阿里云

智能语音技术的定义

通过对语音进行分析、理解和合成，使计算机设备实现“能听会说”、具备自然语言交流的技术能力。其涉及的范围主要有：

- 语音合成技术
- 语音识别技术
- 语音测评技术
- 语音降噪与增强技术
- ……

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

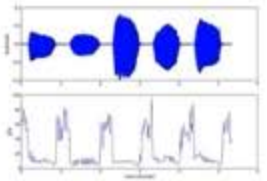
智能语音技术的研究任务

智能语音技术的研究难点

- 一门跨领域的技术——智能语音技术涉及到很多领域，需要掌握各领域的基础知识、掌握很多技能才能实用化。

语言学

心理学



信号处理

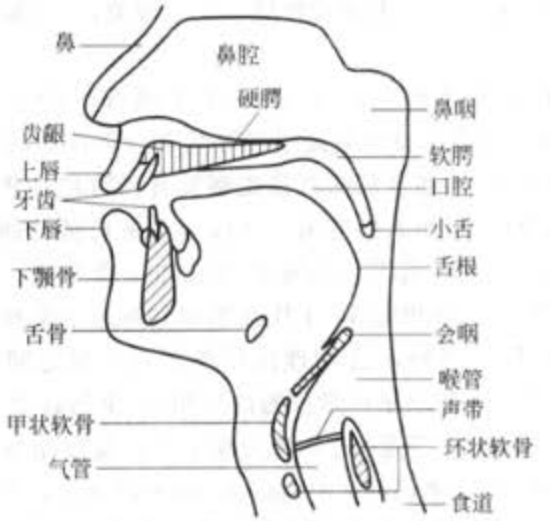
深度学习 … …

语音的产生

- 人的发声：肺部呼出的气流透过支气管到达喉头，引发喉头中声带的颤动，振动产生声音，再由口腔或鼻腔控制发声位置。

声音（Sound）是由物体振动产生的**声波**。是通过介质（**空气或固体、液体**）传播并能被人或动物听觉器官所感知的波动现象。

可以被人耳识别的**声波**（频率在20Hz~20000 Hz之间）。



人的发声器官

语音的物理载体及其特征属性

语音的物理载体是一种声波，声波的特征属性包括：

语音与语言

阿里云

- 语音的内涵：
  - 人类语言的物质表达；
  - 语言的外部形式；
  - 最直接地记录人的思维活动的符号体系；
  - 人的发音器官发出的具有一定社会意义的声音。

- 语音是声音和语言的组合体。
- 语音是一段语音序列携带语言信息的声音。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

音节的介绍

- 能够自然发出和觉察到的最小语音单位。
- 一个音节由一个或几个元音和辅音按照一定的规则组织起来。

英语发音

英语单词发音时，根据字母拼凑进行发音，这几个拼凑起来的字母叫一个音节，如：

- 单词Red，划分成Re-d两个音节；
- 单词speech分成s-pee-ch三个音节；

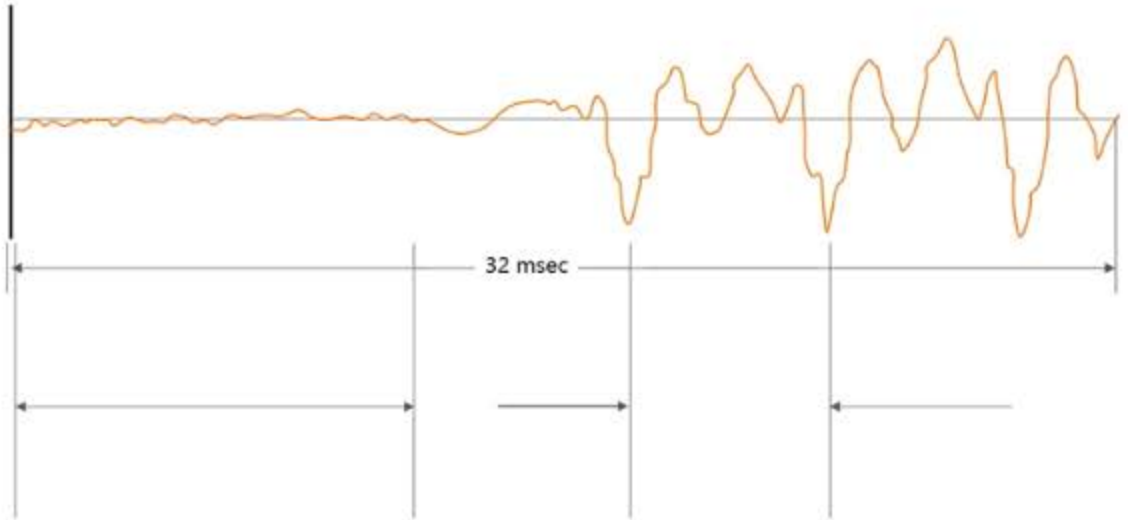
汉语发音

汉语发音则是一个字一个音节。发音示例如下：

- 语音分为“语-音”两个音节；
- 天猫精灵分为“天-猫-精-灵”四个音节。

语音信号的介绍

- 语音是人类交流的发声形式；
- 语音的基本模拟形式为语音信号的声波波形。
- 语音信号在产生过程中与环境和发声器官的联系很紧密，与各种运动都是相关的，语音信号本身是不平稳的信号。



