

第三章 语音识别

目标：

- 1. 掌握语音识别的基本概念；
- 2. 知道语音识别的典型应用；
- 3. 可以熟练掌握语音识别的相关应用；
- 4. 能够利用人工智能开源框架开发语音识别相关应用；
- 5. 建立起利用语音识别解决生活中所碰到的问题的意识；
- 6. 在日常学习生活中，建立起保护隐私信息（声纹等）的意识，养成主动保护自己及他人的个人隐私数据的习惯。

一、语音识别初体验

表 3.1.1 语音识别初体验备选工具

环境	工具名称	功能简介	获取方式
移动终端下（IOS 和 Android）	讯飞输入法	语音输入 1 分钟识别 400 字，识别率高达 98%，支持方言输入，支持英、日、韩、俄实时语音互译。	相应的市场输入讯飞输入法下载。或访问网址： https://srf.xunfei.cn/
Windows、Mac 和移动终端（IOS 和 Android）	剪映	视频编辑软件，支持语音识别，自动生成字幕。	APP 市场下载，或者访问网址下载： https://lv.ulikecam.com/

语音识别体验案例：自动语音识别应用，如语音输入和字幕自动生成。

1.利用讯飞输入法快速语音输入

应用案例 3-1：利用讯飞输入法实现语音输入

（1）在移动终端上，如苹果手机、ipad、Android 手机、平板电脑，访问相应的 APP 市场，并下载讯飞输入法；也可以下载安装电脑版，通过麦克风来体验语音输入功能。下面以 Windows 版为例演示说明。

（2）安装完成后，输入法切换到讯飞输入法。



图 3.1.1 讯飞输入法面板

(3) 在文档中，或者聊天软件的输入框中，点击图 3.1.1 的话筒标志，弹出如图 3.1.2 所示语音输入面板。

(4) 点击中间话筒标志说话，即可进行语音输入。

(5) 尝试使用方言语音输入，体验并填写表 3.1.2。



图 3.1.2 语音输入界面

2.利用剪映自动创建视频字幕

应用案例 3-2：利用剪映电脑版实现自动生成视频字幕功能

(1) 创建字幕是视频编辑者的一个痛点，需要大量的重复劳动，而语音识别可以很好的解决这个痛点问题。按照表 3.1.1 第二行所示下载适合你的设备的剪映，下面以 Windows 版为例进行体验。

