分组演讲报告

我的分组演讲课题是语音识别（ASR），语音识别技术定义是让机器通过识别和理解过程把语音信号转变为相应的文本或命令的高新技术。语音识别技术主要包括特征提取技术、模式 匹配准则及模型训练技术三个方面。语音识别是人机交互的基础，主要解决让机器听清楚人说 什么的难题。人工智能目前落地最成功的就是语音识别技术。语音之所以可以转化成文字，用简单的话叙述如下：正如同英文单词是由英文字母组成的，声音也是由音素组成的，机器先收 到并录入人们的声音，并将声音转化成一段长长的声波。随后机器将长长的声波分成极其小的 一段，每一段叫做一帧。声学模型可以将已知的所有因素与录入的小段声波进行对比，并找出 最有可能匹配的音素。随后将音速串在解码器中，与相应的文字进行对比。对比所用的算法和 模型中行中，马尔可夫模型，是如今最常用最实用的。其他与语音识别相关的应用，如语音控 制系统，智能对话查询系统等使用的方法，都与此相关。从上世纪50年代起，科研人员就对语 音的音素，开始了相关的研究，上世纪70年代，马尔可夫模型在语音识别中的运用，使语音识别真正有了走向百姓人家的可能性。随着大数据，深度学习等技术的发展，让语音识别技术， 在2010年开始快速发展，现如今主流的语音识别系统的准确率均超过了百分之90，业界领导者 科大讯飞的语音识别率更是达到了百分之98，不仅是科大讯飞，包括百度，搜狗，谷歌等厂家 的语音识别系统，均已达到了可以商用的级别。语音识别准确率之所以快速提升，马尔可夫模 型发挥了至关重要的作用。

3.1 隐马尔科夫模型

HMM 是一种用参数表示的用于描述随机过程统计特性的概率模型, 它是由马尔可夫链演 变来的 . 所以它基于参数模型的统计识别方法 . 它是一个双重随机过程——具有一定状态数的 隐式马尔可夫链和显示随机函数集 . 每个函数都与链中一个状态相关联 . 隐式过程通过显示过程所产生的观察符号序列来表示。“它把一个总随机过程看成一系列状态的不断转移。时刻 t 的状态用qt 表示, 它可以是N 种状态集合S = [s1, s2, ⋯, sN ]中的任意一个。马尔可夫模型的特 性主要用“转移概率”来表示。后一状态出现的概率决定于其前出现过的状态次序。即: 状态 qt 出现的概率为 P r[qtg qt- 1, qt- 2, ⋯, q1 ]。如果此概率只决定于前一个状态, 即 P r[qtg qt- 1 ], 则称为一阶马尔可夫过程。它是研究中引用得最多的形式, 即: P r [qtg qt- 1, qt- 2, ⋯, q1 ] = P r [qtgq t- 1 ]。隐马尔可夫模型 (H idden M arkov M odel, HMM ) 则认为模型的状态是不可 观测的 (这便是“隐”得名的由来)。能观测到的只是它表现出的一些观测量(observations)。” [3] HMM模型也有着很大的改建空间，如《隐马尔可夫模型在语音识别中的应用》中所言“HMM 技术之所以在语音识别中应用较为成功, 主要是它具有较强的对时间序列结构的建模能力, 尽管 如此, HMM 仍然是有缺点和局限性的: (i) 对低层次的声学音素建模能力差, 使声学上相似的词 易混淆; (ii) 对高层次语音理解或语义建模能力差, 使其仅能接受有限状态或概率文法等简单场 合应用; (iii) 一阶HMM 假设很难直接用模型描述协同发音(coarticulation) , 因为HMM 假设输 出是相互独立的, 且依赖于当前状态。”

3.2 语音合成

语音合成（text-to-speech）与语音识别原理相同，使用到的技术、方法也大致相同。这种 技术目前可以做到文本转换语音、音乐生成、语音生成、语音支持设备、导航系统等功能，就 目前来看将来这项功能在未来最有可能用于为视障人士提供无障碍服务。语音合成一般可分为 两个步骤：文本处理和语音合成（前端和后端）。 首先第一个步骤文本处理，与语音识别中将音素逐步转化成文字的过程正好相反，语音 合成要将文本转换成音素序列。但这个过程与音素转文字又有所不同，它有一些新的问题需要解决，比如拼写相同但读音不同的词的区分、缩写的处理、停顿位置的确定、情感预测等 等。“This paper has presented WaveNet, a deep generative model of audio data that operates directly at the waveform level. WaveNets are autoregressive and combine causal ﬁlters with dilated convolutions to allow their receptive ﬁelds to grow exponentially with depth, which is important to model the long-range temporal dependencies in audio signals. We have shown how WaveNets can be conditioned on other inputs in a global (e.g. speaker identity) or local way (e.g. linguistic features). When applied to TTS, WaveNets produced samples that outperform the current best TTS systems in subjective naturalness. Finally, WaveNets showed very promising results when applied to music audio modeling and speech recognition.”由此说明WaveNet仍然是语音合成后端的最优方法。随着语音合成的不断发展，人们开始意识到当 时语音识别的操作过于复杂，文字的生成速度过慢。自然，人们就想到，能不能将语音合成的 两个步骤简化为一个步骤，这样一来，文字生成速度必然能大幅增加，之后科研工作者就开始 了End-to-End模型的研究。《一个无条件的端到端神经音频生成模型》说明：“我们提出了 一个新的模型，可以解决在原始声学领域的无条件的音频生成，这通常已经完成，直到最近手 工制作的功能。我们能够证明，时间尺度的层次结构和频繁的更新将有助于克服对非常高分辨 率的时间数据建模的问题。这使得我们可以直接从音频样本中学习数据流形。结果表明，该模 型能很好地泛化和生成三个性质不同的数据集上的样本。我们还表明，该模型生成的样本是人类的首选。”

