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”

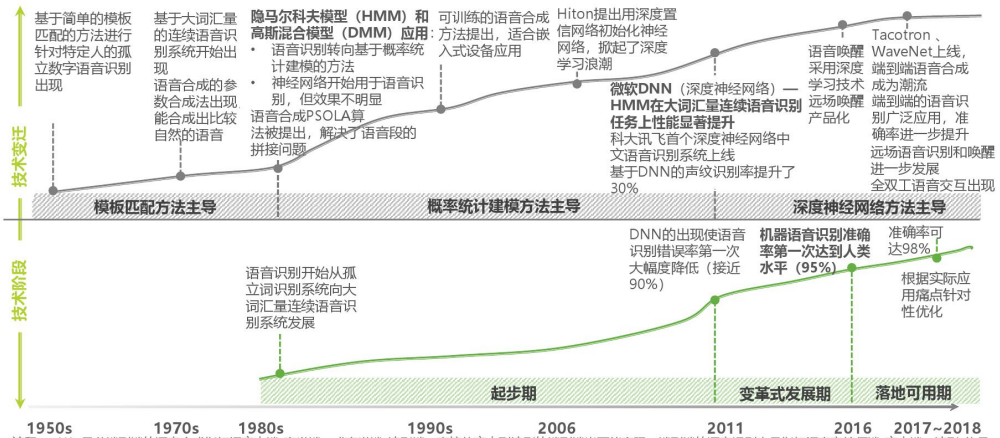
来源：艾瑞根据公开资料自主研究绘制。

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。

智能交互终端分布式场景的应用与展望

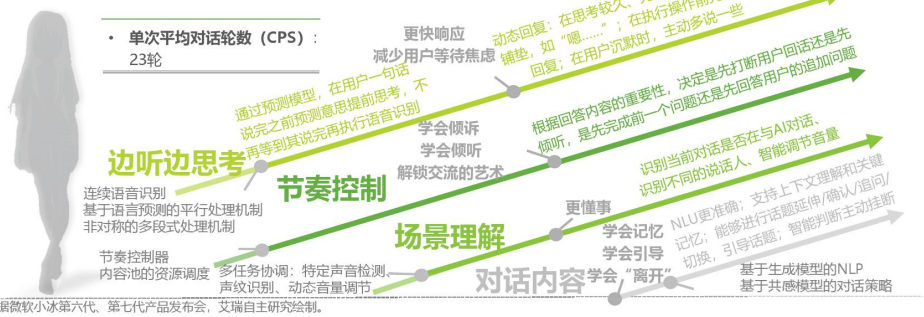
来源：艾瑞根据小米小爱同学3.0，华为HiAI3.0公开资料自主研究绘制。

2020年建议重点关注的技术方向（2/3）

全双工语音：由处理语音消息升级为处理语音流

鉴于目前机器的智能语音交互能力是基于分类任务实现的，其智能程度的提升有赖于技能一项一项地填充补足，最终使交互体验得到质变。上文介绍了人机对话和语音识别（ASR）的基本实现过程，相比于普通以语音消息作为交互的人机对话，全双工则是处理语音流，能够实时预测人类即将说出的内容，实时生成回应，并控制对话节奏。多家厂商在持续投入全双工的研发，全双工、多轮对话、单轮对话对比如下：全双工——只需一次唤醒，保持进行连续的语音流分析（机器保持听+想的状态，即使在它回话的时候也同步在听和想）；多轮对话——只需一次唤醒，听、想、说分离，机器会在它的本句回话完成后才再次开始听用户说话、听完再分析；单轮对话——每一次用户说话前都需要先唤醒设备。除了基本的对话IQ与EQ外，让机器实现跨情景流畅切换的全双工（如内容、导航、查询、设备控制的跨情景切换）也是重要研究方向，目前市场上绝大部分机器都只支持单轮对话或多轮对话，真正搭载了完整、成熟全双工语音能力的产品还很少。

