项目一 智能客服机器人

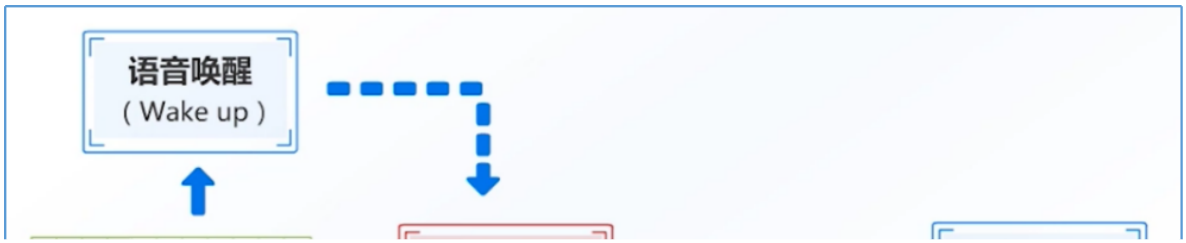
项目情景

智能客服机器人是指用电脑代替人工执行客服的任务，在如今的在线客服系统中日渐成为不可或缺的存在。得益于互联网技术、人工智能、大数据的发展，从最初代的智能客服机器人到如今经历了很多技术革新，功能也不断完善，在机器人语言处理、语义识别、关键词匹配、知识库建设乃至自主学习等方面都有了很大改进，越来越多地被运用于如今人们的工作生活。

某高校针对新生整理了一些问答（Q&A）集合放在了学校微信公众号和官网页面，但是学生和家长觉得问题不够全面而且需要从问答列表中找自己要问的问题，体验感不好。希望对现有的问答集合进行智能化升级为智能校园客服，新生或者新生家长可以语音提问，客服机器人进行语音回答。

项目导览

智能客服机器人技术路线为输入语音，系统进行语音识别并转换成文本，对转换的文本进行语义理解从答案库选择匹配的答案进行回答，回答提供文本和语音两种模式。为了节约资源，没人需要服务的时候让机器人进入休眠状态，所以在语音输入部分增加语音唤醒功能。



