部分）中指定的 ETSI 普通房间配置。

- 本底噪声目标应为 28 dBSPL(A) 并且必须 < 35 dBSPL(A)
- 房间的混响时间应少于 0.7 秒，但多于 0.4 秒，频率范围在 100 hz 和 8 kHz 之间。

表1 典型的环境噪声的场景

场景编号	环境	噪声源与被测终端距离	噪声源与被测终端角度	唤醒词及语音指令在麦克风处声压级	麦克风处的环境噪声声压级	备注
1	安静	1.5m	0°、45°、90°、135°180°	50-55dB（A）	35dB（A）以下	必选

2	中噪	1.5m	0°、45°、90°、135°180°	55-60dB（A）	45-55dB（A）	必选
3	重噪	1.5m	0°、45°、90°、135°180°	60-65dB（A）	50-60dB（A）	可选
4	自噪声	N/A	N/A	50-55dB（A）	播放粉噪70dB（A）	必选

表2　待测设备摆放高度

位置编号	离地距离	备注
1	40cm	必选
2	80cm	必选

表3　目标声源位置

位置编号	与待测设备距离	与待测设备角度	发声位置	备注
1	1m	0°	站姿：距地面 150cm-162cm 坐姿：距地面约 80cm 躺姿：距地面约 40cm	必选
2	3m	同上	同上	必选
3	5m	同上	同上	必选

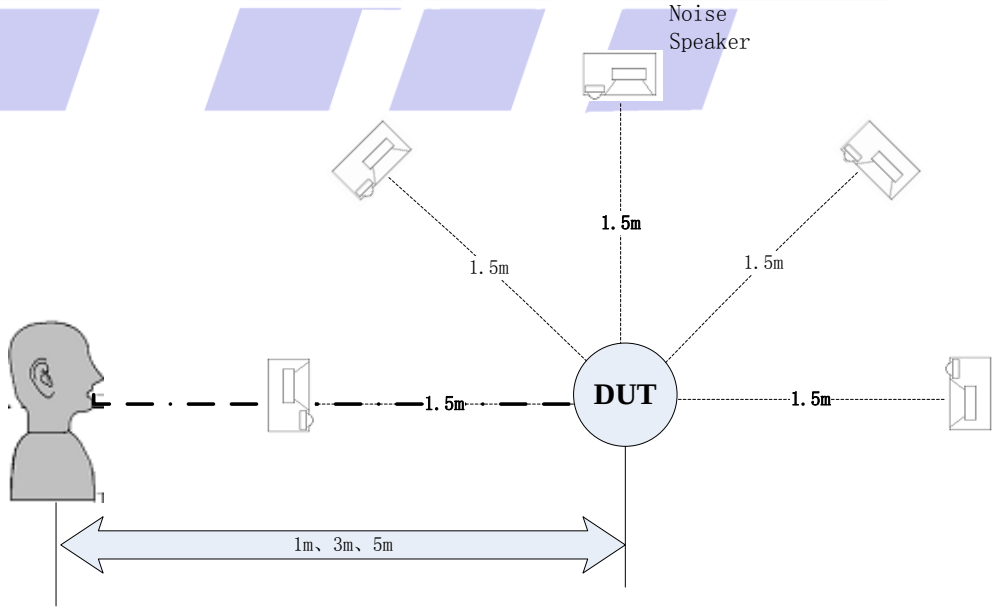

图 4 测试环境搭建

6　测试方法

6.1 唤醒率

唤醒率在智能音箱人机交互应用场景中是最为重要的一个指标。

6.1.1 测试条件

音箱开启，并完成网络配置，处于正常工作状态。

测试环境包括 5.2 节表 1 场景编号 1—4；待测设备摆放高度包括 5.2 节表 2 位置编号 1-2；目标声源位置包括 5.2 节表 3 位置编号 1-3。

