iResearch
艾 瑞 咨 询

凝望璀璨星河：

中国智能语音行业研究报告

2020年

摘要

人类对机器语音识别的探索始于20世纪50年代，迄今已逾70年。2016年，在深度神经网络的帮助下，机器语音识别准确率第一次达到人类水平，意味着智能语音技术落地期到来。不过人们面对“AI”时希望得到自然、类人的交互体验，这是一个宏伟的开放性课题，背后涉及的各学科技术仍有不足，还面临长期的求索方能突破。

消费级智能硬件是最早显示出市场潜力的赛道，市场各方都在瞄准消费级智能交互终端。而智能终端的背后还有广阔的生态，包括语音开放平台、语音操作系统、内容等等，近年行业正在经历从单一商业模式向多元化商业模式的变迁，技术输出的“厚度”增加，“边界”扩大，也带来了技术落地曲线的加速度增加。

智能语音企业级和公共级市场主要有平台化技术输出和解决方案两类商业模式，解决方案业务占比较高。与国外市场以医疗为重头有所差异，我国市场以智能客服、公检法及教育业务份额更高。智能语音为各行业解决了刚需性问题，将促进各行业业务效率的提升。

目前全国约有超过250家企业参与智能语音语义市场。互联网巨头、技术提供方、设备商和行业集成商应分别重视连续性投入支持问题、基础开发模块标准化程度提升与商务团队配置问题、设备后服务增长问题和软件研发能力建设问题，迎接人机交互升级带来的行业价值链扩张。