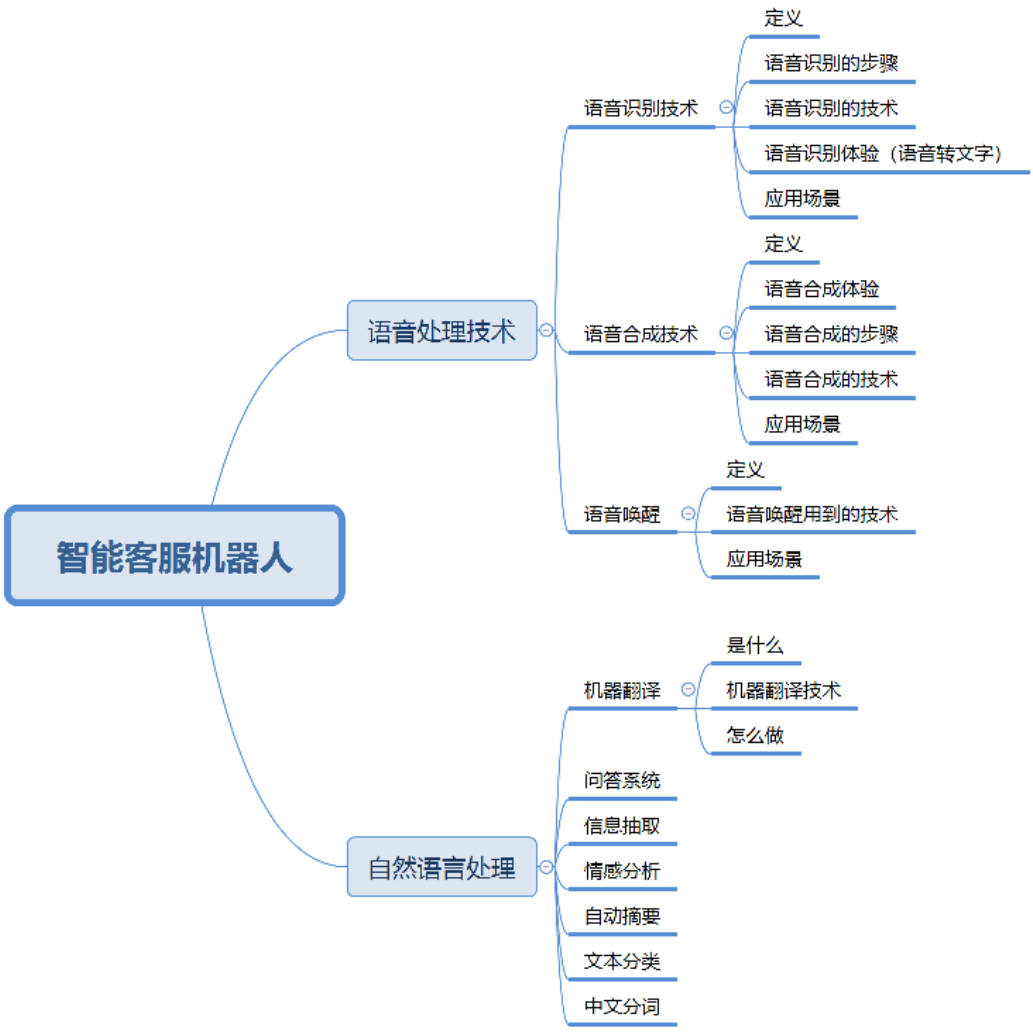
探究主题	探究活动	实现方法
语音数据采集	利用麦克风录音并保存到本地	

探究主题	探究活动	实现方法
语音转文本	将声音文件转化成中文文本	
文本转语音	1. 输入中文，转换成声音文件保存本地并播放 2.开发语音助手：将Word文档中的中文用普通话读出来。	
语音识别应用	1.天气助手 2.聊天机器人	

项目目标

1. 理解语音识别原理
2. 了解语音识别技术应用
3. 掌握语音采集和处理的方法
4. 了解自然语言处理关键技术
5. 能调用API进行语音识别的应用开发

知识导览



知识准备

语音技术是计算机领域中的关键技术，有自动语音识别技术（ASR）和语音合成技术（TTS）。

语音识别技术的研究最早开始于20世纪50年代，1952年在贝尔研究所，Davis等人研制了世界上第一个能识别10个英文数字发音的实验系统。1960年英国的Denes等人研制了第一个计算机语音识别系统。大规模的语音识别研究始于上世纪70年代以后，并在小词汇量、孤立词的识别方面取得了实质性的进展。上世纪80年代以后，语音识别研究的重点逐渐转向大词汇量、非特定人连续语音识别。上世纪90年代以后，在语音识别的系统框架方面并没有什么重大突破。但是，在语音识别技术的应用及产品化方面出现了很大的进展。比如，DARPA是在上世界70年代由美国国防部远景研究计划局资助的一项计划，旨在支持语言理解系统的研究开发工作。进入上世纪90年代，DARPA计划仍在持续进行中，其研究重点已转向识别装置中的自然语言处理部分，识别任务设定为“航空旅行信息检索”。

我国的语音识别研究起始于1958年，由中国科学院声学所利用电子管电路识别10个元音。由于当时条件的限制，中国的语音识别研究工作一直处于缓慢发展的阶段。直至1973年，中国科学院声学所开始了计算机语音识别。进入上世纪80年代以来，随着计算机应用技术在我国逐渐普及和应用以及数字信号技术的进一步发展，国内许多单位具备了研究语音技术的基本条件。与此同时，国际上语音识别技术在经过了多年的沉寂之后重又成为研究的热点。在这种形式下，国内许多单位纷纷投入到这项研究工作中去。1986年，语音识别作为智能计算机系统研究的一个重要组成部分而被专门列为研究课题。在“863”计划的支持下，中国开始组织语音识别技术的研究，并决定了每隔两年召开一次语音识别的专题会议。自此，我国语音识别技术进入了一个新的发展阶段。自2009年以来，借助机器学习领域深度学习研究的发展以及大数据语料的积累，语音识别技术得到突飞猛进的发展。

1. 什么是语音识别

语音识别，又称为自动语音识别(Automatic Speech Recognition，ASR)、语音转文本（Speech to Text，STT），其核心任务就是将人类的语音转换成对应的文字，让机器"听懂"人类的语音。语音识别技术的出现为人机交互的发展提供了新的方向。随着人工智能的发展，智能语音功能早已在车载、智能家居、手机端等场景中实现，语音对话机器人、语音助手、互动工具等智能产品也走进了人们的日常生活。