声学信号 → 模式识别 → 自然语言处理　**全双工语音的智能之处——以微软小冰为例**



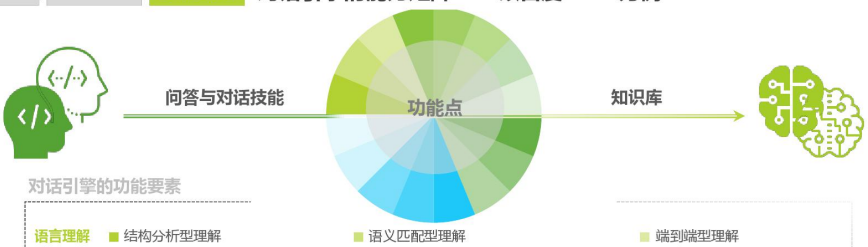
来源：根据微软小冰第六代、第七代产品发布会，艾瑞自主研究绘制。

2020年建议重点关注的技术方向（3/3）

对话引擎：支撑问答与对话内容实现的核心

对话引擎是支撑人机交互中问答和对话内容实现的核心，广泛应用于智能客服、智能交互设备、智能车载系统等领域，核心功能包括语言理解力、对话管理、知识库和帮助开发者定制开发扩展应用的工具。知识的指导对对话引擎十分重要，其中知识图谱及图谱知识库构建工具能够直接从业务文档抽取知识、建立规则，而不局限于整理好的问答对，这不仅可以帮助机器找到直接的答案来源，还可以使机器依据元素的属性与关系理解语义、形成话题推荐等对话策略。

声学信号 → 模式识别 → 自然语言处理　**对话引擎的能力矩阵——以百度UNIT为例**



对话引擎的功能要素

语言理解	结构分析型理解	语义匹配型理解	端到端型理解
对话管理	分发式对话管理	流程式对话管理	异常对话处理与在线学习
知识库	键-值型知识库	实体-关系型知识库	无结构文本知识
开发工具	对话理解开发环境	对话管理编程框架	知识库构建工具	日志分析工具

对话引擎的服务要素　丰富的预置技能与资源　多样化的部署方案　SLA与合理的定价　完善的培训与认证体系

来源：根据百度UNIT3.0，艾瑞自主研究绘制。

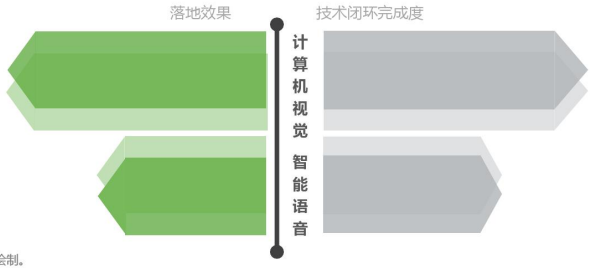
本章小结

技术闭环完成度有待提升，面临长期的求索方能突破

智能语音背后涉及的声学研究、模式识别研究、通用NLP研究及垂直场景的深度语义理解等还未成熟到拼成一个没有明显短板的“木桶”，在交互体验、使用效果、场景性优化等方面都还有很长的路。与人工智能发展最快的分支计算机视觉相比，尽管二者都凭借深度学习取得重大突破，并在识别准确率上达到人类水平，但计算机视觉通过人脸识别这一大技术分支便高完成度地解决1:1或1：N比对问题，快速渗透到了各行各业；智能语音技术要解决的却远远不是1:1或1：N的比对，而是人们面对“AI”时希望得到的自然、类人、甚至高信息密度的交互体验，这是一个宏伟的开放性课题，因此尽管智能语音已取得了一些商业上的成就，但仍面临长期的求索方能突破。

智能语音与计算机视觉的差异

- **计算机视觉**：以人脸识别为代表的图像分类技术主要解决**1:1或1：N比对**问题，解决场景问题的技术闭环基本形成
- **智能语音**：解决人们面对“AI”时希望得到的自然、类人、以及一些场景下需要高信息密度的**交互体验**的问题，相关技术还未成熟到拼成一个没有明显短板的“木桶”