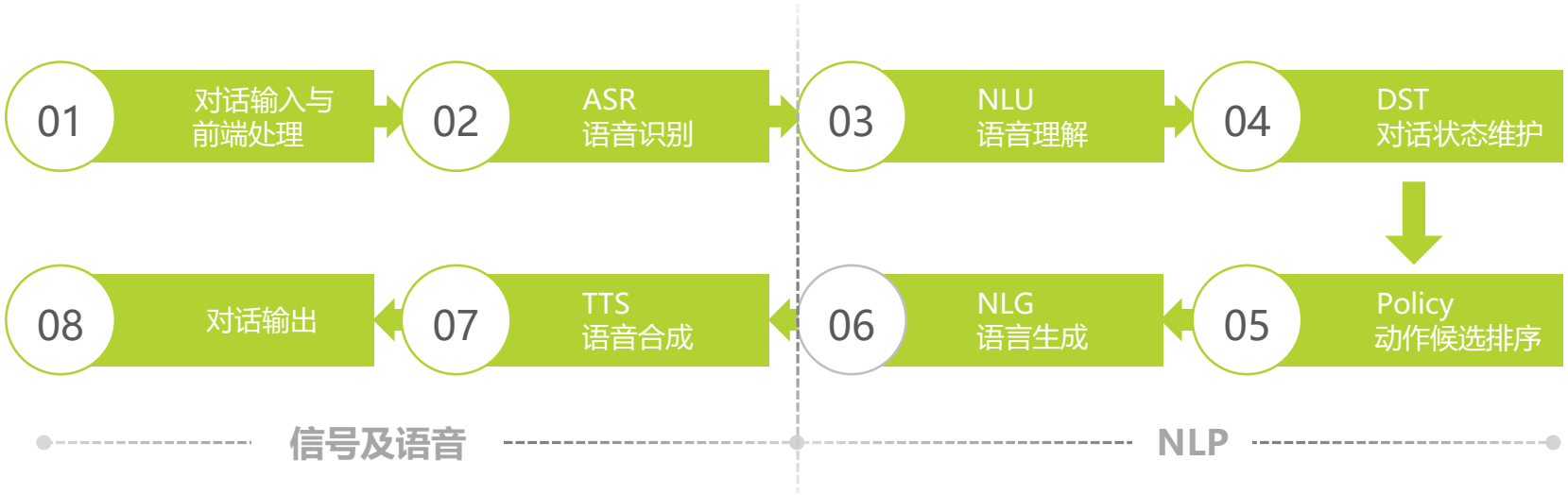
来源：艾瑞自主研究绘制。

智能语音的概念

智能语音即实现人与机器以语言为纽带的通信

智能语音即实现人与机器以语言为纽带的通信。人类大脑皮层每天处理的信息中，声音信息占20%，它是沟通最重要的纽带，人机对话将方便人们的工作与生活。完整的人机对话包括声音信号的前端处理、将声音转为文字供机器处理、在机器生成语言之后，用语音合成技术将文本语言转化为声波，从而形成完整的人机语音交互。

人机对话的实现流程



来源：艾瑞《2018年中国人工智能行业研究报告》；百度AI。

智能语音的前情提要（1/3）

机器“听觉”本质上是对声音特征和文本的分类任务

人的听觉形成过程是将声能转变为机械能、再转为生物电信号，在听觉中枢加工、分析的结果，而机器的“听觉”则经过声音信号-音频信号-电信号-特征向量-解码为文字-理解的过程，本质是对声音特征和文本的分类任务（将字音分类对应为文字、将文字对应为潜在语义），如果需要机器感知声音的起止和音色等特征，还需要另外进行信号处理与特征分类任务。

人与机器的“闻音知意”

外耳中耳：声音传递

- 声音收集
- 将声能转化成机械能并放大，把声波传送至内耳

内耳：声波转换

- 通过毛细胞将声波机械能转化为生物电信号，经由听神经传导

听觉中枢：主观听觉感受

听觉中枢纵跨脑干、中脑、丘脑的大脑皮层

- 听觉掩蔽——噪声分离
- 空间听觉——声源定位、声源分离和环境感知
- 判断声音起止
- 感知声音特征
- 与管理“读”、“写”、“说”的语言中枢关联

“人们之所以能听到声音、现解言语，是依赖于由耳、听神经、听觉中枢组成的听觉通路。其中，听觉的形成部位是听觉中枢”

麦克风阵列拾音器

- 接收声音信号转化为音频信号并放大
- 声源定位
- 去混响
- 语音增强
- 声源信号提取/分离

语音识别

- 将模拟信号转化为电信号，用向量描述信号的多维特征（如用MFCC特征刻画声音的频谱、共振峰等）
- 将特征识别为状态，组合成音素（声学模型）
- 根据语言表达的特点，选取音素所对应的正确的文字，最终完成由向量转化为文字的过程（语言模型）

语言理解

- 根据文本，通过领域识别模块识别对话发生的背景领域，再通过人的意图检测模块识别出人希望触发的行为
- 通过槽填充，补全人的意图，将隐式的意图转化为显式的指令，为机器提供一个规范的语义表示

“机器的闻音知意本质上是对声音特征和文本的分类任务，当然通过声学技术保障拾音效果同样重要。如果需要机器感知声音的起止和音色等特征，还需要另外进行信号处理与特征分类任务”

