慧科
只为赋能人才

机器人语音

——张洪涛

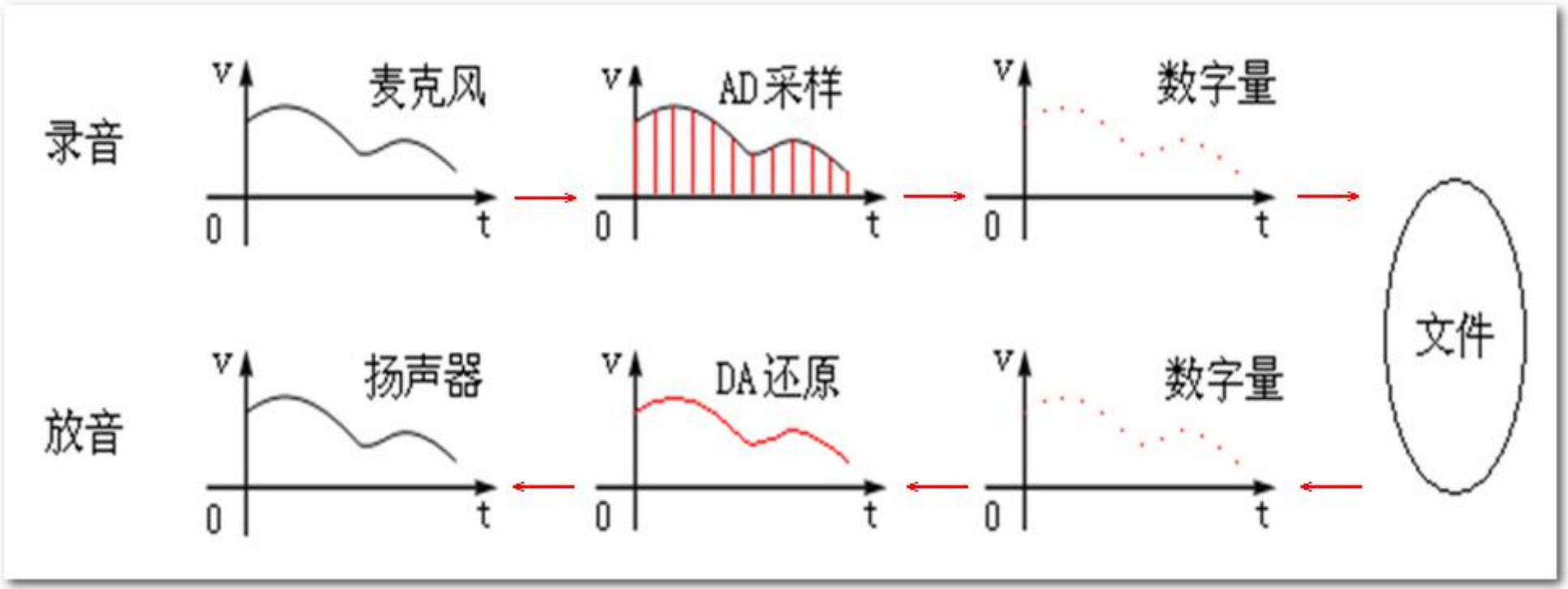
音频概述

➢音频原理

- 由自然音源通过采集电路（麦克风）得到的音频信号是一种连续变化的模拟信号。
- 但计算机只能处理和记载数字信号。音频信号必须变换为数字信号之后，才能由计算机进一步处理。

慧科
只为赋能人才

音频录音放音过程



慧科
只为赋能人才

录音及放音

- 录音时，先设置好采样频率、采样位数、声道数，再启动声卡的AD芯片，将音频的模拟信号转化为数字信号，最后将音频对应的数字信号存储到文件中。
- 放音时需要依据录音时声卡的采样频率、采样位数、声道数去设置声卡，再将音频的数字信号通过声卡的DA芯片转化为音频对应的模拟信号。

音频三要素

- 音频三要素

- 采样频率、采样位数、声道数

1、采样频率

- 每秒钟读取声音幅度样本的次数。
- 正常人听觉的频率范围大约在20Hz～20kHz之间，根据奈奎斯特采样理论，为了保证声音不失真，采样频率应该在40kHz左右。
- 常见的采样频率有11025Hz（11kHz）、22050Hz（22kHz）和44100Hz（44kHz）等。