来源：结合艾瑞《2017年中国计算机视觉行业研究报告》，自主研究绘制。

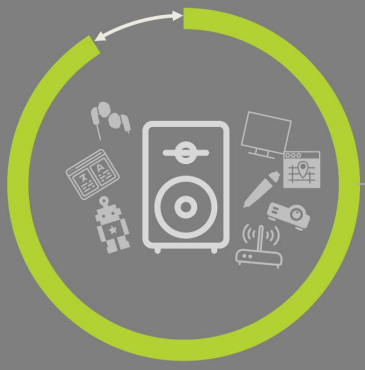
子研究（1/3）

消费级市场

消费级智能硬件

智能音箱研究单元

语音输入法



2019年**市场**
超两千亿元

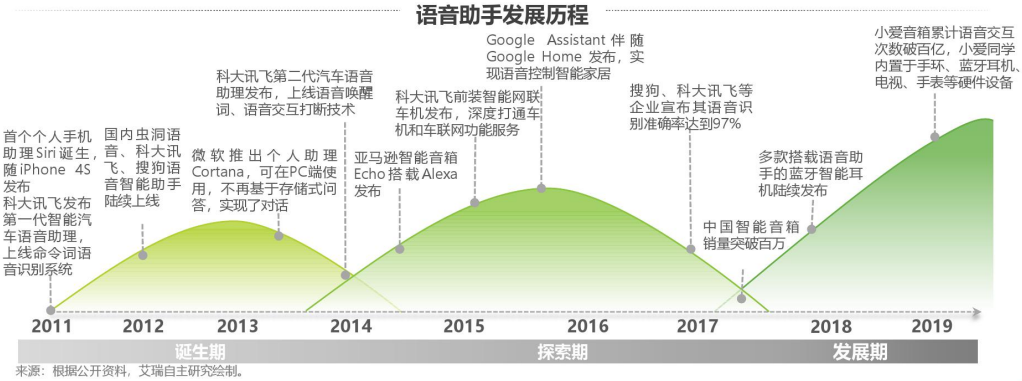
语音助手及其发展历程

智能语音助手赋能多类智能终端，构建全产业生态链

消费级智能语音交互是人们接触智能语音最普遍的渠道，从手机语音助手、家庭智能音箱、智能耳机、智能电视、故事机到智能车载等等，根据艾瑞《2020年中国智能物联网（AIoT）白皮书》，2018年消费级AIoT在总AIoT市场中占比68%，市场规模达到1753亿元，作为最早显示出市场潜力的赛道，无论是硬件设备厂商还是互联网公司、AI公司都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，包括面向开发者的语音开放平台、语音操作系统及音频内容等等。

语音助手是用于终端的语音控制程序，通过智能对话与即时问答的智能交互，让智能机器助手帮助完成用户指派的任务。2011 年第一款手机语音助手 Siri 伴随 iPhone 4S 亮相，各大厂商纷纷入局。从 2017年下半年开始，通过开放语音生态系统，进行产业内合作，语音助手向家居、车载、可穿戴设备等领域不断延伸和迁移，构建出全产业生态链。

语音助手发展历程



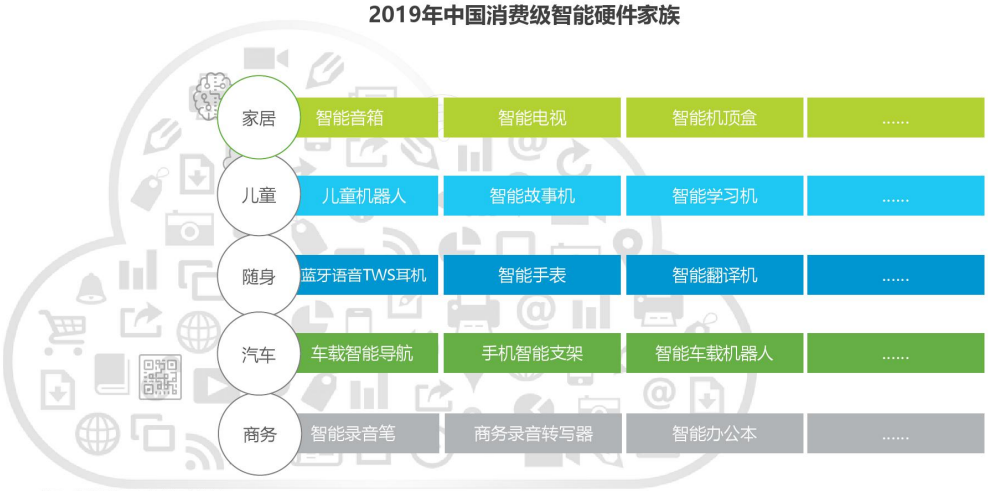
来源：根据公开资料，艾瑞自主研究绘制。

消费级智能硬件家族

通过语音助手或语音转写能力提供音频内容与任务处理服务

目前带有智能语音能力的消费级硬件大体可划分为智能家居、儿童产品、随身产品、车载设备、商务产品等。部分产品的交互特性更强，需要通过语音交互为用户提供音频内容和某些任务处理操作，例如智能音箱与车载设备可用于控制开关、收听FM、导航等；部分产品的功能性更强，例如智能录音笔的核心功能是为用户提供语音转文字服务。