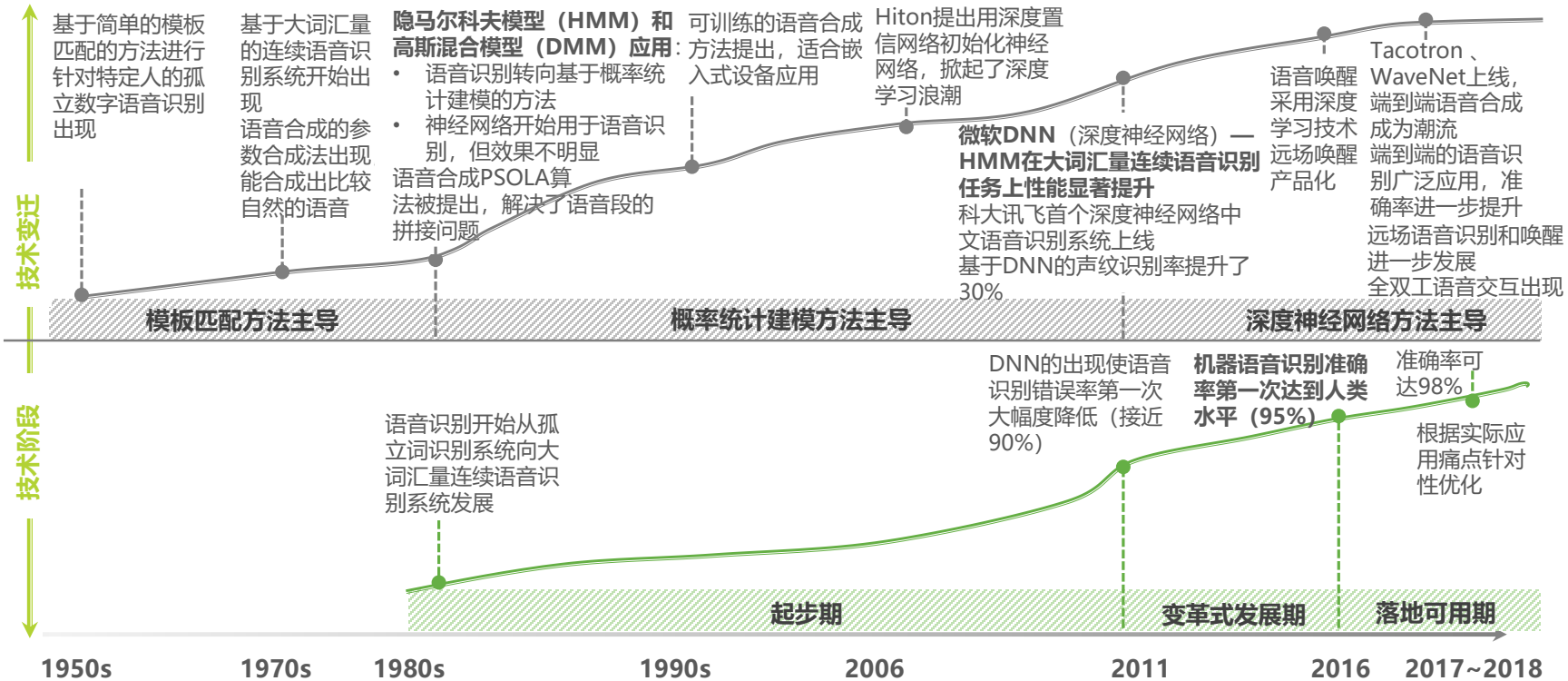
来源：艾瑞根据公开资料自主研究绘制。

智能语音的前情提要（2/3）

深度神经网络是智能语音技术近年达到落地可用的推动器

2011年，微软研究院提出的基于*上下文相关深度神经网络*和*隐马尔可夫模型*的声学模型在大词汇量连续语音识别任务上获得了显著的性能提升，从此大量研究人员开始转向深度学习在智能语音领域的研究，2016年，机器语音识别准确率第一次达到人类水平，意味着智能语音技术的落地期到来。近年，研究方向主要是端到端神经网络及针对实际应用中的算法优化。

智能语音技术发展历程示意图 *（以语音领域模式识别为主）*



注释：（1）目前端到端的语音合成指打通文本端-声学端，或声学端-波形端，直接从文本到波形的端到端尚不能实现；端到端的语音识别也是指打通声音特征端-文本端，波形-信号处理-声学模型-语音模型-文本的端到端尚不能实现。端到端的方法有助于训练效率和效果提升。