语音信号的特点

- 通过麦克风转换成电子波形；
- 通过模拟/数字信号处理操作；
- 由扬声器或耳机转换回声学形式。

语音信号处理

- 将一种语音信号表示形式转换为另一种语音信号，以揭示语音信号的各种数学或实际性质，并进行适当的处理，以帮助解决基本问题和深层问题。

- 语音信号处理的目的:
  - 理解语音是一种交流的手段；
  - 语音的传播和复制；
  - 对语音进行分析，以便自动识别和提取信息；
  - 发现说话者的一些生理特征。

音频文件的参数介绍

阿里云

声道

录制声音时，在不同空间位置采集的相互独立的音频信号。声道数也就是声音录制时的音源数量。常见的音频数据为单声道或双声道（立体声）。

比特率

数据传输时单位时间传送的数据位数，也就是每秒的传输速率。比特率越高，传送数据速度越快。

音频采样率

音频采样率是指录音设备在一秒钟内对声音信号的采样次数，采样频率越高声音的还原就越真实自然。

音频采样位数

采样值或取样值，即是将采样样本幅度量化。用来衡量声音波动变化的参数，或是声卡的分辨率。数值越大、分辨率越高，发出声音的能力越强。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

音频编码

- 语音数据存储和传输的方式。
- 在调用智能语音交互服务之前需确认语音数据编码格式是服务所支持的。

01 PCM

PCM（脉冲编码调制）是Pulse Code Modulation的缩写。特点是音质好，但体积大。

02 WAV

WAV是常见的声音文件格式之一，是微软公司专门为Windows开发的一种标准数字音频文件。

03 MP3

MP3是常见的声音文件格式之一，是有损压缩的格式，适用于要求满足高比特率及兼容性的场景。

04 AAC

AAC是新一代的音频有损压缩技术，是一种专为声音数据设计的文件压缩格式。

05 OGG

类似于MP3等的音频压缩格式，但OGG是完全免费、开放和没有专利限制的。

06 FLAC

无损音频压缩编码，FLAC的特点是无损压缩。不会破坏任何原有的音频资讯。

语音降噪与增强技术的定义与作用

阿里云

定义

- 尽可能地从带噪声的语音信号中**提取有用语音信号**，**抑制或降低噪声干扰**的技术。

作用

- 降低背景噪声干扰，改善语音质量，提升听者的舒适感；
- 提高语音信息传达的可懂度。

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

语音降噪与增强技术的研究思路——传统信号处理方法

阿里云

实现原理：

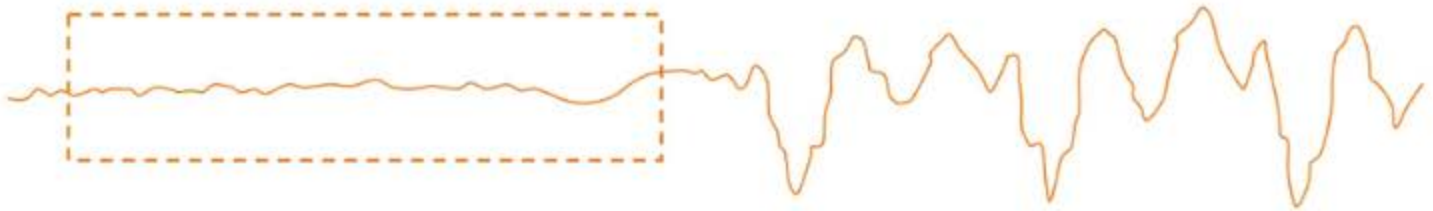
基于物理和数学原理推导，适用性强，所以系统一般有比较好的鲁棒性。

使用环境：

传统信号处理方法一般具有小计算量、低延迟等优势，容易满足实时性要求。

1.基于单通道的语音降噪与增强方法

2.基于麦克风阵列的语音降噪与增强方法



阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

阿里云

语音降噪与增强技术的研究思路——深度学习方法

实现原理：

利用**大量**的语音数据或噪声数据，训练网络学习相关的特征从而实现降噪，性能变化范围较大，系统在新环境下鲁棒性较差。

使用环境：

模型及计算资源等问题一方面会限制其在计算资源有限的系统中的使用，另一方面难以保证实时通信需求。

语音识别技术的定义

- 语音识别技术就是**“机器的听觉系统”**；
- ASR，Automatic Speech Recognition；
- 让机器通过识别和理解，把语音信号转变为相应的文本或命令。

