语音识别的知识体系

语音的知识体系可以划分为三个大的部分：专业基础、支撑技能和应用技能。语音识别的专业基础又包括了算法基础、数据知识和开源平台，其中算法基础是语音识别系统的核心知识，包括了声学机理、信号处理、声学模型、语言模型和解码搜索等。

01

专业基础

1.1 算法基础

声学机理：包括发音机理、听觉机理和语言机理，发音机理主要探讨人类发声器官和这些器官在发声过程中的作用，而听觉机理主要探讨人类听觉器官、听觉神经及其辨别处理声音的方式，语言机理主要探究人类语言的分布和组织方式。这些知识对于理论突破和模型生成具有重要意义。

信号处理：包括语音增强、噪声抑制、回声抵消、混响抑制、波束形成、声源定位、声源分离、声源追踪等。具体如下：

- 语音增强：这里是狭义定义，指自动增益或者阵列增益，主要是解决拾音距离的问题，自动增益一般会增加所有信号能量，而语音增强只增加有效语音信号的能量。
- 噪声抑制：语音识别不需要完全去除噪声，相对来说通话系统中则必须完全去除噪声。这里说的噪声一般指环境噪声，比如空调噪声，这类噪声通常不具有空间指向性，能量也不是特别大，不会掩盖正常的语音，只是影响了语音的清晰度和可懂度。这种方法不适合强噪声环境下的处理，但是足以应付日常场景的语音交互。
- 混响消除：混响消除的效果很大程度影响了语音识别的效果。一般来说，当声源停止发声后，声波在房间内要经过多次反射和吸收，似乎若干个声波混合持续一段时间，这种现象叫做混响。混响会严重影响语音信号处理，并且降低测向精度。
- 回声抵消：严格来说，这里不应该叫回声，应该叫“自噪声”。回声是混响的延伸概念，这两者的区别就是回声的时延更长。一般来说，超过 100 毫秒时延的混响，人类能够明显区分出，似乎一个声音同时出现了两次，就叫做回声。实际上，这里所指的是语音交互设备自己发出的声音，比如 Echo 音箱，当播放歌曲的时候若叫 Alexa，这时候麦克风阵列实际上采集了正在播放的音乐和用户所叫的 Alexa 声音，显然语音识别无法识别这两类声音。回声抵消就是要去掉其中的音乐信息而只保留用户的人声，之所以叫回声抵消，只是延续大家的习惯，其实是不恰当的。
- 声源测向：这里没有用声源定位，测向和定位是不太一样的，而消费级麦克风阵列做到测向就可以，定位则需要更多的成本投入。声源测向的主要作用就是侦测到与之对话人类的声音以便后续的波束形成。声源测向可以基于能量方法，也可以基于谱估计，阵列也常用 TDOA 技术。声源测向一般在语音唤醒阶段实现，VAD 技术其实就可以包含到这个范畴，也是未来功耗降低的关键因素。
- 波束形成：波束形成是通用的信号处理方法，这里是指将一定几何结构排列的麦克风阵列的各麦克风输出信号经过处理（例如加权、时延、求和等）形成空间指向性的方法。波束形成

主要是抑制主瓣以外的声音干扰，这里也包括人声，比如几个人围绕 Echo 谈话的时候，Echo 只会识别其中一个人的声音。

端点检测：端点检测，英语是 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 通常都用信号处理的方法来做，之所以这里单独划分，因为现在 VAD 的作用其实更加重要，而且通常 VAD 也会基于机器学习的方法来做。

特征提取：声学模型通常不能直接处理声音的原始数据，这就需要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，然后将这些序列输入到声学模型。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。

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，每次所提到的语音识别进展，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，我们把每个阶段应用最为广泛的模型介绍一下，其实现在很多模型都是在混用，这样可以利用各个模型的优势，对于场景的适配更加鲁棒。

- GMM，Gaussian Mixture Model，即高斯混合模型，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。

- HMM，Hidden Markov Model，即隐马尔可夫模型，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。
- DNN，Deep Neural Network，即深度神经网络，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。
- RNN，Recurrent Neural Networks，即循环神经网络，CNN，Convolutional NeuralNetworks，即卷积神经网络，这两种神经网络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。其中，RNN 模型主要包括 LSTM（多隐层长短时记忆网络）、highway LSTM、Residual LSTM、双向 LSTM 等。CNN 模型包括了时延神经网络（TDNN）、CNN-DNN、CNN-LSTM-DNN（CLDNN）、CNN-DNN-LSTM、Deep CNN 等。其中有些模型性能相近，但是应用方式不同，比如双向 LSTM 和 Deep CNN 性

能接近，但是双向 LSTM 需要等一句话结束才能识别，而 Deep CNN 则没有时延更适合实时语音识别。

语言模型：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。

解码搜索：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。

1.2 语音识别数据知识

数据采集：主要是将用户与机器对话的声音信息收集起来，一般分为近场和远场两个部分，近场采集一般基于手机就可完成，远场采集一般需要麦克风阵列。数据采集同时还有关注采集环境，针对不同数据用途，语音采集的要求也很不一样，比如人群的年龄分布、性别分布和地域分布等。

数据清洗：主要是将采集的数据进行预处理，剔除不合要求的语音甚至是失效的语音，为后面的数据标注提供精确的数据。

数据标注：主要是将声音的信息翻译成对应的文字，训练一个声学模型，通常要标注数万个小时，而语音是时序信号，所以需要的人力工时相对很多，同时由于人员疲惫等因素导致标注的错误率也比较高。如何提高数据标注的成功率也是语音识别的关键问题。