慧科
只为赋能人才

音频三要素

- 2、采样位数
- 每个采样点其幅度数据的二进制位的个数。采样位数越高，数字化后的音频信号就越可能接近原始信号。

3、声道数

- 录音或放音时硬件线路的数量，即录音时的音源数量或放音时的扬声器数量。通常语音只用一个声道。而对于音乐来说，既可以是单声道（mono），也可以是双声道（即左声道右声道，叫立体声 stereo），还可以是多声道，叫环绕立体声， 多用于影院中。

慧科
只为赋能人才

有损压缩

- 前面提到音频的码率和实际的音质有关，高音质的音频文件一般都很大。但是由于人的听觉范围是 20-20000Hz，有的声音人是听不到的，所以在音频压缩的时候可以考虑去掉人听不到的那些声音以达到减少音频文件大小的目的。这就是有损压缩。
- 常见的有损音频压缩格式有：MP3，AAC，WMA，Ogg Vorbis

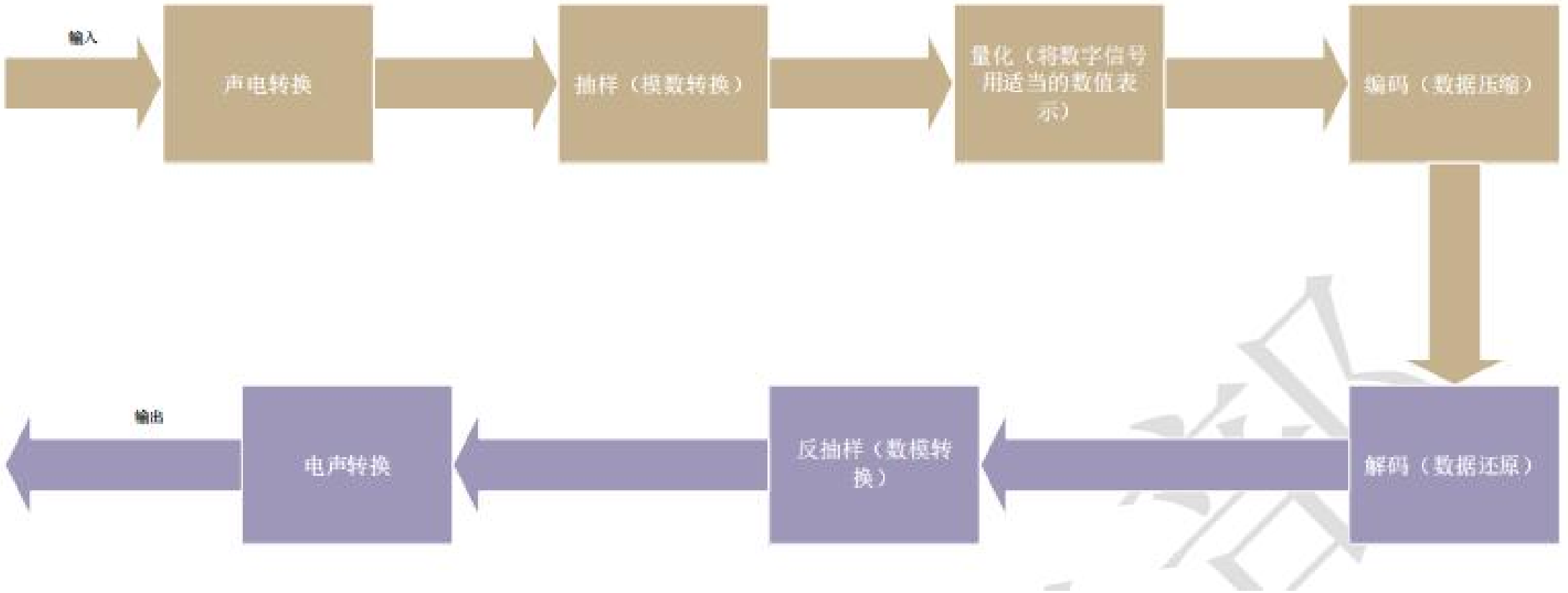
慧科
只为赋能人才

无损压缩

- 无损压缩主要是利用信息冗余进行数据压缩。无损压缩是一个可逆的过程。无损压缩比较适合在需要反复存档、读取的工作上。
- 无损的音频格式解压时不会产生数据或质量上的损失，解压产生的数据与未压缩的数据完全相同。
- 如需要保证音乐的原始质量，应当选择无损音频编解码器。常见的无损压缩格式有：
- WAV，PCM，ALS，ALAC，TAK，FLAC，APE，WavPack(WV)

慧科
只为赋能人才