(2) 打开剪映软件，点击开始创作，导入素材，点击文本菜单，点击识别字幕按钮。如图 3.1.3 所示。

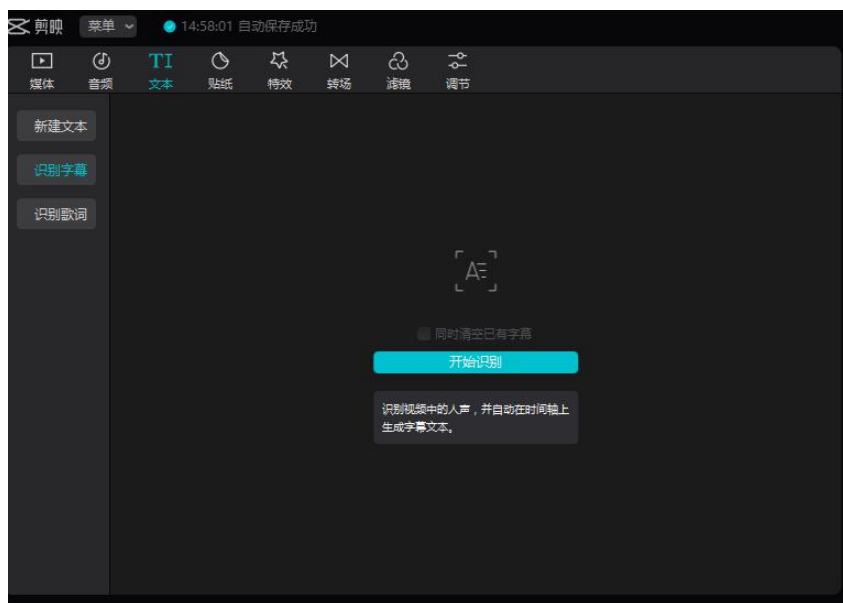


图 3.1.3 剪映识别字幕界面

(3) 点击开始识别，程序开始自动识别，完成后，在下方轨道上自动生成字

幕文件，视频中也自动生成了字幕。

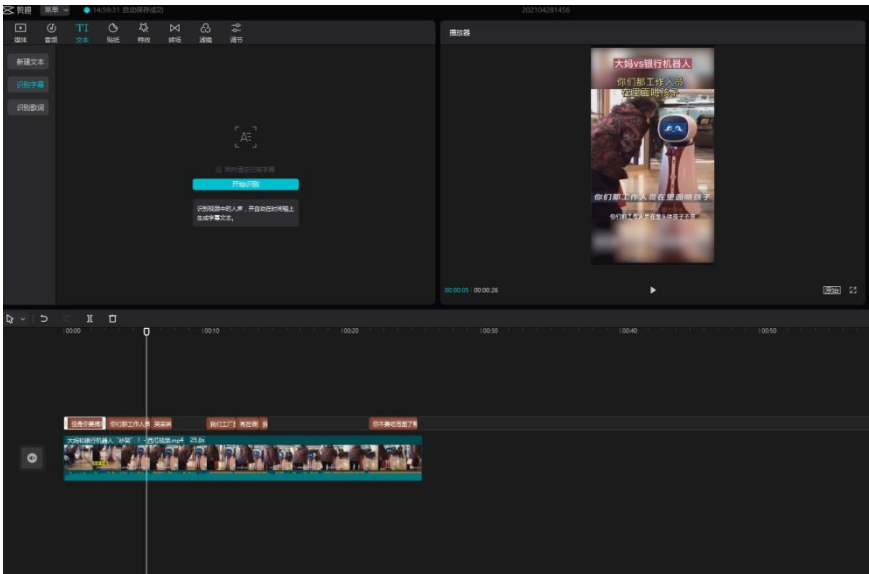


图 3.1.4 剪映自动生成字幕

(4) 体验完成后，填写 3.1.2 的记录表。

表 3.1.2 语音识别体验任务记录表

项目	记录
体验的工具名称	
体验的功能	
错误数（所识别语音中文字错误的个数）	
错误率（错误数/总数）	

二、语音识别的概念

1.语音识别的概念

声音是重要的信息载体，人类通过语言来传递信息，人们对声音已经有了深入的了解，并用响度、音调、音色、波形图、频谱图等来对声音进行量化描述，其中响度、音调、音色为声音的三要素。在语音识别之前，学者们研究了将声音从自然状态转为数字化的技术，也就是经过采样、量化、编码，实现对声音的数字化，但这仅仅是存在状态的改变，并没有改变声音的波形本质。

自动语音识别，简称语音识别，与声音的数字化不同，是将声音转换为相应

的文本或命令的技术，是实现将信息由声音波形转变为数字化字符的过程，同时语音识别还有语义理解层面的内涵，不仅仅识别出文字，还能够理解文字所承载的信息，也就是能够听得到、听得懂。传统的语音识别系统主要包含特征提取、声学模型、语言模型以及字典与解码四大部分。

2.语音识别的发展水平

语音识别目前已经实现了产品化应用，如智能音箱、翻译机、会议记录、同传设备等，识别的正确率已经和人类差别不大。2016 年 10 月，微软人工智能与研究部门发表了论文《Achieving Human Parity in Conversational Speech Recognition》中，提出语音识别系统已经实现了与专业转录员相当的错误率水平，当时的错误率仅有 5.9%。¹

在特定场景下，语音识别的效果确实已经能满足日常应用。但是在一些复杂场景下，语音识别还是无法和人类的能力相匹敌，如鸡尾酒会场景（嘈杂场景）、远场、方言等多语言混杂场景、人名识别、专有名词识别等。

3.语音识别的核心问题

语音识别的发展经历了高斯混合模型和隐马尔科夫模型的混合模型阶段，即 GMM-HMM 模型，到结合深度神经网络的模型阶段（深度神经网络（DNN）、循环神经网络（RNN）、长短时记忆网络（LSTM）），再到近期的端到端的模式阶段，识别的准确度已经越来越高，普遍认为语音识别已经达到 97%及以上的准确度，语音识别已经具备了可用性。

语音识别的核心问题为“是什么”、“是谁”和“生成语音”三个问题，即自动语音识别（ASR）、声纹识别（SR）、语音合成（TTS）。自动语音识别是将声音转为文字或指令，声纹是识别说话者的特征，语音合成是将文字转为声音。语音识别的问题指向人类的听得到、听得懂和说的出，最终目标是能够实现同人类一样的对话能力。

¹ Xiong W , Droppo J , Huang X , et al. Achieving Human Parity in Conversational Speech Recognition[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2016, PP(99).