2019年中国消费级智能硬件家族



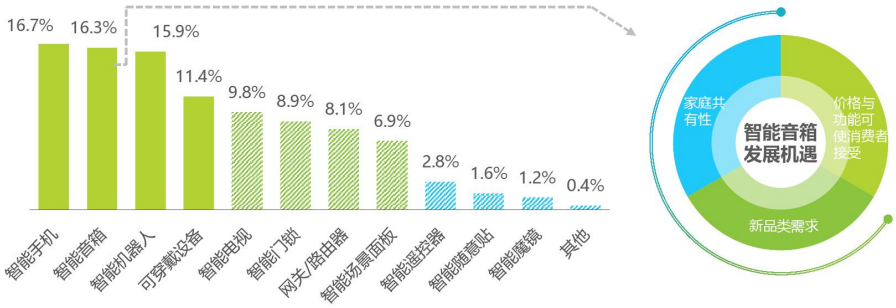
来源：艾瑞根据公开资料研究绘制。

智能音箱

为语音交互蓝图铺路，成为智能生活入口

近年，智能音箱作为智能生活的“入口” 的地位逐渐被夯实，主要得益于三点：（1）智能生活入口是一个偏重的服务，因此基于已有较重服务的品类上延伸发展（例如电视、遥控）在产品逻辑上不太通畅，这就给了新兴家庭智能终端发展机会，智能音箱虽然仍定位为“音箱”，但旧瓶新酒，更像是简易形态的机器人；（2）家庭场景是服务于全部家庭成员的，个人私有设备不太适宜执行与整个家庭成员交互的功能，而一些可穿戴设备在芯片等硬件配置上仍有不足，因此需要一款家用设备承载这类场景需求；（3）智能音箱产品价格适中，近年来价格下降趋势明显，且随着远场语音识别、基于NLP的对话和问答能力逐渐成熟，功能达到可用。与智能手机相似，智能音箱在初期也采取了大量补贴的策略，加速在市场上“占位”成为第一要务，便宜的价格使用户心理预期不会过高，这也为厂商持续优化赢得 “宽容” 时间。

2018年中国智能家居从业者最看好的用户入口



注释：图中用户入口指用于操控智能家居的入口。
样本：N=100，于2018年4月通过艾瑞专家渠道网络问卷调研获取。
来源：调研数据来自艾瑞《2018年中国智能家居行业研究报告》。

国内智能音箱卡位家庭流量迁移（1/3）

从PC与移动互联网看流量迁移模型

目前移动数据及互联网业务收入达到固定数据及互联网业务收入的三倍，说明互联网流量大量迁移到移动端。智能音箱厂商则希望在智能音箱从用户家庭场景流量中分一杯羹，成为家庭场景流量入口。借鉴移动互联网的经验，有三个关键因素将促成设备端口的流量增长：终端可得性、接入便利性、应用丰富性。

流量迁移模型及智能音箱成为家庭流量入口的优劣势分析

终端可得性——设备渗透率	★ ★ ★	手机：2012年，3G手机大量出货，渗透率达到**21%**，2013年Q1移动移动数据及互联网业务收入开始超过固定数据及互联网业务收入；2018年底，移动数据及互联网业务收入达到固定数据及互联网业务收入的三倍，互联网流量已大量迁移到移动端。 智能音箱：预计到2019年底，中国智能音箱累计出货量超过7200万台，在我国城镇住房中渗透率达到**20%**，且产品价格适中有望推动智能音箱进一步渗透，智能音箱作为家庭接网入口的条件初步具备。
终端可得性——设备活跃度	★	手机：2012年，手机上网率首次超过电脑（达到74.5%），**93%**的手机网民每天通过手机上网；使用手机成为超高频场景，2018年12月，手机上网率达到98.6%，手机网民平均每天上网时长达5.69小时。 智能音箱：智能音箱主打语音交互，由于用户体验尚未取得质变，使用频率仍然较低，只有**少数**用户会每天使用智能音箱；2018年底，每月会登录智能音箱App 的用户数只有智能音箱保有量的15.8%。
应用丰富性——应用数量	★	手机：2012年移动App如雨后春笋般涌现，当年8月仅App Store中国区应用数量就有**50.4万**个。截至2018年12月，我国总App在架数量为449万款，人均App数超40个。 智能音箱：目前智能音箱应用数量大约在**万位数**，潜在的应用想象力空间还比较空白。
应用丰富性——流量质量	★ ★	手机：相较于PC时代，移动互联网带来的短视频、直播、支付、网文、O2O丰富了人们的生活，让用户对手机的使用频率更高，而这些应用更多的是大流量应用，信息丰富度和用户吸引力较强。 智能音箱：目前以无屏款音箱居多，故用户应用大多是查询提醒、有声内容、家居控制，音频得到的流量入口时长也较短，信息量大的服务不易通过语音交互，成为智能音箱应用引流的考验，抓住时间窗口精准推荐成为势在必行的选择；不过其服务内容能较大程度满足老人和孩子的需求、吸引他们的关注，成为智能音箱产品的机会。
接入便利性	★ ★ ★ ★ ★	手机：2015年开始的提速降费解决了移动终端上网“管道”的问题，为移动互联网爆发带来契机。 智能音箱：不仅自身入网便利，且基本兼容所有连接通信协议，随着家庭场景中智能家居数量增加，智能音箱作为普及程度和交互性较好的控制设备需求提升。