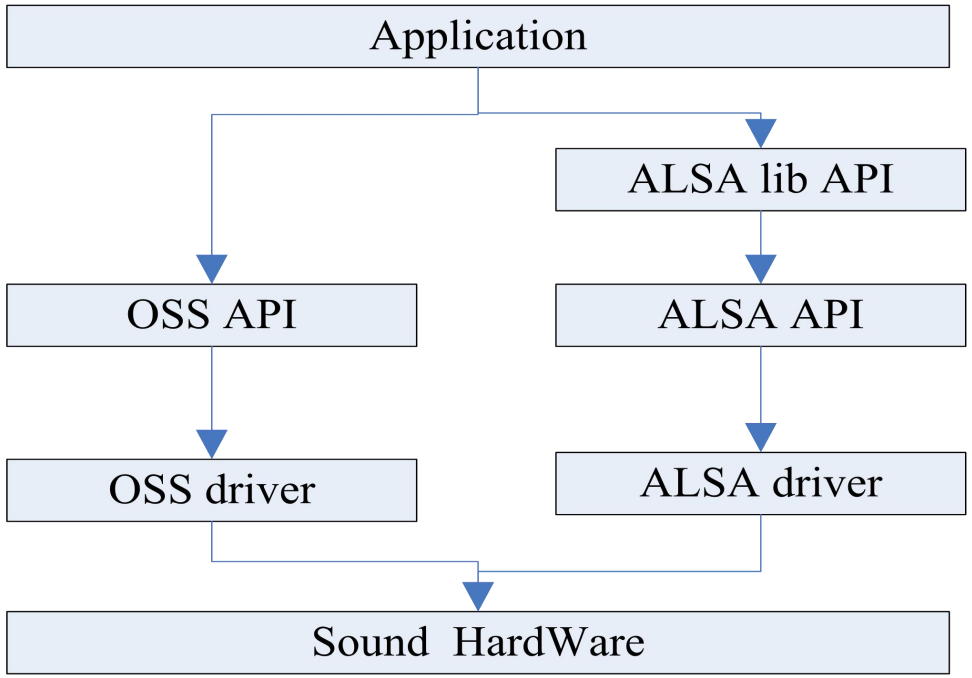
音频编解码基本原理



慧科
只为赋能人才

底层音频开发库

- OSS可移植性好，但具有声卡独占问题，闭源（商业版本）。
- ALSA具有声卡共享的特性，兼容OSS接口，开源。
- 2.6版本的Linux内核用ALSA代替了OSS。但为了兼容以前的程序，ALSA提供了OSS模拟。



慧科
只为赋能人才

python库pyaudio

- https://people.csail.mit.edu/hubert/pyaudio/docs/

慧科
只为赋能人才

练习

- 根据资料或示例程序，完成录音
- 播放声音
- 播放器：Aplay、Mpg123、sox、mplayer等
- 使用sox：
- 安装：sudo apt-get install sox
- play xxx.wav

慧科
只为赋能人才

机器人语音识别

- 现在 AI 技术发展得非常快，AI 技术在日常生活中应用也越来越广。其中语音技术在很早的时候就尝试走入大家的生活。从亚马逊的 Echo 到微软的 Cortana，从苹果的语音助手 Siri 到谷歌的Assistant 等等，语音识别技术的广泛应用让我们的生活便利了很多，但是在使用的过程中，是否会对这些智能化的语音回答感到兴趣满满呢？在这个课程中，我们将介绍一个经典的开源语音识别程序（CMU 的 pocketsphinx）来学习语音识别的基本知识。通过这个案例我们将会学会和理解语音识别的基本方法等。