4.语音识别的一般过程

自动语音识别系统主要包含特征提取、声学模型，语言模型以及字典与解码四大大部分，其中为了更有效地提取特征，往往还需要对所采集到的声音信号进行滤波、分帧等预处理工作，把要分析的信号从原始信号中提取出来；之后，在特征提取工作阶段，将声音信号从时域转换到频域，为声学模型提供合适的特征向量；声学模型中再根据声学特性计算每一个特征向量在声学特征上的得分；而语言模型则根据语言学相关的理论，计算该声音信号对应可能词组序列的概率；最后根据已有的字典，对词组序列进行解码，得到相应的文本并输出。

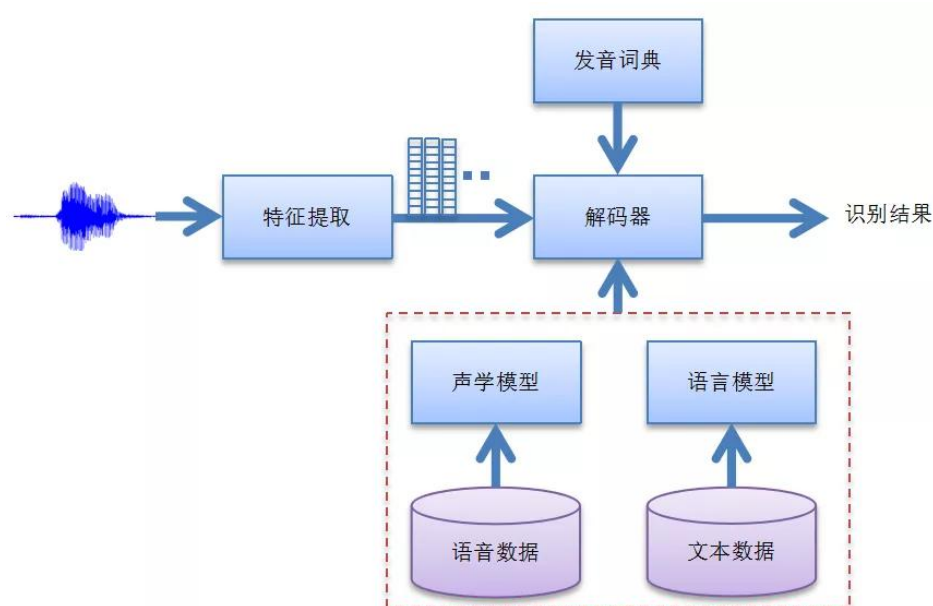


图 3.2.1 语音识别过程

三、语音识别的应用

语音识别在日常生活中有诸多的应用，最简单也是最常见的就是语音输入法；然后就是语音特征识别，即声纹识别，实现语音解锁功能，也可以复刻人的声音，让机器代替你配音；最后是语音合成应用，可以实现将文字转为声音，广泛用于智能客服领域；在教育领域，可以使用语音识别的技术模式实现智能语音测评，普通话水平测试，发音训练等场景。下面就以自动语音识别、语音合成、语音测评为例介绍语音识别的典型应用。

1. 自动语音识别（ASF）：微信语音输入&语音转文字

自动语音识别是语音识别的主体，是自动实现识别语音，并将语音转为文字或指令的过程，一般可用于语音输入法，会议记录，实时字幕，同声翻译等。

应用案例 3-3：微信的强大语音输入和语音转文字功能

表 3.3.1 自动语音识别体验工具

环境	工具名称	功能简介	获取方式
Windows、Mac 和移动终端 (IOS 和 Android)	微信	聊天、社交工具。	APP 市场下载，或者访问网址下载： https://weixin.qq.com/

场景描述：有时候我们双手都在忙，而又不得不快速回应一个朋友的聊天信息，朋友又不方便接听语音，这时候就需要用到了微信自带的语音输入了。同样，如果朋友发了一条语音消息给你，而你又不想让别人听到声音，就可以使用语音转文字功能了。

体验步骤：

- (1) 打开微信；
- (2) 找到朋友或者文件传输助手，点击下方的加号，找到语音输入，口述聊天内容。
- (3) 将语音识别的内容发送。
- (4) 同样的聊天窗口，使用按住说话功能，发送语音。
- (5) 在语音内容上，长按，选择转为文字，将所转的文字与语音内容进行对比，看是否有错误。
- (6) 完成体验后思考如下问题：
 - 该应用解决了什么问题？
 - 解决的效果如何？