（2）准确率数据指近场语音识别准确率。

来源：艾瑞根据Economist、公开资料、专家访谈，整理研究绘制。

智能语音的前情提要（3/3）

所涉学科及其研究任务

声学信号	声源定位	用于确定声源方向和距离，主要应用于语音交互设备对声源进行定位和海洋声学中的声源定位/方位估计。主流方法包括波束形成，超分辨谱估计和TDOA等
	语音增强	当语音信号被各种各样的噪声干扰后，深度神经网络模型利用大量数据，对噪音成分和语音成分进行有效估计，从含噪声的语音信号中提取出纯净语音，对于智能语音的完成非常重要
	去混响	弱化混响引起的不同步的语音相互叠加、从而提升语音识别效果。主要方法有基于盲语音增强的方法、基于波束形成的方法、基于逆滤波的方法
	回声抵消	即自噪声抑制，去除语音交互设备自己发出的声音，而只保留用户的人声
	其他方向	将机器学习应用进生物声学、地质探测等
模式识别	声纹识别	生物识别技术的一种，从应用方向看包括说话人辨认（匹配特定说话人）、确认与聚类（区分不同说话人音频片段），需要用到声学处理和深度神经网络处理人说话时的短时频谱、声源、时序动态、韵律等特征
	语音唤醒	属于信号处理（SSP）的一部分。在连续语流中实时检测出说话人特定片段，将设备从休眠状态激活至运行状态。实现方法有基于置信度、基于识别和基于垃圾词网络的唤醒；目前主流应用类型有：先唤醒再指令、将唤醒词和指令一同说出、将常用用户指令设置为唤醒词等。目前远场的智能硬件设备如机器人、智能音箱可支持3-5米的远场唤醒
	语音识别	通过将人类语音转换为计算机可读的输入，由特征提取、声学模型、语言模型组成，包括近场识别、远场识别，近年的应用中还涉及切分说话人、全双工语音等
	特定声音检测	通过特征提取与算法训练，使机器能够完成对不同人群、不同乐音等特定声音检测
	谎言检测	提取谎言中微颤抖所引起的语谱局部能量变化，将所提取的特征作为神经网络输入进行谎言识别
自然语言处理	自然语言理解	将用户的输入映射到预先根据不同场景定义的语义槽中，让机器理解语言的意思。通常包括三个任务：领域检测、意图识别和语义槽填充
	对话管理	考虑历史对话信息和上下文的语境等信息进行全面地分析，决定系统要采取的相应的动作，如追问、澄清和确认等。主要任务有：对话状态跟踪和生成对话策略。实现途径上，目前有检索模型、生成模型等。
	自然语言生成	将机器输出的抽象表达转换为句法合法、语义准确的自然语言句子
语音合成	语音合成	把文字智能地转化为自然语音流，也就是输入是文本，输出是波形；近年个性化TTS、带有情绪的TTS成为热点

来源：艾瑞根据CSDN、中科院声学研究所、《计算机学报》、知乎专栏《子鱼说声学》等公开资料研究绘制。

2020年建议重点关注的技术方向（1/3）

声学感知空间环境：解决多智能设备无法配合的困扰

随着智能语音算法基础性能不断提升，识别准确率、时延问题已不再是交互体验的核心痛点，人们希望让智能设备具备更多的基本能力，例如能够感知环境，当同一个房间里有多个智能交互设备或多台智能交互设备分布在不同的房间时能准确唤醒，过去通过设备间蓝牙通信可以解决由哪台设备被唤醒与人对话，但无法解决相关的家居控制执行问题。2019年，业内玩家开始重视将声学感知空间的能力与交互系统结合起来，实现多智能交互设备的就近唤醒应答，避免多设备重复响应和执行指令，在这种情形下并不存在某个中心交互设备，因此也被称为分布式场景。

未来，设备之间的隔阂可能被进一步打破，如使任何形态、任何配置的终端设备通过连接协议实现AI能力共享、算力共享（而不仅限于目前用一个设备通过连接协议对其他设备语音控制），就可能使场景内适宜拾音的设备与人交互、适宜功放的设备配合放音，使多设备的协同达到效率最优。

声学信号 → 模式识别 → 自然语言处理

智能交互终端分布式场景的应用与展望