慧科
只为赋能人才

语音识别

- 语音识别在这里我们指把声音变成文本。声音实际上是一种波，而我们常见的 MP3 格式的是压缩过的格式。在语音识别的过程中，我们通常使用非压缩的格式来处理。比如 PCM 文件。 比如下图就是一个 PCM 文件的声音波形。

慧科
只为赋能人才

语音识别

- 在语音识别开始时，需要把首尾端的静音切除，降低对后续步骤造成的干扰。这个静音切除的操作一般称为 VAD。在对声音分析的时候，需要对声音分帧，这里可以通俗地理解为把一个连续的声音切成一小段一小段，每一小段就是一帧。但是这里要理解的是上一帧与下一帧是有重叠的。
- 在分帧后，需要对每一帧建模。这个过程叫声学特征提取。

慧科
只为赋能人才

音素

- 音素，可以这么理解。单词的发音由音素构成，下面我们用到的例子中就是由 CMU定义的一套音素集。详细情况参见 The CMU Pronouncing Dictionary。
- 从语音识别开始分析声音开始到识别到单词经过了如下过程：

慧科
只为赋能人才

音素

- 在人的发声中，在一个音转向另一个音的时候是渐变的，所以后一个音的频谱与在其他情况下的频谱并不完全一样。为了解决这个问题，需要根据上下文来辨别音素。这里说的状态指的就是得到音素的中间状态。音素的第一部分是与它之前的音素存在关联，中间是稳定的部分，最后一部分是与下一个音素存在关联。这就是通常在使用 HMM 模型时的音素三状态。可以粗略地认为这是为了更准确地描述语音的一个手段。

慧科
只为赋能人才

非语言声音

- 另外在识别的时候，也有一些非语言学的声音，比如呼吸声，um,uh,咳嗽声等。在识别的时候可以用来做为停顿。这样的非语言学的声音叫做 utterances

慧科
只为赋能人才

Sphinx

- Sphinx 是由美国卡内基梅隆大学开发的大词汇量、非特定人、连续英语语音识别系统。
- 一个连续语音识别系统大致可分为四个部分：特征提取，声学模型训练，语言模型训练和解码器。
- 在语音识别中，会用到三个模型。声学模型（acoustic model），语音字典（phonetic dictonary）和语言模型（language model）。

慧科
只为赋能人才

声学模型

- 声学模型是根据训练语音库的特征参数训练出的结果。
- 在识别时可以将待识别的语音的特征参数和声学模型进行匹配，得到识别结果。
- 目前的主流语音识别系统多采用隐马尔可夫模型 HMM进行声学模型建模。

慧科
只为赋能人才

语音字典

- 语音字典包含了单词与音节的映射关系。
- 在识别的时候字典并不是唯一的用来表示单词与音节关系的手段，也可以便用更复杂的手段来表示它们之间的关系。

慧科
只为赋能人才

语言模型

- 语言模型是用来约束单词查找的。
- 它主要用于决定哪个词序列的可能性更大，或者在出现了几个词的情况下预测下一个即将出现的词语的内容。语音模型可以简单地理解为消除多音字的问题。
- 语言建模方法主要有基于规则模型和基于统计模型两种方法。统计语言模型是用概率统计的方法来揭示语言单位内在的统计规律，其中 N-Gram 模型简单有效，被广泛使用。它包含了单词序列的统计。

慧科
只为赋能人才

N-Gram

- N-Gram 模型基于这样一种假设，第 n 个词的出现只与前面 N-1 个词相关，而与其它任何词都不相关，整句的概率就是各个词出现概率的乘积。这些概率可以通过直接从语料中统计 N 个词同时出现的次数得到。
- 常用的是二元的 Bi-Gram 和三元的 Tri-Gram。
- Sphinx 中是采用二元语法和三元语法的统计语言概率模型，也就是通过前一个或两个单词来判定当前单词出现的概率 P(w2| w1)，P(w3| w2, w1)。