● 你是否有改进建议？或者你是否有更好的解决方案？

【拓展阅读】

我们目前所使用的语音识别技术，给我们很惊艳的感觉，语音识别的准确度这么高，貌似很简单的一门技术，但实际上学者们在语音识别领域已经进行了半个多世纪的攻坚了，我们目前所享受的强大功能，都是历代学者努力研究突破的结果。

早在 1952 年的贝尔实验室的研究中，学者们首次实现了 Audrey 英文数字识别系统，可以识别单个数字 0~9 的发音，而且识别准确率已经能够达到 90% 以上。

1971 年美国国防部研究所资助的语音理解研究项目，使得语音识别有了一次大的发展。IBM、卡内基梅隆大学、斯坦福大学等企业机构加入到语音识别的研究中，当时最令人振奋的是卡耐基梅隆大学研发的 harpy 语音识别系统，能够识别 1011 个单词。

1980 年随着两个重量级算法的出炉：隐马尔科夫模型（HMM）、N-gram 语言模型，这两个算法的出现让语音识别能够从独立词组的识别发展到能够识别连续词。

1990 年，大词汇量连续词识别技术获得了持续的进步。当时提出了区分性的模型训练方法 MCE 和 MMI，使得语音识别的精确度日益提高，尤其适用于长句子的情况下，与此同时还提出了模型自适应方法 MAP 和 MLLR。


2006 年神经网络之父 Hinton 提出深度置信网络（DBN），2009 年 Hinton 和学生 Mohamed 将深度神经网络应用与语音识别，在小词汇量连续语音识别任务 TIMIT 上获得成功。

目前，最新的语音识别采用模式识别的基本框架，按照数据准备、特征提取、模型训练、测试应用等步骤，经过这几个步骤可以训练出进行语音识别的模型，在实际的识别任务中，将语音特征对照模型进行匹配，经过匹配、判决等过程，完成语音识别的过程。

语音识别还有一个非常好玩的应用就是听声识曲，大家可以下载酷狗音乐 APP（听歌识曲世界冠军），在 APP 主界面点击我的，点击右上角的菜单按钮，看

到听歌识曲按钮，进入后可以体验听歌识曲、连续识曲、哼唱识别三个功能。

表 3.3.2 语音合成体验工具


环境	工具名称	功能简介	获取方式
Windows、Mac 和移动终端（IOS 和 Android）	酷狗音乐播放器	拥有海量歌库的音乐播放器，拥有听歌识曲功能。	各大市场搜索酷狗下载； 网址： https://www.kugou.com/ 或扫描以下二维码： 

2. 语音合成（文字转语音，TTS）：文本转语音机器人

语音合成是一个神奇的存在，可以将文字转成声音信号并输出。可以实现阅读听书，智能配音，辅助语言障碍人士发音等功能。语音合成目前在语言的流畅度、拟人性等方面还有一段路要走，虽然语音合成较生硬，但已经能够实现语音交互和语音信息传递，具备了应用级的能力。

应用案例 3-4：文字转声音机器人

表 3.3.3 语音合成体验工具

环境	工具名称	功能简介	获取方式
Windows、Mac 和移动终端（IOS 和 Android）浏览器	Texttospeechrobot	文字转语音工具，支持语音下载，支持英语、德语、西班牙语、中文等文字转语音功能。	网址： https://texttospeechrobot.com/tts-player/ 或扫描以下二维码： 

场景描述：声音是很多人赖以生存的手段，但有时候喉咙会不舒服，甚至讲不了话，比如老师讲课讲的太多了，声音嘶哑，而此时又迫切需要准备视频微课并配音，这时候就可以使用语音合成功能了。

体验步骤:

- (1) 电脑浏览器中打开表 3.3.2 所示网址。
- (2) 点击 v4 (只有 v4 支持中文), 选择语言为中文 (两个女声, 一个男声), 如图 3.3.1 所示。
- (3) 输入需要转为声音的文字。
- (4) 点击转换 (CONVERT)。
- (5) 在播放器上右键点击音频另存为, 选择本地位置并保存, 注意下载完成后需要修改文件的扩展名为 mp3, 如图 3.3.2 所示。
- (6) 完成体验后思考如下问题:
 - 该应用解决了什么问题?
 - 解决的效果如何?
 - 你是否有改进建议?

