

硕士论文

论文名称：智能家居语音控制系统的设计与实现

姓 名：张加杰

学 号：ZF1821334

专业方向：人工智能

学院指导教师：王丽华

企业指导教师：王宁

企业导师单位：百度时代网络技术（北京）有限公司

论文起止时间：2020年11月 ~ 2021年 8 月

2020 年 11 月

**目录**

[1. 课题来源和意义 3](#_Toc56462362)

[1.1 课题来源 3](#_Toc56462363)

[1.2 课题背景 3](#_Toc56462364)

[1.3 选题意义 4](#_Toc56462365)

[2. 国内外相关研究现状 4](#_Toc56462366)

[2.1 国外研究现状 4](#_Toc56462367)

[2.2 国内研究现状 4](#_Toc56462368)

[3. 研究目标及内容 5](#_Toc56462369)

[3.1 研究目标 5](#_Toc56462370)

[3.2 研究内容 5](#_Toc56462371)

[3.2.1 嵌入式芯片和系统的选型 5](#_Toc56462372)

[3.2.2 语音唤醒功能 6](#_Toc56462373)

[3.2.3 语音识别 6](#_Toc56462374)

[3.2.4 控制系统 6](#_Toc56462375)

[3.3 研究方法与技术线路 6](#_Toc56462376)

[4. 关键技术难点和技术解决方案 7](#_Toc56462377)

[4.1 系统需求分析 8](#_Toc56462378)

[4.1.1功能需求分析 8](#_Toc56462379)

[4.1.2 非功能需求分析 10](#_Toc56462380)

[4.2 系统架构初步方案设计 11](#_Toc56462381)

[4.3 系统功能结构初步方案设计 12](#_Toc56462382)

[4.4 关键技术难点和解决方案 13](#_Toc56462383)

[5. 预期成果及可能的创新点 14](#_Toc56462384)

[6. 论文工作计划 14](#_Toc56462385)

[7. 参考文献 15](#_Toc56462386)

# 课题来源和意义

## 课题来源

本课题主要是受当前流行的智能语音对话音箱启发，结合我当前从事的公司相关项目，我相继在百度智能生活事业群组小度智能音箱业务部和语音技术部门工作，主要工作内容是小度智能音箱的开发，先后开发了小度智能音箱的语音SDK，OTA（Over the Air）空中升级，蓝牙配网，底层Linux系统和驱动等功能。当前智能音箱普遍重云轻端，控制功能薄弱，强依赖网络，将主要算法如ASR（Automatic Speech Recognition），NLP（Natural Language Processing）等放在云端，音箱端仅负责数据采集和TTS（Text To Speech）播放，端云之间通过一套自研协议进行通信，基于这种架构形式可以有效削减硬件成本，相应的音箱端可实现的功能受到限制，无法发挥嵌入式系统控制优势，音箱端可新增功能受限，降低了灵活性。基于此，设计一款基于深度学习的嵌入式高性能语音控制系统非常有必要，应用场景可涵盖手机，车载，智能家居，行业智能终端等，该系统可将唤醒，识别和控制功能全部放在本地系统中，可实现离线语音唤醒和识别功能，并能极大程度增加系统的控制功能，对该控制系统进行有效设计和实现是本课题的主要内容。

## 课题背景

当前，信息技术领域正在进入新的一轮发展周期——移动互联网蓬勃发展。通过移动互联网，人们可以利用各种移动终端设备随时随地的访问网络。在桌面互联网时代，键盘、鼠标和显示器很好的解决了人机交互问题。但是伴随着移动互联网的快速发展，社会的信息化、智能化也得到了快速的发展，原有信息获取与沟通方式已经不能满足人们的需求了，语音作为人类最自然、最便捷的信息沟通方式，逐渐进入人们的视线并引起了业界的广泛关注。随着深度学习技术的突破，大数据技术以及自然语言理解能力的提升，带动了一波产业热潮。国内继科大讯飞、捷通华声之后，阿里，百度，腾讯等巨头都在智能语音领域发力。语音交互的智能音箱已经走入千家万户，越来越多的人们习惯于用口语化的指令查询和播放音乐，和智能音箱聊天对话来获取了解天气情况、设置日程提醒、预订机票酒店等实用的生活服务。随着智能家居产业的蓬勃发展，照明、安防、工控等主要制造商已经推出多款智能家居产品，通过无线传感技术来增加使用舒适度。但因消费端需求逐渐成熟，通过智能语音进行集中控制更多的智能产品已日渐成为主流趋势。