语音识别技术的实现流程

- 语音识别技术的本质是一种**基于语音特征参数的识别**；
- 通过学习，系统能够把输入的语音按一定模式进行分类，进而依据判定准则找出最佳匹配结果。

语音唤醒技术的定义

- 语音识别任务的一个分支；
- 又称**关键词检测**（KWS，Key Word Spotting）；
- 在一串语音流中，检测出预先定义的**激活词**或**关键词**，而不需要对所有的语音进行识别。

语音唤醒模型的实现流程

语音唤醒能力主要依赖于语音唤醒模型的支持，这也是语音唤醒技术的核心。

“天猫精灵”

定义唤醒词需要依据：

- 易唤醒
- 低误唤醒
- 品牌性
- 易记易读性

方言版唤醒词

收集唤醒词发音时，一定要注意以下两点：

- 发音的清晰程度
- 收集相近的音节

深度学习

准备完数据后，需要构建模型并训练语音数据。

唤醒效果

上线后可收集用户的唤醒词，接着需要标注和重复训练并迭代。

语音合成技术的定义

- 一种通过机械的、电子的方法产生人造语音的技术；
- 又称**文语转换**（TTS，Text To Speech）；
- 可将任意输入文本转换成相应语音；
- 可将基本语音信息数字化，并利用计算机系统仿真出人类的声音。

文本输入　　语音合成技术　　语音输出

语音合成技术的原理——传统语音合成

阿里云

传统语音合成

语言分析部分

语言分析部分主要是根据输入的文字信息进行分析，生成对应的语言学规格书。主要有以下阶段：输入文本、句子结构分析、文本正则、文本转音素及韵律预测。

声学系统部分

声学系统部分需要根据语言分析部分的分析结果，通过一定的方法生成语音波形，但目前仍需要人工介入制定很多挑选规则和参数。

语音合成技术的原理——端到端语音合成

阿里云

端到端语音合成

- 直接输入文本或者注音字符，系统直接输出音频波形。
- 降低了对语言学知识的要求，可以很方便在不同语种上复制，批量实现更多语种的合成系统。
- 并且端到端语音合成系统表现出强大丰富的发音风格和韵律表现力。
- 但灵活性降低、效果不稳定。