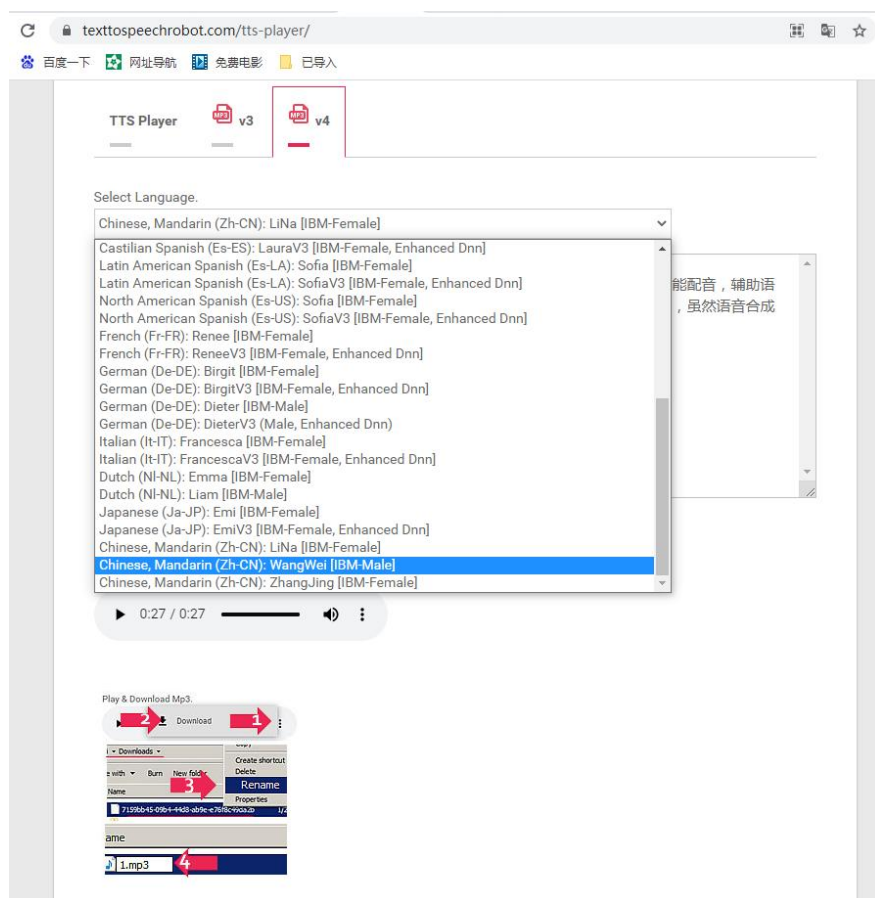


图 3.3.1 选择语言和发音者界面

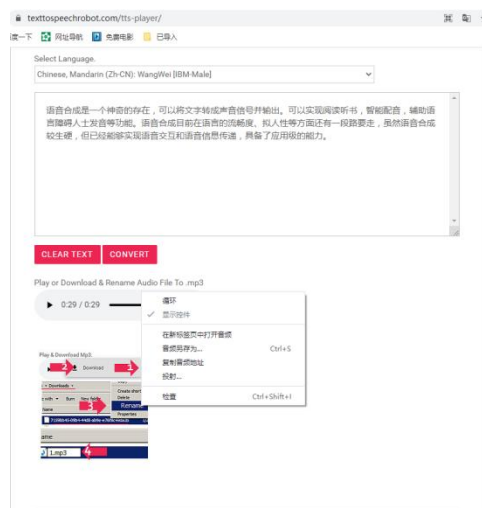


图 3.3.2 文本输入及音频另存为界面

【拓展阅读】

语音合成是语音识别的逆过程，就是按照目的实现自动生成语音的功能。除了通过输入文字实现文字转语音，目前，学者们还在研究如何通过脑电波、唇语等实现语音合成。脑电波语音合成可以帮助失语症患者获得发音的能力，唇语合成语音可以应用于警务监听等特殊场景，也可以帮助失聪者“听到”声音。语音合成可以弥补人类的缺陷。此外，语音合成在语音导航、机器人技术等领域有着深入的研究和应用。下面就简单介绍下读唇语音合成和脑电波语音合成。

嘴唇一动就能合成语音。2020 年 github 上开源了一个比较有趣的项目——lip2wav，一个嘴唇一动就能合成语音的项目，这个项目的效果比较惊艳，这个项目采用了五位演讲者作为训练数据，通过训练识别面部的唇部动作，项目基于 face_alignment 模型开发，利用 LSTM 识别唇部动作生成文字，然后合成语音。该项目目前仅仅限于在训练数据集上表现良好，在通用场景中表现差强人意，但是给唇语翻译场景提供了一种非常有效的思路。²

github 开源地址：<https://github.com/Rudrabha/Lip2Wav>

大家可能还记得著名的科学家霍金，虽然身患“渐冻症”，丧失了说话能力，依然可以借助一个外置的装置实现说话，那么现在的这个项目应该算是这个装置的升级版，仅仅通过脑电波的解码就能生成语音，实现语言交流的目的。该项目

² K R Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, C V Jawahar. Learning Individual Speaking Styles for Accurate Lip to Speech Synthesis[EB/OL]. <https://arxiv.org/pdf/2005.08209.pdf>, 2010.