2. 语音识别的原理

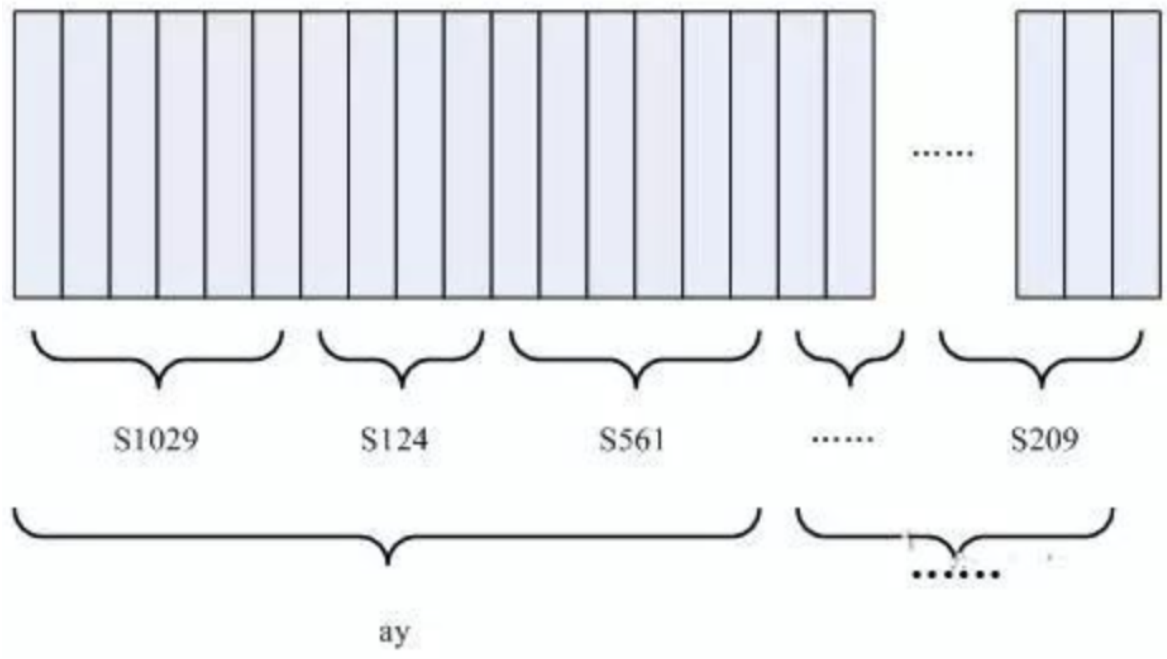
语音识别技术拆分下来，主要可分为“输入—编码—解码—输出 ”4个流程。

第1步：通过硬件输入声音信号，声音是一种波，其实就是输入一段声波文件。常见的音频文件mp3等格式都是压缩格式，必须转成非压缩的纯波形文件来处理，比如Windows PCM文件，也就是wav文件。wav文件里存储的除了一个文件头以外，就是声音波形的一个个点了，如图所示。

第2步：将输入的音频进行信号处理，帧（毫秒级）拆分，图中每个竖条是一帧，对拆分出的小段波形按照人耳特征变成多维向量信息，若干个帧信息识别成状态。这个过程叫做声学特征提取。

第3步：将第2步中的状态组合形成音素，通常3个状态组合成1个音素。

第4步：将音素组成字词并串连成句 。



经过以上四个步骤这实现由语音转换成文字了。

3.语音识别技术

（1）端点检测

端点检测（Voice ActivityDetection，简称VAD），主要作用是区分一段声音是有效的语音信号还是非语音信号。VAD是语音识别中检测句子之间停顿的主要方法，同时也是低功耗所需要考虑的重要因素。VAD通常用信号处理的方法和基于机器学习的方法来做。

（2）特征提取

特征提取就是把时域的声音原始信号通过某类方法提取出固定的特征序列，为训练声学模型准备输入。事实上深度学习训练的模型不会脱离物理的规律，只是把幅度、相位、频率以及各个维度的相关性进行了更多的特征提取。

（3）声学模型

声学模型是语音识别中最为关键的部分，是将声学和计算机学的知识进行整合，以特征提取部分生成的特征作为输入，并为可变长的特征序列生成声学模型分数。声学模型核心要解决特征向量的可变长问题和声音信号的多变性问题。事实上，语音识别的发展基本上都是指声学模型的进展。声学模型迭代这么多年，已经有很多模型，比较有代表性的是高斯混合模型（GMM）、隐马尔可夫模型（HMM）和深度学习。