## 选题意义

本选题在智能音箱普遍重云轻端的大背景下，实现一套在线离线一体的语音控制系统，相较于市面语音智能音箱具有更强力的控制功能，即使在一些没有网络的场景下仍然能离线识别控制，稳定性强，识别率高，适应场景能力更强。另一方面，基于算法和工程的完整系统可以回馈算法，更好的提升算法能力和效率。基于当前工作项目，目前主要实现工程化及周边内容，与算法和深度学习关系密切但没有深入的设计和实现过程，基于此选题，可以更深入的学习深度学习在智能语音交互系统中的实现，通过自己的思考和实现可以巩固自己所学和工作任务，更希望能在这个领域做出自己的一份贡献。

# 国内外相关研究现状

## 国外研究现状

现在国外智能语音音箱已经被广泛的推广，Amazon的echo，谷歌的GoogleHome，苹果的HomePod等纷纷上市抢占市场。2014年亚马逊的智能音箱Echo让人眼前一亮，它最大的特点是嵌入名为 Alexa语音交互大脑，它赋予了音箱语音交互的能力，让音箱能够像你的朋友一样与你交流，更重要的是，不管是购物或是信息查询等所有功能都是由语音交互来完成。可以说，Echo就是一位家庭私人助理，Echo的成功又一次带动了语音交互的发展。随后造型各异的语音交互机器人陆续推向市场，从手机端到智能音箱这种真实垂直场景的过度，需要解决的是远场拾音与固定场景语义理解的问题， 而亚马逊首先突破了这个垂直场景语义理解瓶颈，使语音交互机器人面向真实客户群。

## 国内研究现状

当前国内各厂家的音箱普遍具有语音控制功能，尤其是小米的家居生态做的比较好，而家居生态中均以在线识别为核心。2015年，京东与科大讯飞联手推出在功能上类似Echo的智能语音音箱“叮咚”。而2016年下半年开始，语音交互市场的突然爆发，几乎每隔一个多月，各类语音交互产品的用户体验效果都会出现较大的提升，而推动其发展归根到底还是依赖于语音交互技术链条的成熟。国内科大讯飞和声智科技在随后补齐了语音信号前端处理这个空缺。不管是手机还是音箱，毫无疑问产品的落地到运作带来大量真实场景下的语音数据；面对如此规模的数据量，云端能力再次彰显出来，其存储能力足以保存终端产品在日常使用过程中所产生的数据；既然云端已经拥有海量数据，剩下便是如何处理数据，而深度学习算法与硬件计算能力相结合在这方面发挥了惊人的威力，在大量数据加持下不断训练进化模型，语音识别和自然语音理解方面不断取得突破性进展，如今讯飞、搜狗、百度等语音识别准确率高达97%。在大数据时代下，国外的微软、谷歌、亚马逊，国内的科大讯飞、百度、阿里巴巴等一批实力强劲的公司已经深入部署并提供一系列语音解决方案，如亚马逊的语音服Alexa，百度的语音交互系统 DuerOS，科大讯飞的AIUI等。

# 研究目标及内容

## 研究目标

本论文的研究目标是实现一套基于深度学习的智能语音控制系统，夯实所学人工智能知识，并增进工程落地能力。具体的研究目标如下：

* 嵌入式芯片和控制系统的选型和可行性分析
* 关键词唤醒功能设计实现
* 自动语音识别ASR的设计实现
* 控制系统的设计实现
* 语音合成TTS功能设计实现

## 研究内容

该智能语音控制系统的研究内容根据研究目标可分为具体下列几个方面:

### 嵌入式芯片和系统的选型

依据市面已有的智能音箱进行芯片对比，确定芯片型号，结构设计上对麦克风采集到的信号的影响，包括结构震动，腔体效应，导音孔，音腔谐振自激放大，现在大多数产品同时有麦克风和喇叭，而喇叭放出的声音会被麦克风采集到，回采（硬件或者软件实现）到的喇叭信号对最终的AEC（回声消除）影响比较大。另外操作系统可选择Linux系统来搭建开发和执行环境。

### 语音唤醒功能

对于智能产品的用户来说,唤醒就是语音交互的第一入口,唤醒效果的好坏直接影响到用户的第一体验。用户使用时通过指定唤醒词将系统唤醒，如”小贾小贾”。该功能依托音频的声学算法和深度学习，对声学算法AEC进行研究和调优，设计关键词唤醒训练模型并进行调参训练。