在 2019 年在 Nature 杂志上发布，当时论文的标题为：“Speech synthesis from neural decoding of spoken sentences”，该项技术可以恢复人类口语交流能力。该项目需要一个外部装置采集头部、眼睛或者大脑皮层活动信息，然后将这些信息转换为文字，然后转为语音信息。核心技术采用了循环神经网络（RNN）。³


语音合成还有一个非常神奇的应用，就是能够复刻你的声音，如果感兴趣可以利用微信访问小程序——讯飞留声来体验。还有也可以访问讯飞 AI 体验栈小程序，来体验声音鉴定、拍照阅读等功能。

3. 语音测评：基于英语流利说的英语朗读打分

语音测评是自动语音识别的扩展和延伸，利用训练机器语音识别的模式，对人类的发音进行测评，从而帮助人类更好的进行发音学习，如普通话、外语等语言类学习时的测评打分与示范纠正功能。目前在师范生普通话水平考试，英语口语测评等场景下，已经开始了使用，人类专家只需要对机器测评结果进行抽样复核即可。

应用案例 3-5：英语口语测评

表 3.3.4 语音测评体验工具

环境	工具名称	功能简介	获取方式
移动终端（IOS 和 Android）浏览器，微信小程序	英语流利说	一款英语学习工具，交互式训练，包含语法、对话、阅读等多个维度的能力训练，采用语音评分技术，帮助学习者学习英语。	网址： https://www.liulishuo.com/ 或扫描以下二维码 体验微信小程序： 

场景描述：同学们，你们在学习英语的时候，是不是对自己的发音没有信心？需要有人教授你准确的发音？而你的父母要么工作繁忙，要么在英语发音方面也

³ Gopala K. Anumanchipalli, Josh Chartier, Edward F. Chang. Intelligible speech synthesis from neural decoding of spoken sentences[EB/OL]. <https://www.biorxiv.org/content/10.1101/481267v1.full>, 2018.

不擅长，这时候如果有一个智能工具来帮助你，肯定会让你的英语发音能够纯正标准的吧？下面我们来看一下英语流利说是否能够满足你的需求。

体验步骤（以 APP 为例，微信小程序可自行体验）：

- （1）APP 市场中搜索“英语流利说”，获取并安装。
- （2）注册并登陆。
- （3）点击“我的”，拖到底部找到“英语水平测试”。
- （4）点击“开始测试”，利用 5 分钟时间精准定位英语听说读写的全方位能力。
- （5）按照提示完成各个题目，完成朗读任务，听说任务。
- （6）出具测试报告。
- （7）完成体验并思考以下问题：
 - 该应用解决了什么问题？
 - 他比传统模式的改进点在哪里？
 - 你是否有更好的解决方案？