解码器

- 最后解码器是指在给定了根据语法、字典对马尔科夫模型进行连接后的搜索的网络（网络的每个节点可以是一个词组等）后，在所有可能的搜索路径中选择一条或多条最优（通常是最大后验概率）路径（字典中出现词组的词组串）作为识别结果。

慧科
只为赋能人才

安装Pocketsphinx

- SphinxBase 提供了一些 CMUSphinx 的一些基础的功能。Pocketsphinx 它依赖于 SphinxBase,
- 所以要安装 Pocketsphinx,就必须先安装 SphinxBase.
- 关于 Sphinx 的源码详细信息可以参看：
- https://cmusphinx.github.io/wiki/download/

慧科
只为赋能人才

安装Sphinxbase

- cd ~/
- wget http://sourceforge.net/projects/cmusphinx/files/sphinxbase/5prealpha/sphinxbase-5prealpha.tar.gz
- tar -zxvf ./sphinxbase-5prealpha.tar.gz
- cd ./sphinxbase-5prealpha
- ./configure --enable-fixed
- make clean all
- sudo make install

慧科
只为赋能人才

安装PocketSphinx

- cd ~/
- wget
- http://sourceforge.net/projects/cmusphinx/files/pocketsphinx/5prealpha/pocketsphinx-5prealpha.tar.gz
- tar -zxvf pocketsphinx-5prealpha.tar.gz
- cd ./pocketsphinx-5prealpha
- ./configure
- make clean all
- sudo make install

慧科
只为赋能人才

设置环境变量

- 在编译好 SphinxBase 和 PocketSphinx 后，请记得设置环境变量。 不然有可能得到如下错误
- "error while loading shared libraries: libpocketsphinx.so.3"。

- cd ~/
- export LD_LIBRARY_PATH=/usr/local/lib
- export PKG_CONFIG_PATH=/usr/local/lib/pkgconfig
- 也可以通过修改/etc/ld.so.conf 然后执行 ldconfig, 让设置永久有效。

慧科
只为赋能人才

创建语言模型

- 创建一个文本文件（corpus.txt）包含包下信息：
- resources
- application
- configuration
- alternatively
- language
- information
- depending
- paramters

慧科
只为赋能人才

在线生成语言模型

- 上传文本文件（corpus.txt）到：http://www.speech.cs.cmu.edu/tools/lmtool-new.html
- lmtool 是 CMUSphinx 提供的一个在线的生成语言模型的工具，它适合一些小的语料场景。
- 上传后，会得到如下结果。 下载 dic 和 lm 这两个文件。

慧科
只为赋能人才

运行

- cd ~/
- export LD_LIBRARY_PATH=/usr/local/lib
- export PKG_CONFIG_PATH=/usr/local/lib/pkgconfig
- pocketsphinx_continuous -hmm /usr/local/share/pocketsphinx/model/en-us/en-us -lm 3199.lm -dict 3199.dic -samprate 16000/8000/48000 -inmic yes

慧科
只为赋能人才

练习

- 自行创建文本文档，在线获取模型，用于识别
- 多次试验，检验结果
- 通过多次测试，试着分析部分不精确的识别结果产生的原因

慧科
只为赋能人才

实验任务

- 使用python控制机器人制作语音唤醒程序
- 唤醒成功则打印一句话

慧科
只为赋能人才

参考博客

- https://blog.csdn.net/zouxy09/article/details/7942784/
- https://www.cnblogs.com/woshijpf/articles/3633300.html
- http://blog.51cto.com/feature09/2300352

慧科
只为赋能人才

Yanshee SDK中的语音检测接口

名称	int ubtDetectVoiceMsg(char* buf, int timeout)
描述	检测是否听到
参数	【buf】 文字内容（utf-8编码）【timeout】检测超时时间，最小10s,最大10min;
返回值	-1获取状态超时、0检测ok
调用方式	同步