### 语音识别

用户唤醒系统后，发起语音query控制家居设备，如”打开电灯”，系统收到指令音频，经过一系列声学算法，最后送入识别引擎转成文字。该模块同样唤醒模块一样，需要研究声学算法并进行语音识别模型搭建。经过声学算法对音频进行降噪增强，在语音识别中声学算法+深度学习模型具有非常好效果，这一部分主要研究语音去混响(Dereverberation)WPE降噪等算法，处理后的音频送与ASR模型进行训练。

### 控制系统

ASR识别结果为文本内容，通过字典匹配查找待执行指令，匹配成功后通过局域网通信协议将指令发与外设，电灯成功打开。这一部分需要研究局域网通信协议如MQTT或蓝牙MESH组网，红外控制等，需要在芯片选型时考虑芯片是否支持相关功能,电灯打开后发送打开成功指令给系统，系统通过TTS合成打开成功音频并播放，该模块可采用默认音频+简易TTS合成引擎结合实现。

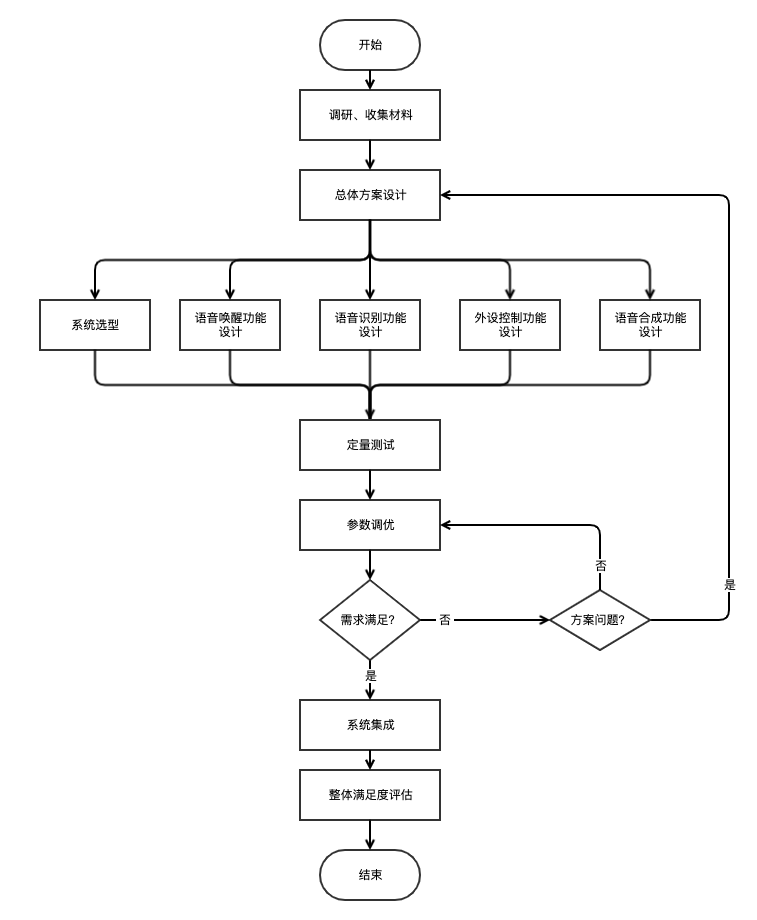
## 研究方法与技术线路

基于百度现有音箱方案对比市场主流音箱，如百度的小度音箱，小米的小爱同学，阿里的天猫精灵等选择可支撑在嵌入式系统上实现唤醒和识别算法的能力芯片。由于当前主流音箱普遍性能偏低，以上现有方案只能作为参考，需要在彻底了解后进行性能评估。

论文前期主要通过文献调查法通过大量阅读人工智能相关材料和文献，结合比较研究法进行初期系统结构构造，根据研究目标形成理论基础，从而全面地、正确地了解掌握智能语音控制系统各个模块所需知识点并形成一个统一的整体，为后续高效准确完成目标打下基础。

论文中期主要通过统计分析法对唤醒，识别和语音合成模块进行量化调优，并通过模拟法在功能和性能上对比市面已有方案，进行对比研究不断改善升级

技术路线如下图3-2:



**图3-2** **研究方法与技术路线**

# 关键技术难点和技术解决方案

## 系统需求分析

### 4.1.1功能需求分析

* 外围通信控制功能

该系统硬件上应具备最基本的控制功能模组，如红外，蓝牙，WIFI。芯片选型时应当注意。

* 麦克风音频数据采集功能

该系统处于安静环境中应处于静默状态，麦克风正常拾音，送入离线唤醒引擎。该阶段需要注意硬件上支持参考路信号回采，否则做不了AEC降噪，无法在有内噪的情况下唤醒系统。