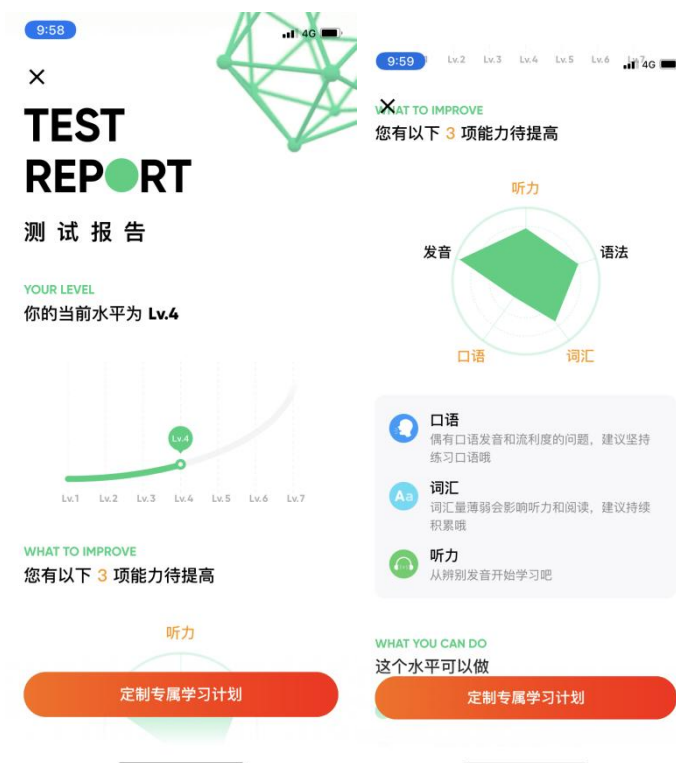


图 3.3.3 英语流利说英语水平测试功能

【拓展阅读】

通过语音测评的了解，我们发现语音识别涉及到了语音的基本知识，在语音识别过程中，需要识别出特征序列，将特征序列转为状态序列，然后识别出音素序列，将因素序列转变为次序，最终合并成句子，完成语音识别的过程。如图 3.3.4 所示。

人类的大脑从婴儿一出生开始就在不断地学习声音，经过长时间的浸润，最终才能够听得懂人类的语言，而专家们在训练机器的时候，需要学习语言的共性和发音规律，其中主要的概念为：音素、音节。

音素是构成语音的最小单位。英语有 48 个音素（20 个元音和 28 个辅音）。采用元音和辅音来分类，汉语普通话有 32 个音素，包括元音 10 个，辅音 22 个。

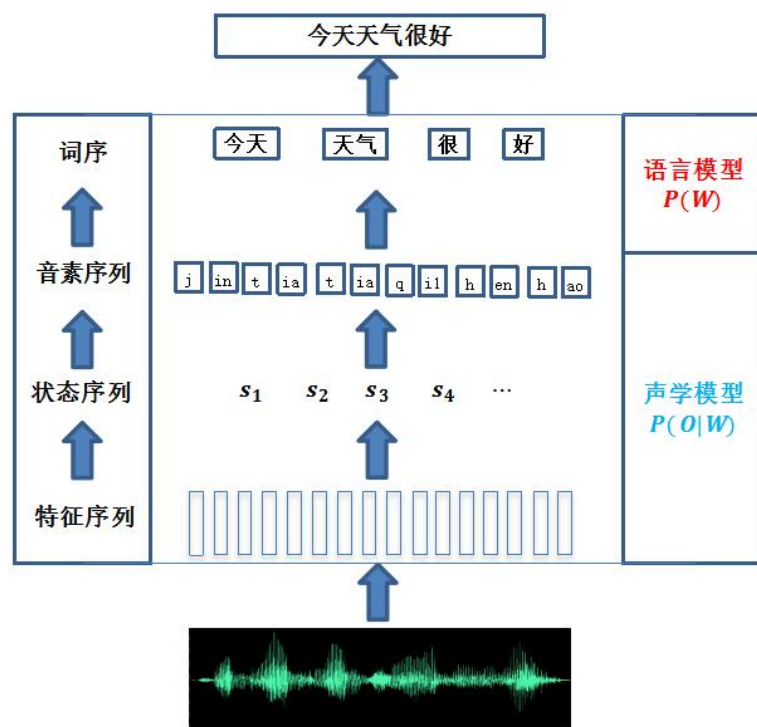


图 3.3.4 语音识别的过程

音节是听觉能感受到的最自然的语音单位，有一个或多个因素按照一定的规律组合而成。英语音节可单独由一个元音构成，也可由一个元音和一个或多个辅音构成。汉语的音节由声母、韵母和音调构成，其中音调信息包含在韵母中。汉语音节结构可以简化为声母+韵母。

四、语音识别的开发

到目前为止，我们已经知道语音识别是什么、怎么用了，那么你是否跃跃欲试想要开发属于自己的应用了？你又想到了哪些金点子呢？带着你的金点子，先来看一下图形化编程和 Python 编程的开发案例，给你提供一些开发参考。

1.图形化编程（采用慧编程工具，工具信息参加表 2.4.1）

前驱知识准备：图形化编程的界面布局；知道角色背景造型等基本概念；图形化编程的事件（当绿旗被点击）、外观（说**2 秒）、控制（等待 1 秒、重复执行、如果那么）等基本积木，连接运算符；知道图形化编程的基本方法。

硬件准备：带有摄像头和麦克风的笔记本电脑（含 Mac 笔记本）、台式机电脑。注：平板电脑暂不支持机器学习模块。慧编程 mBlock5. 3.0 支持 Windows7、Windows10、macOS 10.12+版本的操作系统，推荐 64 位操作系统。