人机交互方式的趋势

阿里云

第一台 通用电子计算机	Windows	智能手机 手绘板	智能音箱 语音输入法	手势识别
				
• 手工操作 • 依赖机器	• 图形用户界面 • 鼠标键盘输入	• 触摸式交互 • 手写输入	• 语音交互 • 语音输入	• 体感交互 • 智能用户界面

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能语音交互的定义

- 基于语音输入的新一代交互模式；
- 人类通过语音交流与机器进行信息传递的活动；
- 基于语音识别、语音合成、自然语言理解等技术。

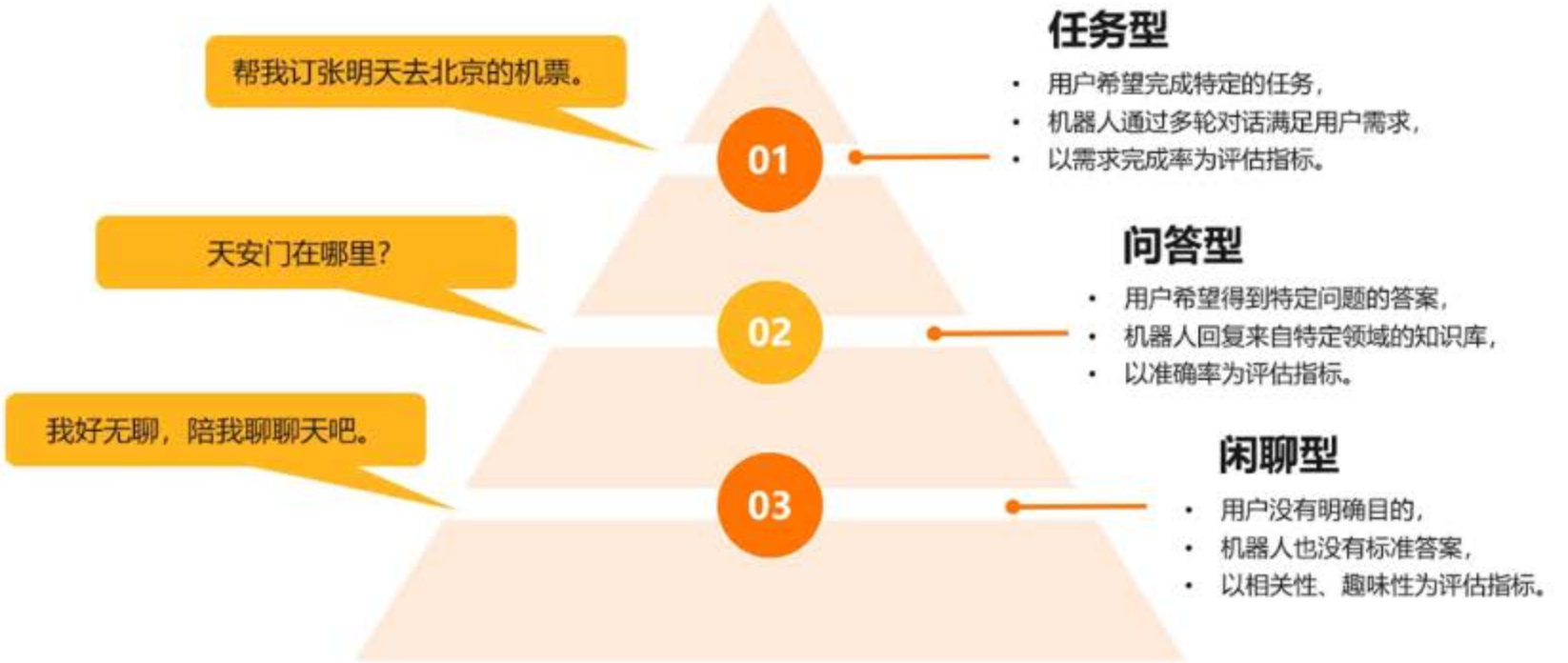

阿里云
智能语音交互的优劣势
信息传递效率
相比于键盘输入，语音输入在4类场景中，速度及准确率方面更具优势。
使用门槛低
语音交互学习成本低，且可为障碍群体带来极大的便利。
传递声学信息
依靠声音可以判断性别、年龄层、情绪等信息。
优势
劣势
信息接收效率
语音交互无法像文字一样可以跳读，同时会增加用户的记忆负担。
嘈杂环境的影响
嘈杂环境使得人声的提取变得非常困难。
公开环境的影响
公开环境下进行语音交互具有心理负担。
阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能对话系统的定义

- 人与机器可以通过自然语言进行对话交互的系统；
- 用准确、简洁的自然语言，回答用户用自然语言提出的问题；
- 注重与人的交互、对人意图的理解、对对话氛围的感知，以及回答的多样性和个性化。

智能对话系统的分类

根据用途分为任务型、问答型、闲聊型对话系统，这3类对话机器人的涵盖范围从低到高，相应的精度要求则是从高到低。



智能对话系统的发展趋势

阿里云

快速适应

- 有能力从机器与人的交互中主动学习；
- 快速适应用户的需求。
- **“根据用户性格自主训练，提供个性化方案”**

深度理解

- 目前模型产生的回复仍然缺乏多样性；
- 能够更加有效地深度理解语言和真实世界。
- **“理解‘意思意思’的意思”**

保护隐私

- 对话助手可能存储了一些较为敏感的信息；
- 因此加强对用户隐私的保护是非常重要的。
- **加密“地址、联系方式”等隐私信息**

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能对话系统的要素

阿里云

01

用户： 指产品或服务的使用者。

“喜欢看电影的年轻人”

02

对话代理人： 既可以是真人，如客服人员；也可以是虚拟人，如机器人。

“天猫精灵”

03

对话轮次： 一来一回称一轮，来回多次称为多轮对话。

“最近有啥热门电影”
“为您推荐以下电影”

04

会话： 由用户发起的某次多轮对话。会话是对话代理人与用户之间发生的一次连续对话。

“最近有啥热门电影”
“为您推荐以下电影”
“我想看这个”……

05

意图： 意图是系统能够识别的最小的用户目的，是系统决策的基本元素之一。

“看电影”

06

槽位： 存储会话过程中提取到的所有实体/槽值信息，用于后续对话系统的决策。

“电影种类”

07

实体/槽值： 特殊领域相关的实体需要单独定义，通用实体则可以由平台统一支持。

“喜剧” “情感剧”

08

动作： 理解用户的意图后，对话代理人除了回复消息外，可能需要做其他动作。

“订电影票”

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能对话系统的实现流程

阿里云

电影查询助手

"年轻人"
"从业人员"
目的性强

"我想看电影"
"好的，为您推荐
下列电影"

不同人的
对话风格

对话数据
再训练

调研 01	定位 02	设计 03	测试 04	完善 05
• 需求分析 • 场景选择	• 创建人物画像 • 选择对话风格	• 设计对话流程 • 对话体验脚本 • 情感化设计	• 模拟测试 • 场景测试	• 用户反馈 • 数据分析 • 迭代更新

阿里云培训中心
ALIBABA CLOUD TRAINING CENTER

智能对话系统的技术应用