- 高斯混合模型（GMM）
  高斯混合模型（英文Gaussian Mixture Model，简称GMM），是基于傅立叶频谱语音特征的统计模型，可以通过不断迭代优化求取GMM中的加权系数及各个高斯函数的均值与方差。GMM模型训练速度较快，声学模型参数量小，适合离线终端应用。深度学习应用到语音识别之前，GMM-HMM混合模型一直都是优秀的语音识别模型。但是GMM不能有效对非线性或近似非线性的数据进行建模，很难利用语境的信息，扩展模型比较困难。
- 隐马尔可夫模型（英文Hidden Markov Model，简称HMM），用来描述一个含有隐含未知参数的马尔可夫过程，从可观察的参数中确定该过程的隐含参数，然后利用这些参数来进一步分析。HMM是一种可以估计语音声学序列数据的统计学分布模型，尤其是时间特征，但是这些时间特征依赖于HMM的时间独立性假设，这样对语速、口音等因素与声学特征就很难关联起来。HMM还有很多扩展的模型，但是大部分还只适应于小词汇量的语音识别，大规模语音识别仍然非常困难。
- 深度神经网络（英文Deep Neural Network，简称DNN),是较早用于声学模型的神经网络，DNN可以提高基于高斯混合模型的数据表示的效率，特别是DNN-HMM混合模型大幅度地提升了语音识别率。由于DNN-HMM只需要有限的训练成本便可得到较高的语音识别率，目前仍然是语音识别工业领域常用

的声学模型。循环神经网络（RNN）和卷积神经网络（CNN）在语音识别领域的应用，主要是解决如何利用可变长度语境信息的问题，CNN/RNN比DNN在语速鲁棒性方面表现的更好一些。

- 通过训练语料学习词之间的关系来估计词序列的可能性，最常见的语言模型是N-Gram模型。近年，深度神经网络的建模方式也被应用到语言模型中，比如基于CNN及RNN的语言模型。
- 解码是决定语音识别速度的关键因素，解码过程通常是将声学模型、词典以及语言模型编译成一个网络，基于最大后验概率的方法，选择一条或多条最优路径作为语音识别结果。解码过程一般可以划分动态编译和静态编译，或者同步与异步的两种模式。目前比较流行的解码方法是基于树拷贝的帧同步解码方法。

4.语音识别开源平台和开放平台

语音识别的开源平台很多，但是部署应用相当复杂，特别是基于深度学习的开源平台，需要大量的计算和数据以训练引擎，这个对于一般的用户来说也是一个非常高的技术门槛。因此对于一般的创业型公司来讲，显然自己部署语音识别引擎也不划算，那么免费的开放平台就是很好的选择。

（1）Nuance NVP

Nuance是语音识别领域的老牌劲旅，除了语音识别技术外，还包扩语音合成、声纹识别等技术。Nuance Voice Platform(NVP)是Nuance公司推出的语音互联网平台，这是一个开放的、基于统一标准的语音平台产品。它能够支持客户公司已有的IT投资和基础设备，同时可以加入语音的应用。

（2）Microsoft Speech API

微软的Speech API是微软推出的包含语音识别（SR）和语音合成（SS）引擎的应用编程接口，SAPI支持多种语言的识别和朗读，包括英文、中文、日文等。但是，微软总有个问题，就是任何一个产品都得和Windows绑定。

（3）Google Speech API

这个领域自然不能少了苹果和谷歌，但是苹果打死也不会免费的，而谷歌打死也不会收费的。但是，这没有意义了，因为不管你的引擎多么优秀，现在的语音识别还是要基于云的。所以国内的众多创业用户压根用不了，甚至也访问不到。但是如果开发的产品主要部署在国外，Google Speech API可以备选的，因为这个API调用起来更加方便。

（4）科大讯飞语音

科大讯飞1999年成立，作为中国最大的智能语音技术提供商，在智能语音技术领域有着长期的研究积累，并在中文语音合成、语音识别、口语评测等多项技术上拥有国际领先的成果。科大讯飞目前提供语音识别、语音合成、声纹识别等全方位的语音交互技术。目前也是国内创业团队使用最为广泛的开放语音识别平台。

（5）百度语音

百度语音自从和中科院声学所合作以后，在贾磊带领下短时间内建立起来自己的引擎，而且打出了永久免费的口号，在很多领域抢占了一定的市场。

国内的语音识别开放平台还很多，和国外有所不同，国内开放的都是语音识别的专业公司，比如云之声、思必驰、捷通华声等等。

5.Python语音识别库

对于 Python 使用者而言，一些语音识别服务可通过各大平台提供的开源 API 在线使用，且其中大部分也提供了 Python SDK。 我们不需要从头开始构建任何机器学习模型，PyPI中有一些现成的语音识别软件包，包括：