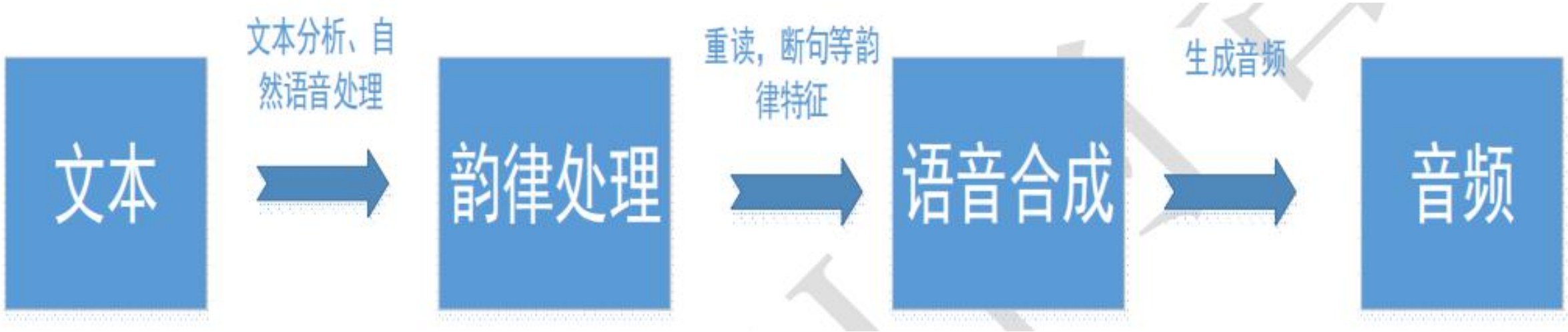
慧科
只为赋能人才

语音合成

- 语音合成就是通常我们说的 TTS(text to speech)，它涉及到了语言学、语音学、自然语言处理、信号处理、模式识别等，是一门典型的交叉学科。
- 一般来说，可以分为文本分析、韵律处理和语音合成三大模块。可以理解为通过一系列的数学方法建模，从原始的语音库中取出对应的最小单元，再利用合成技术对最小单元进行重音，语调等调整和修改，最后得到一个音频。

慧科
只为赋能人才

过程



慧科
只为赋能人才

文本分析

- 对输入文本进行语言学分析，逐句进行词汇的、语法的和语义的分析，以确定句子的低层结构和每个字的音素的组成，包括文本的断句、字词切分、多音字的处理、数字的处理、缩略语的处理等。

慧科
只为赋能人才

韵律处理

- 是指语音合成系统所输出的语音的质量，一般从
- 清晰度(或可懂度)、自然度和连贯性等方面进行主观评价。清晰度是正确听辨有意义词语的百分率；
- 自然度用来评价合成语音音质是否接近人说话的声音，合成词语的语调是否自然；
- 连贯性用来评价合成语句是否流畅。自然度取决于韵
- 律因素，像措辞和重读，声音的中断和持续，这决定音节的长度和语音的节奏。 这些特征可以指出句子强调的重点，在合适的地方断句等。

慧科
只为赋能人才

语音合成

- 把处理好的文本所对应的单字或短语从语音合成库中提取，把语言学描述转化成言语波形。 常用的语音合成技术有共振峰合成、LPC 合成、PSOLA 拼接合成和 LMA 声道模型技术合成等。这些语音合成技术各有特点，也各有不足。

慧科
只为赋能人才

共振峰合成

- 首先由于它是建立在对声道的模拟上，因此，对于声道模型的不精确势必会影响其合成质量。
- 实际工作表明，共振峰模型虽然描述了语音中最基本最主要的部分，但并不能表征影响语音自然度的其他许多细微的语音成分，从而影响了合成语音的自然度。
- 共振峰合成器控制十分复杂，对于一个好的合成器来说，其控制参数往往达到几十个，实现起来十分困难。

eSpeak

- 现在很多公司都开发了各自的语音合成程序，如科大讯飞、百度、腾讯、阿里等。但是这些商业的语音合成程序我们无法深入地用来学习。
- eSpeak 是一个开源的语音合成软件，它代码精简、支持多种语言并且合成快速。最重要的是他的代码是开放的，我们可以从 eSpeak 学习一个完整的语音合成。eSpeak 使用了上面介绍的共振峰的方法来合成语音，所以 eSpeak 合成的声音清晰、快速，但是并不真实平滑。详细信息可以参见：http://espeak.sourceforge.net/