* 唤醒功能

当用户喊出关键词时可以唤醒系统。含有关键词的音频经训练好的唤醒引擎后，触发唤醒，系统唤醒后，继续将录音结果送入下一阶段，进行ASR识别。

* 网络检测功能

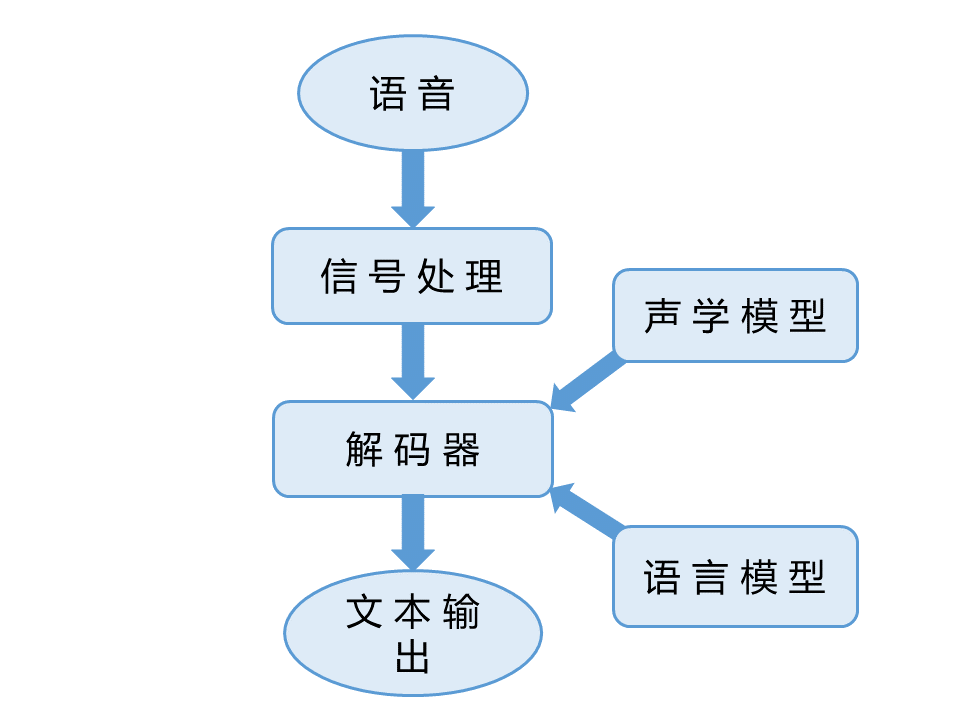
该功能可以是一个独立任务，每隔一分钟进行网络状态判断并设置网络状态标志位。

* 在线语音识别功能

进行语音到文字转化，可以接入百度的在线语音识别服务，百度提供了API接口。

* 离线识别功能

该系统应同时支持在网络状况不好或未接入网络的情况下选择离线识别引擎。该离线识别引擎基于深度学习，通过采集语料训练识别模型。语音识别系统框架如图4-1



**图4-1 语音识别系统框架图**

* 识别结果关键词分类和匹配功能

系统应对转化后的文字进行有限的关键词分类和匹配，匹配到可控制设备关键词和动作关键词后查找对应设备的控制指令，这一块可通过预先设置字典记录设备关键词，关键动作和控制指令。

* 外设通信控制功能

系统通过匹配到的设备关键词，关键动作和控制指令对设备发起控制指令。这一块要实现对应的内设的驱动和与外部设备交互的协议。

控制模块可以接受外部设备发送回来的状态，同样的通过字典的形式匹配对应状态文本。结合唤醒与外设控制，用户可以使用的功能如图4-2所示



**图4-2 使用控制用例图**

* 离线在线TTS文本转音频功能

在处理状态文本时可以通过判断网络状态选择在线或离线TTS将文本转化为可播放的音频文件。在线TTS同样可接入百度的TTS服务器，离线TTS需要通过深度学习训练TTS模型。控制结果以TTS音频与用户交互，如图4-3