软件工具准备：参照表 2.4.1 所示第一行的获取方式下载并安装慧编程工具。完成账号注册并登陆（可以应用更多功能）。

任务：搭建一个识别你性别并能够重复你说的话的对话机器人

- （1）识别性别，并让角色说出。
- （2）识别普通话，并重复你说的话。（此处使用了连接运算符）

参考代码如下：



图 3.4.1 简单对话机器人代码

拓展任务 1 (难度: 简单): 如何让角色一直识别你说的话, 并复述出来? (提示: 循环积木)

拓展任务 2 (难度: 难): 如何让角色用声音和你对话呢? (提示: 使用人工智能服务扩展)

2. Python 编程

前驱知识准备: 编程的基本模式如 IPO 模式, 知道输入、处理、输出的概念; Python 的基本语法: 运算符与表达式、常见数据类型、控制流 (顺序、分支、循环)、输入与输出、函数, 库安装和库引用, Python 的编程环境配置 (如 Anaconda 或者单独的 Jupyter notebook)、安装与使用。

硬件准备: 带有摄像头和麦克风的笔记本电脑 (含 Mac 笔记本)、台式机电脑。

软件准备: Python 3.8.5 及以上版本。直接下载 Python 安装包或者通过 Anaconda 安装时一起安装 Python。语音识别需要的依赖库包括: portaudio、pyaudio (录音、播放、生成 wav 文件)、SpeechRecognition

安装命令如下:

```
pip3 install portaudio
```

```
pip3 install pyaudio
```

```
pip3 install PocketSphinx
```

```
pip3 install SpeechRecognition
```

或者:

```
conda install portaudio
```

```
Conda install pyaudio
```

```
Conda install PocketSphinx
```

```
Conda install SpeechRecognition
```

然后安装最新的中文声学模型、语言模型和字典文件, 下载地址如下:

<https://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20>

0Language%20Models/Mandarin/cmusphinx-zh-cn-5.2.tar.gz/download

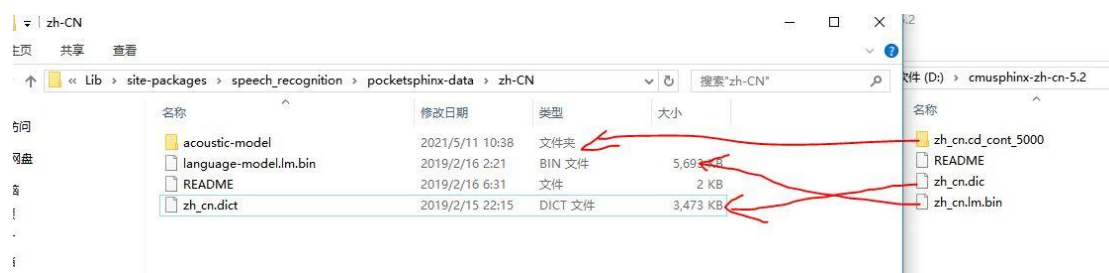


图 3.4.2 中文语言模型文件对应关系

先在 C:\Anaconda3\Lib\site-packages\speech_recognition\pocketsphinx-data\创建 zh_CN 的文件夹，将下载的模型按照图 3.4.2 的对应关系修改文件名。这样就配置好了中文语音识别的环境。下面就可以开始开发任务了。

任务：利用开源框架实现音频转写功能

参考代码如下：

```
import speech_recognition as sr

r=sr.Recognizer()

test=sr.AudioFile("D:/test1.wav")

with test as source:

    audio=r.record(source)

type(audio)

c=r.recognize_sphinx(audio, language=' zh-cn' )

print(c)
```

拓展任务 1（难度：简单）：如何实现实时语音转文字？

拓展任务 2（难度：难）：编写程序，设置语音唤醒指令为“老铁”，开始侦测语音指令，如果指令为“天气情况”播放当前天气，如果指令为“空气质量”，播放当前位置空气质量。（注：需要调用第三方接口读取天气和空气质量数据，可参考网上代码实现）

思考与练习

1. 什么是语音识别？语音识别的常见应用有哪些？
2. 应用计算机语音识别会不会导致个人隐私数据的泄露？如果会，该如何避免？
3. 请使用合适的工具或编程语言，搭建一个接收语音指令“当前时间”并文字播报当前时间的程序。有余力的同学可以实现语音播报当前时间。