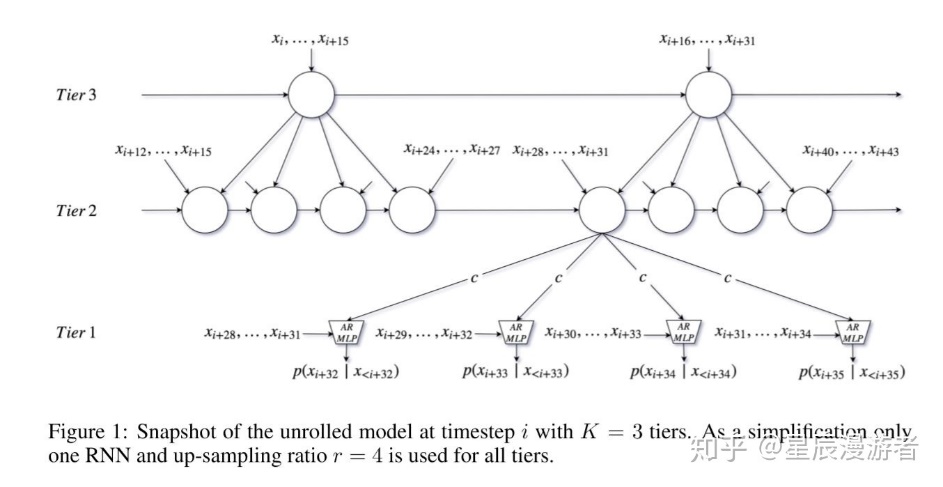


图: ent to end 初始模型

3.3 语音识别的商业盈利模式

现如今，使用场景越来越广阔。2014 年，全球智能语音识别市场规模达 2750亿美元，2016年 达4210亿美元，同比增长百分之53.1。有经济学家指出，这一市场规模到2025年将超过10000亿 美元。即使如此，市场上的语音识别企业，大多仍然处于亏损状态，曾有数据显示，2018年全 年，近百分之90的人工智能公司处于亏损状态，而百分之10赚钱的企业基本是技术提供商。在极少数可以盈利的企业中，讯飞便是其中之一，但它并不典型，它既是产品提供商也是技术 提供商。C段业务住在成为讯飞输入的主力，打开讯飞的首页，我们可以看到它不同于其他语 音公司的软件，它有着讯飞翻译机，学习本录音笔等多种智能硬件，软件中以讯飞输入法为 主等多种软件共同实现营收，以讯飞听见为例，讯飞听见通过翻译时长需要购买的方式来创造 收入，该公司之所以敢于采取这种营收方式，源于他对自己实力的自信。讯飞听见可以快速转 写（约四十字每分钟），以及中英文实时翻译的功能都是业界领先的。由此见得，创造收入的 前提是自身技术的优越。应用声纹识别技术来确认身份,高精度的身份确认,可以使军事系统的 安全性大大增强。一些计算机产品,为了安全起见,也使用了声纹识别技术,例如在普通的移动存 储设备上增加声纹认证功能,对电脑系统进行语音加密保护,用以保护个人隐私和军事机密。在 保密方面，语音识别的优越性不言而喻。安保方面在未来也注定成为语音识别企业的盈利方面 之一。随着智能化时代的到来，智能家居终将普及，家居智能的实现，离不开语音识别。想要 增强智能家居消费者的体验感，舒适度。语音识别的精度，速度将会有着更高的需求。不仅是 智能家居，未来大量的智能产品都将以语音识别为基础，这无疑会增加语音识别企业接受的投 资，企业以此来提升自己的产品实力，也是公司最终实现盈利的方式之一。 声，聆音千里之臻，洞悉万象之源，重塑数据价值，强化机器学习，真实环境下语音识别 率将全面提升，为更自由的人机语音交互提供保障，感你所言，知你所想，在纷繁的环境下， 敏锐地感知你的声音，全方面无延迟地取景，每一声呼唤都闻声即映。领先的声学技术可以感 知你的方向，倾听你的表达，消除空间混响，抵消声源回声，用灵敏的“耳朵”让人机交互更 加随心自如，声纹让交流专属独享，识别语调情绪，让沟通快速高效。聚合云端应用，让智能 自由分发，智能汽车，玩具，机器人，智能家居，医疗，安防……语音识别将引领智能化时代， 构建沟通的桥梁。繁华而有序的智能生活将在未来实现已成为了必然，在这种大趋势，语音识 别的盈利问题自然会迎刃而解。