慧科
只为赋能人才

eSpeak安装

- eSpeak 可以从源码安装，也可以简单地通过 Raspbian 的 APT 源安装。命令如下:
- sudo apt install espeak

- eSpeak 的源码可以从下面网站上得到：
- https://sourceforge.net/p/espeak/code/HEAD/tree/trunk/

慧科
只为赋能人才

eSpeak使用

- espeak -ven+f3 -k5 -s150 "Hello, I am Yanshee"

- python 调用
- cmd_start=" espeak -ven-us+m7 -a 200 -s180 -k20 --stdout '"
- cmd_end="' | aplay"
- call([cmd_start+text+cmd_end],shell=True)

慧科
只为赋能人才

练习

- 使用if语句实现和机器人固定的短暂对话

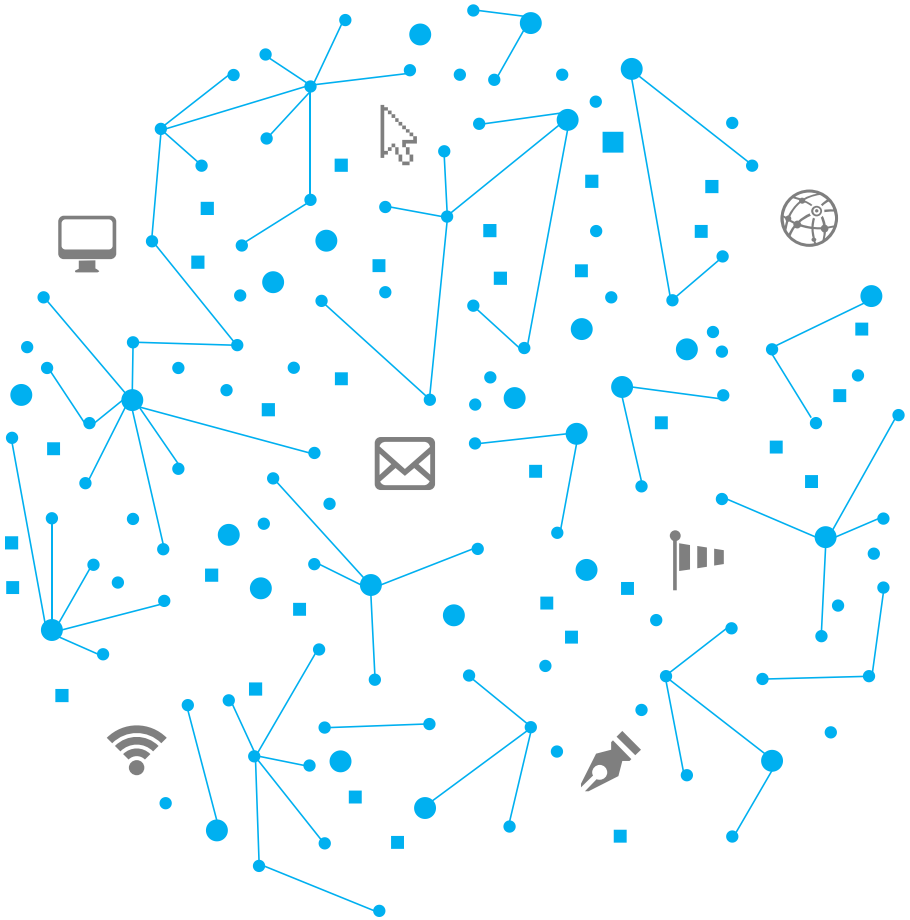
慧科
只为赋能人才

Yanshee SDK语音合成接口

名称	bool　ubtVoiceTTS (int isInterrputed, char *tts)
描述	将文字转成声音播放
参数	【isInterrputed】0 不打断前面的语句 1 打断前面的语句【tts】文字信息
返回值	-1-- 操作失败；0 -- 操作成功
调用方式	异步

慧科

只为赋能人才



THANK YOU

谢谢观看!