数据管理：主要是对标注数据的分类管理和整理，这样更利于数据的有效管理和重复利用。

数据安全：主要是对声音数据进行安全方便的处理，比如加密等，以避免敏感信息泄露。

1.3 语音识别开源平台

目前主流的开源平台包括 CMU Sphinx、HTK、Kaldi、Julius、iATROS、CNTK、TensorFlow 等，CMU Sphinx 是离线的语音识别工具，支持 DSP 等低功耗的离线应用场景。由于深度学习对于语音识别 WER 的下降具有明显的作用，所以 Kaldi、CNTK、TensorFlow 等支持深度学习的工具目前比较流行，Kaldi 的优势就是集成了很多语音识别的工具，包括解码搜索等。具体的开源平台汇总如表 1 所示。

支撑技能

2.1 声学器件

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，这是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。

2.2 计算芯片

DSP，Digital Signal Processor，数字信号处理器，一般采用哈佛架构，具有低功耗运算快等优点，主要应用在低功耗语音识别领域。

ARM，Acorn RISC Machine，是英国公司设计的一种 RISC 处理器架构，具有低功耗高性能的特点，在移动互联网领域广泛应用，目前 IOT 领域，比如智能音箱也是以 ARM 处理器为主。

FPGA, Field - Programmable Gate Array，现场可编程门阵列，是 ASIC 领域中的一种半定制电路，既解决了固定定制电路的不足，又克服了可编程器件门电路有限的缺点。FPGA 在并行计算领域也非常重要，大规模的深度学习也可以基于 FPGA 计算实现。

GPU，Graphics Processing Unit，图形处理器，是当前深度学习领域最火的计算架构，事实上深度学习领域用到的是 GPGPU，主要是进行大规模计算的加速，GPU 通常的问题就是功耗过大，所以一般应用到云端的服务器集群。

另外，还有 NPU、TPU 等新兴的处理器架构，主要为深度学习算法进行专门的优化，由于还没有大规模使用，这里先不详叙。

2.3 声学结构

阵列设计，主要是指麦克风阵列的结构设计，麦克风阵列一般来说有线形、环形和球形之分，严谨的应该说成一字、十字、平面、螺旋、球形及无规则阵列等。至于麦克风阵列的阵元数量，也就是麦克风数量，可以从 2 个到上千不等，因此阵列设计就要解决场景中的麦克风阵列阵型和阵元数量的问题，既保证效果，又控制成本。

声学设计，主要是指扬声器的腔体设计，语音交互系统不仅需要收声，还需要发声，发声的质量也特别重要，比如播放音乐或者视频的时候，音质也是非常重要的参考指标，同时，音质的设计也将影响语音识别的效果，因此声学设计在智能语音交互系统也是关键因素。

03

应用技能

语音识别的应用将是语音交互时代最值得期待的创新，可以类比移动互联时代，最终黏住用户的还是语音应用程序，而当前的人工智能主要是基础建设，AI 的应用普及还是需要一段时间。虽然 Amazon 的 Alexa 已经有上万个应用，但是从用户反馈来看，目前主要还是以下几个核心技术点的应用。

3.1 语音控制

事实上是当前最主要的应用，包括了闹钟、音乐、地图、购物、智能家电控制等等功能，语音控制的难度相对也比较大，因为语音控制要求语音识别更加精准、速度更快。

3.2 语音转录

这在比如会议系统、智能法院、智能医疗等领域具有特殊应用，主要是实时将用户说话的声音转录成文字，以便形成会议纪要、审判记录和电子病历等。

3.3 语言翻译

主要是在不同语言之间进行切换，这在语音转录的基础上增加了实时翻译，对于语音识别的要求更高。

下面这三种识别，可以归为语音识别的范畴，也可以单独列成一类，这里我们还是广义归纳到语音识别的大体系，作为语音识别的功能点更容易理解。

声纹识别，声纹识别的理论基础是每一个声音都具有独特的特征，通过该特征能将不同人的声音进行有效的区分。声纹的特征主要由两个因素决定，第一个是声腔的尺寸，具体包括咽喉、鼻腔和口腔等，这些器官的形状、尺寸和位置决定了声带张力的大小和声音频率的范围。第二个决定声纹特征的因素是发声器官被操纵的方式，发声器官包括唇、齿、舌、软腭及腭肌肉等，他们之间相互作用就会产生清晰的语音。而他们之间的协作方式是人通过后天与周围人的交流中随机学习到的。声纹识别常用的方法包括模板匹配法、最近邻方法、神经元网络方法、VQ 聚类法等。

情感识别，主要是从采集到的语音信号中提取表达情感的声学特征，并找出这些声学特征与人类情感的映射关系。情感识别当前也主要采用深度学习的方法，这就需要建立对情感空间的描述以及形成足够多的情感语料库。情感识别是人机交互中体现智能的应用，但是到目前为止，技术水平还没有达到产品应用的程度。

哼唱识别，主要是通过用户哼唱歌曲的曲调，然后通过其中的旋律同音乐库中的数据进行详细分析和比对，最后将符合这个旋律的歌曲信息提供给用户。目前这项技术在音乐搜索中已经使用，识别率可以达到 80%左右。