6.1.2 测试方法

通过人工嘴发出唤醒词唤醒，查看智能音箱的反应（如控制灯闪烁，声音提示），判断是否唤醒成功，测试唤醒率。

若智能音箱共进行了 W 次唤醒，其中 SW 次成功唤醒。则：

唤醒率=SW/W×100%。

6.2 误唤醒率

在语音交互过程中，误唤醒是指设备听到与唤醒词相近的音或设备自身出现故障而被误触发的情况，其误唤醒率越低越好。

6.2.1 测试条件

音箱开启，并完成网络配置，处于正常工作状态。

噪声源距离待测音箱1.5米；待测音箱和噪声源距离地面高度为80cm。

6.2.2 测试方法

在普通干扰环境(背景聊天，电视播放等)下的连续 24*10 小时实测。

若智能音箱在实测中共出现了W次唤醒。则：

误唤醒率= $\frac{W}{24*10} * 100\%$。

6.3 唤醒延迟

6.3.1 测试条件

音箱开启，并完成网络配置，处于正常工作状态。

测试环境包括5.2节表1场景编号1；待测设备摆放高度包括5.2节表2位置编号2；目标声源位置包括5.2节表3位置编号1。

6.3.2 测试方法

对于语音唤醒，若语音输入的结束时刻为 t_e；智能音箱的响应时刻为 t_r（响应可取音箱的灯效响应或音箱的声音响应，此处建议取声音响应）。则：

唤醒延迟时间=t_r-t_e，根据唤醒测试集数量取平均值。

6.4 识别响应准确率

该指标用于评价智能音箱对语音识别任务的正确响应情况。

6.4.1 测试条件

音箱开启，并完成网络配置，处于正常工作状态。

测试环境包括 5.2 节表 1 场景编号 1—3；待测设备摆放高度包括 5.2 节表 2 位置编号 1-2；目标声源位置包括 5.2 节表 3 位置编号 1-3。

6.4.2 测试方法

若智能音箱在既定的识别轮数内完成了语音识别任务，则此次语音识别成功。若智能音箱共进行了 R 次特定的语音识别任务，其中 SR 次识别成功，FR 次识别出现误操作（包括未在既定的识别轮数内完成的识别、未完成识别前退出、识别无响应和错误识别）。则：

识别响应准确率=SR/R×100%；

误操作率=FR/R×100%；

识别响应准确率+误操作率=1。

6.5 识别响应时间

该指标用于评价智能音箱对语音识别任务的响应速度。

6.5.1 测试条件

音箱开启，并完成网络配置，处于正常工作状态。

测试环境包括 5.2 节表 1 场景编号 1；待测设备摆放高度包括 5.2 节表 2 位置编号 2；目标声源位置包括 5.2 节表 3 位置编号 1。

6.5.2 测试方法

对于特定的语音识别任务，若语音输入的结束时刻为t_e；智能音箱的响应时刻为t_r。则：

响应时间=t_e-t_r，根据测试集语句数量取平均值。

6.6 用户意图识别准确率

该指标用于评价智能音箱具体功能支持的广度与深度。

6.6.1 测试条件

音箱开启，并完成网络配置，处于正常的工作状态。

6.6.2 测试方法

在1m安静环境下通过人工嘴播放语音库中用户意图测试集合用例，并查看音箱是否给出正确的反馈。若智能音箱共进行了W次用户意图测试，其中SW次成功响应。则：

用户意图识别率= SW/W×100%。

附 录 A
（规范性附录）
标准修订历史

修订时间	修订后版本号	修订内容

附 录 B
（资料性附录）
测试集技能领域参考

技能	子技能
闹钟	时间明确
	时间模糊，需二次询问
时间	时钟、日期
	农历
	节日、假期
	时区
汇率	/
股票	/
查找手机	/
计算	四则运算
	幂运算
	周长、面积……
换算	/
限行尾号	/
备忘录	/
语音留言	/
天气	日出
	赏月
	空气
	温度
	台风
	穿衣
	洗车
生活辅助	计算器
	日程提醒
	百家姓
	我的快递
	邮编查询
	区号查询
	万年历
	常用号码查询
	手机充话费
	翻译
	网速测试

技能	子技能
出行	地图功能
	订票
	查询飞机票
	订酒店
社交	微博
餐饮	/
购物	/
游戏	/
声音	/
有声资源	音乐（指定人名/歌曲/流派/心情/语言/场景等）
	新闻
	广播
	小说
	评书
	相声
	脱口秀
	通话
	其他
讲笑话	/
闲聊	/
课程学习	/
古诗词	查询古诗词作者
	查询诗词
	诗词接龙
百科问答	/
播放控制	切换
	暂停/继续播放
	播放模式控制
	快进
	收藏
	定时
	音量
	查询
无线控制	蓝牙控制
软件	/

参 考 文 献