**图4-3 控制结果返回图**

* 音频播放功能

系统将音频文件送入Audio驱动，通过喇叭播放音频数据。

### 4.1.2 非功能需求分析

* 唤醒率与误唤醒率需求

系统唤醒率与误唤醒率非常重要，唤醒是该语音控制系统的入口，若唤醒率低，会严重影响系统的可用性，要保证唤醒率在90%以上（100次唤醒有90次以上可以唤醒）。另一方面，误唤醒率高，也会在用户没有操作的情况下对用户造成干扰，误唤醒率应控制在2次每12小时（电视噪音下）以下。

* 识别率，字准句准需求

系统的识别率是影响用户控制体验的另一个关键指标，在线识别率依赖于服务商，识别率会比较高，离线识别字准应在90%以上，句准应在82%以上。

* 安全性需求

安全性需求是一个比较重要的问题，因为系统会不断录音（送入唤醒引擎），会有泄露用户隐私的风险。本地录音数据应使用AES加密存储。另外防止系统入侵，应关闭相关登录接口，如串口，telnet，ssh等服务。

* 外设控制满足度需求

外设控制应满足用户需求，在用户表达不够明确的情况，应尽量揣摩用户心理，如”小贾，小贾，空调”，若当前空调处于关闭状态，可以发送空调打开指令，用户控制满足度应在80%以上。

* 完整性需求

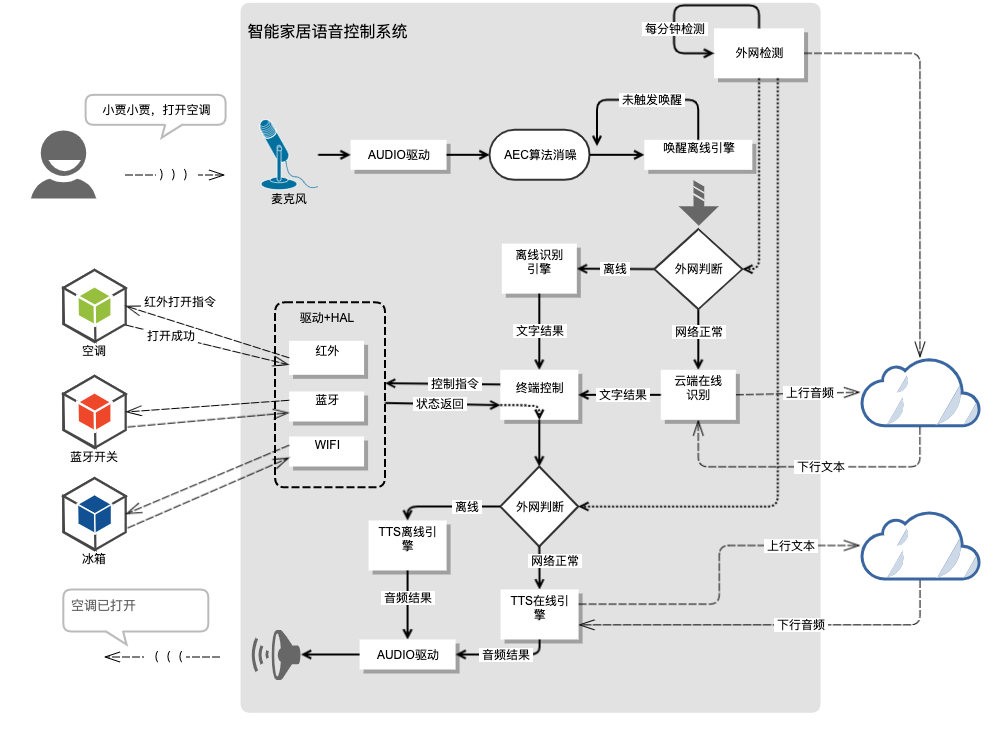
该系统应满足至少三个智能家电的控制，并且控制能力包括开、关、一项自带功能控制三种能力。

