# 项目预研

AI智能音箱第八组

一款智能音箱需要集成多种功能，其中包括最关键的三个模型：语音识别，智能聊天和语音合成。对比模型进行选择时，应该着重注意多点因素，确保准确率，模型速度理想，模型规模合理等客观要求。

1. **语音识别模型： 这种模型用于将用户的语音输入转换为文本。常见的语音识别模型包括深度学习模型如循环神经网络 (RNN)、卷积神经网络 (CNN) 和转录解码器模型（如CTC或Attention机制），用于将声音波形转换为文字。**

模型对比

1. 准确性和性能： 查看模型的准确率和性能指标，包括识别错误率、词误率等。一个高度准确的模型可以更好地理解用户的语音指令。
2. 环境适应能力： 了解模型在不同噪音环境下的表现，以及它的噪音抑制和环境适应能力。
3. 实时性： 模型是否能够在实时或接近实时的速度内完成识别，以确保用户能够得到及时的回应。

以上是模型所必须具备的基本特性，符合要求的模型如下

1. DeepSpeech： 由Mozilla开发的开源语音识别引擎，基于深度学习技术，支持多种语言，包括中文。
2. Kaldi： 一个强大的语音识别工具包，提供多种中文语音识别模型和工具。
3. AIShell-ASR： 由中国科学院自动化研究所开发的中文语音识别数据集和模型，适用于中文语音识别任务。
4. HCHS-30： 一个中文语音数据集，包含30小时的中文普通话语音数据，适用于训练中文语音识别模型。

考虑到模型的准确性、资源消耗、实时性和适应性等因素，根据需求，选择以下两个模型：

DeepSpeech： DeepSpeech是一个由Mozilla开发的开源语音识别引擎，采用了深度学习技术。它具备不错的准确性，并且在轻量级方面也有一些优化，适用于嵌入式设备。

Kaldi： Kaldi是一个功能强大的语音识别工具包，支持各种定制和优化。虽然Kaldi的学习曲线可能较陡，但一旦掌握，就可以根据轻量化需求自定义模型和流程，使其适应智能音箱的资源限制。

最后决定选择为DeepSpeech模型，原因为优良的环境适应能力和实时性和优秀的中文适配性。安装链接：<https://blog.csdn.net/weixin_36071429/article/details/86554065>

1. **语音合成模型：语音合成模型将文字转换为自然语音。模型需要能够产生自然流畅的语音，具备多种音色和情感表达能力。WaveNet、Tacotron、Google Text-to-Speech等都是常见的语音合成模型。**

模型对比

1. 自然流畅度： 生成的语音应该听起来自然、流畅，不像机器生成的声音。模型应该能够捕捉语音的音调、韵律和情感。
2. 多音色支持： 模型应该支持多种音色，以便在不同的场景中选择适合的声音。
3. 情感表达： 能够根据文本的语气和情感，生成具有相应情感色彩的语音。

以上是模型所必须具备的基本特性，符合要求的模型如下

1. Free ST Chinese Mandarin Corpus10万余条语音文件，大约100余小时。内容以平时的网上语音聊天和智能语音控制语句为主，855个不同说话者，同时有男声和女声，适合多种场景下使用。下载链接：[ST-CMDS-20170001\_1-OS.tar.gz](https://openslr.org/38/)
2. AISHELL开源版包含178小时的开源版数据。包含400个来自中国不同地区、具有不同的口音的人的声音。录音质量高，通过专业的语音注释和严格的质量检查，手动转录准确率达到95％以上。下载链接：[data\_aishell.tgz](https://openslr.org/33/)
3. iFLYTEK TTS： 科大讯飞提供的中文语音合成API可以轻松地集成到你的智能音箱中。它是一个商业化的解决方案，适用于快速启动和部署。安装链接指南：

<https://download.csdn.net/download/zhulei2012/12889022?ops_request_misc=&request_id=&biz_id=103&utm_term=FLYTEK%20TTS%E4%B8%8B%E8%BD%BD&utm_medium=distribute.pc_search_result.none-task-download-2~all~sobaiduweb~default-1-12889022.142^v93^chatgptT3_2&spm=10>

最后选择为iFLYTEK TTS模型，理由为语音自然流畅的同时启动速度很快，还拥有多声线，多情感的表达，更加适合本项目轻量级智能音箱的聊天功能。

1. **聊天机器人模型： 聊天机器人模型负责生成自然语言响应，与用户进行交互。这些模型可以基于检索式方法、生成式方法或两者的结合，以产生有意义的对话。开源的Rasa、Microsoft Bot Framework、以及各种自然语言生成模型都可以用于此。**

模型对比

1. 自然语言理解和生成： 模型应具备良好的自然语言理解和生成能力，能够准确地解析用户的输入并生成流畅自然的回复。
2. 多模态支持： 考虑到智能音箱可能有屏幕、音响等不同的输出和输入渠道，模型应该能够支持多种输入和输出模式。
3. 上下文感知和连贯性： 模型应能够理解对话上下文，确保生成的回复与之前的对话一致，保持连贯性。

以上是模型所必须具备的基本特性，符合要求的模型如下

1. ChatGLM-6B 是一个开源的、支持中英双语问答的对话语言模型，基于 General Language Model (GLM) 架构，具有 62 亿参数。结合模型量化技术，用户可以在消费级的显卡上进行本地部署（INT4 量化级别下最低只需 6GB 显存）。ChatGLM-6B 使用了和 ChatGLM 相同的技术，针对中文问答和对话进行了优化。相关链接：<https://blog.csdn.net/qq128252/article/details/129625046?ops_request_misc=%257B%2522request%255Fid%2522%253A%2522169267063816800227437281%2522%252C%2522scm%2522%253A%252220140713.130102334..%2522%257D&request_id=169267063816800227437281&biz_id=0&utm_medium>
2. 讯飞星火认知大模型SparkDesk：SparkDesk是以中文为核心的新一代认知智能大模型，能够与人自然的对话互动的过程中，同时提供多种能力，包括内容回复能力、语言理解能力、知识问答能力和推理能力等。相关链接：<https://xinghuo.xfyun.cn/desk>

以上两个聊天语言模型都是优质之选，原因为快速的相应，相当丰富的API接口和准确率，优秀的上下文连贯性。