- apiai
- google-cloud-speech
- pocketsphinx

- SpeechRcognition
- watson-developer-cloud
- wit

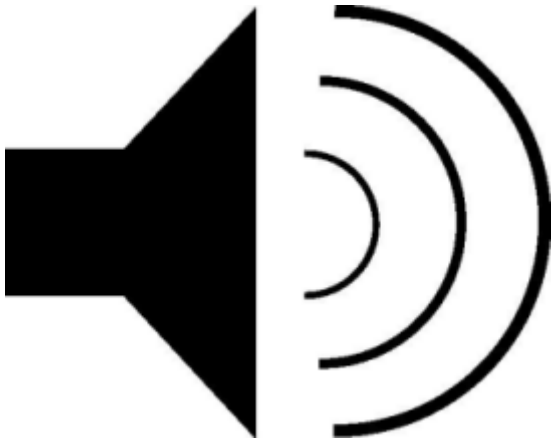
一些软件包（如 wit 和 apiai ）提供了一些超出基本语音识别的内置功能，如识别讲话者意图的自然语言处理功能。其他软件包，如谷歌云语音，则专注于语音向文本的转换，其中SpeechRecognition 就因便于使用脱颖而出。

pyaudio库可以进行录音，播放，生成wav文件等等。PyAudio 提供了 PortAudio 的 Python 语言版本，这是一个跨平台的音频 I/O 库，使用 PyAudio 你可以在 Python 程序中播放和录制音频。为PoTaTudio提供Python绑定，跨平台音频I/O库。使用PyAudio，您可以轻松地使用Python在各种平台上播放和录制音频，例如GNU/Linux、微软Windows和苹果Mac OS X/MACOS。

6.语音识别中的硬件

传声器：通常称为麦克风，是一种将声音转换成电子信号的换能器，即把声信号转成电信号，其核心参数是灵敏度、指向性、频率响应、阻抗、动态范围、信噪比、最大声压级（或AOP，声学过载点）、一致性等。传声器是语音识别的核心器件，决定了语音数据的基本质量。

扬声器：通常称为喇叭，是一种把电信号转变为声信号的换能器件，扬声器的性能优劣对音质的影响很大，其核心指标是TS参数。语音识别中由于涉及到回声抵消，对扬声器的总谐波失真要求稍高。



激光拾声：这是主动拾声的一种方式，可以通过激光的反射等方法拾取远处的振动信息，从而还原成为声音，这种方法以前主要应用在窃听领域，但是目前来看这种方法应用到语音识别还比较困难。

微波拾声：微波是指波长介于红外线和无线电波之间的电磁波，频率范围大约在 300MHz至300GHz之间，同激光拾声的原理类似，只是微波对于玻璃、塑料和瓷器几乎是穿越而不被吸收。

高速摄像头拾声：利用高速摄像机来拾取振动从而还原声音，这种方式需要可视范围和高速摄像机，只在一些特定场景里面应用。

7.语音信号文件WAV格式

wav格式，是微软开发的一种文件格式规范，整个文件分为两部分。第一部分是“总文件头”，就包括两个信息，chunkID，其值为“RIFF”，占四个字节；ChunkSize，其值是整个wav文件除去chunkID和ChunkSize，后面所有文件大小的字节数，占四个字节。第二部分是Format，其值为“wave”，占四个字节。它包括两个子chunk，分别是“fmt ”和“data”。在fmt子chunk中定义了该文件格式的参数信息，对于音频而言，包括：采样率、通道数、位宽、编码等等；data部分是“数据块”，即一帧一帧的二进制数据，对于音频而言，就是原始的PCM数据。从语音识别的原理可以知道，我们语音数据文件存储为WAV格式是最好的。

8.语音识别技术应用场景

语音识别已经深入应用到众多垂直领域中，概括起来，智能语音识别主要应用于以下3个方面，这也是语音识别商业化发展的主要方向。

- 语音输入系统：将语音识别成文字，摆脱生僻字和拼音障碍，让用户更加便捷，如微信中语音转文字、讯飞输入法、直播等视频实时字幕等。
- 语音控制系统：通过语音控制设备进行相关操作，彻底解放双手，如智能音箱、智能汽车系统、智能家居、智能穿戴。
- 语音对话系统：这是结合了语音识别与自然语言处理的技术，根据用户的语音实现交流与对话，保证回答的内容正确，对语义理解要求较高。目前，语音对话系统已经广泛应用于各类服务场景，如电商平台的智能客服、智能服务机器人等。相比于语音输入系统和语音控制系统，语音对话系统更加复杂，代表着语音识别的未来方向。