* 可扩展性需求

在满足最基本完整性需求下，系统应具备充分的扩展性，可扩展能力如下：一唤醒词定制；二离线识别内容扩展；三可控制外设的扩展等。

## 系统架构初步方案设计

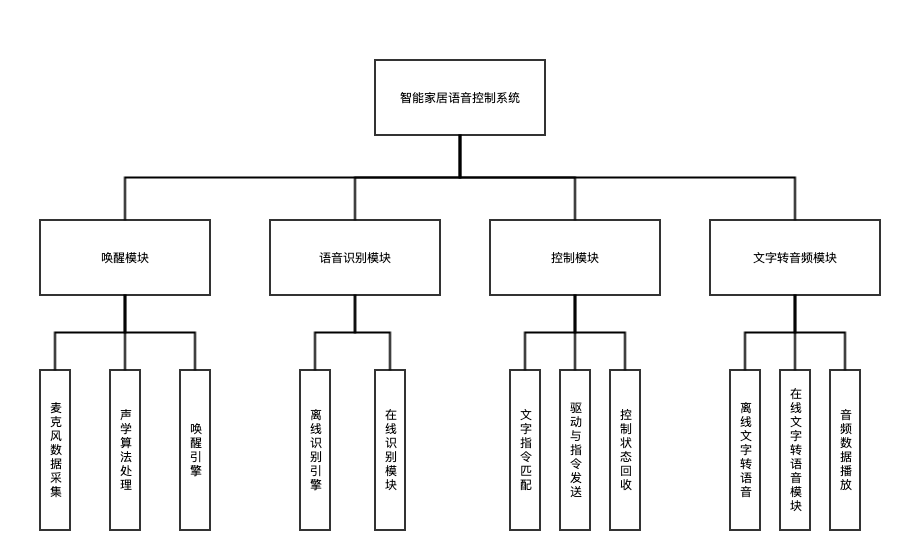
如图4-4是整个语音控制系统的初步方案设计图，示例音频为用户通过”小贾小贾”唤醒系统，指令是”打开空调”，系统将空调打开后播放”空调已打开”。



**图4-4 智能家居语音控制系统方案设计图**

## 系统功能结构初步方案设计

结合语音控制和智能家居的实际情况，并通过对该系统的需求分析设计出的系统的功能结构如图4-5所示



**图4-5 系统功能结构图**

唤醒模块：该模块包含麦克风数据采集并通过生效算法模块进行处理，最终将数据送入唤醒引擎

语音识别模块：该模块主要包含离线识别和在线识别引擎。在线识别引擎主要通过三方接口高质量的完成ASR转化；离线识别通过深度学习训练的识别模型，当在线识别不可用时采用离线识别，大大提高系统的可用性。

控制模块：主要包含文字指令匹配，外设的驱动控制和指令发送，另外该模块还接受外设的指令执行结果，最终用来通知用户指令的执行情况。

文字转音频模块：该模块包含离线文字转语音模块，在线文字转语音模块和音频数据播放。

## 关键技术难点和解决方案

本课题的关键技术难点在于离线语音识别模块。

该模块计划采用深度学习方法解决。因为语音识别的目的是对给定的波形序列，可以得到相应的单词或者字符序列。因此语音识别可以被看作是一个信道解码或者模式分类问题。统计建模是目前主流的语音识别方法。基于统计建模框架，对于给定语音波形序列O，我们可以采用贝叶斯决策的最大后验概率（Maximum A Posteriori，MAP）估计得到最优的输出序列W\*。其中条件概率P(O|W) 表示模型生成观察序列的概率，对应语音识别系统的声学模型（Acoustic Model，AM）。似然值P(W)则表示序列W出现的一个先验概率，称之为语言模型（Language Model，LM）。如图4-6是一个标注的语音识别系统的结构框图，其主要由前端处理，声学模型，语言模型，解码器四个模块组成。解码的过程主要是利用所训练的声学模型和语言模型，搜索得到最佳的输出序列。



**图4-6 语音识别系统结构图**

# 预期成果及可能的创新点

通过本系统的研究，预期达到的目标如下：

1. 完整系统可通过语音控制智能设备，如开关，家电，窗帘，汽车空调等。
2. 支持离线语音唤醒、离线语音识别和离线TTS功能
3. 可根据实际需求出具具体的系统解决方案等。

# 芯片选型和系统搭建

# 降噪

# 语音唤醒

# 语音识别

# 控制

# 测试与效果评估

# 参考文献

[1]陈孝良.语音交互为何迭代如此之快[J].科技导报,2017, 35(3): 92-92.

[2]王智国.嵌入式人机语音交互系统关键技术研究[D].中国科学技术大学,2014.

[3]陈振锋,吴蔚澜,刘加,等.基于Mel倒谱特征顺序统计滤波的语音端点检测算法 [J]. 中国科学院大学学报, 2014, 31(4): 524-529.

[4]Li X, Li G, Li X. Improved voice activity detection based on iterative spectral subtraction and double thresholds for CVR[C]//Power Electronics and Intelligent Transportation System, 2008. PEITS'08. Workshop on. IEEE, 2008: 153-156.

[5]薛胜尧.基于改进型双门限语音端点检测算法的研究[J].电子设计工程, 2015 (4): 78-81