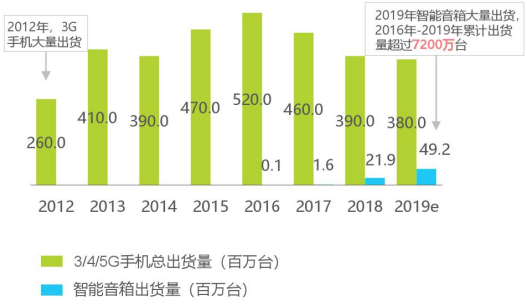
来源：图中引用数据来自艾瑞Mobile App Index 、CNNIC、工信部公报、互联网消费调研中心、 中国网络视听节目服务协会《2019中国网络视听发展研究报告》，图表由艾瑞自主研究绘制。

智能音箱的渗透情况

在我国城镇住房中渗透率达到20%

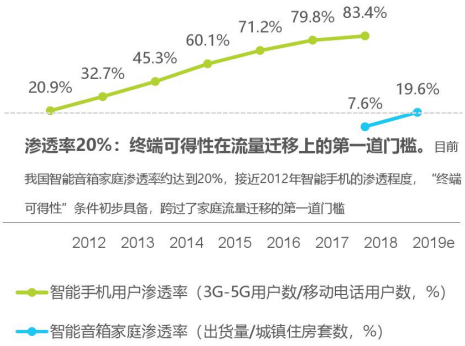
预计到2019年底，中国智能音箱累计出货量超过7200万台，在我国城镇住房中渗透率达到20%，接近2012年智能手机的渗透程度，“终端可得性”条件初步具备，跨过了家庭流量迁移的第一道门槛。

2012-2019年中国智能手机与智能音箱出货情况



来源：艾瑞根据工信部、Canalys、国家统计局数据整理绘制。

2012-2019年中国智能手机与智能音箱渗透情况



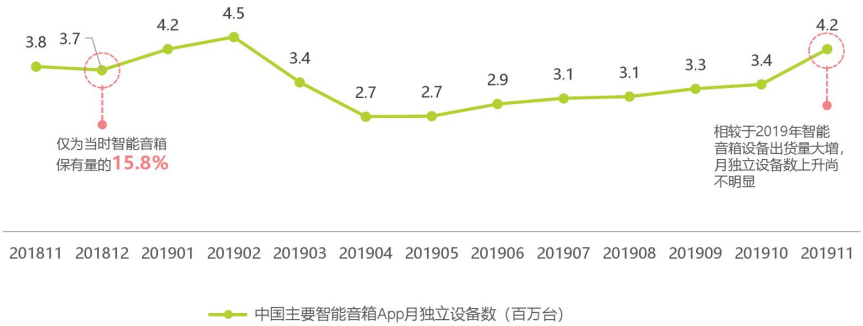
注释：因资料有限，智能音箱家庭渗透率数据是以出货量而非销量为依据的，且计算方式中并没能剔除我国智能音箱出货量中销往海外和乡村的部分，同时未考虑一户城镇住房拥有多台智能音箱的情况，仅做参考。
来源：艾瑞根据工信部、Canalys、国家统计局数据整理绘制。

智能音箱App活跃情况

与智能音箱累计出货量未成正比

智能音箱主打语音交互，由于使用体验等问题，使用频率仍然较低，只有少数用户会每天使用智能音箱进行交互；而智能音箱App作为未来流量变现的重要一环，其活跃度也不够乐观：2018年底，每月会登录智能音箱App 的用户只有智能音箱设备保有数的15.8%，后期由于新奇退去，活跃度在2019年上半年还产生了一定下滑，至2019年底，智能音箱App的活跃情况相对于大幅增长的出货量依然未成正比，背后的原因主要是智能音箱应用数量有限、品类较少，潜在的应用想象力空间还比较空白，同时信息量大的服务不易通过语音交互，也成为智能音箱应用引流的考验。

2018年11月-2019年11月中国主要智能音箱App月独立设备数



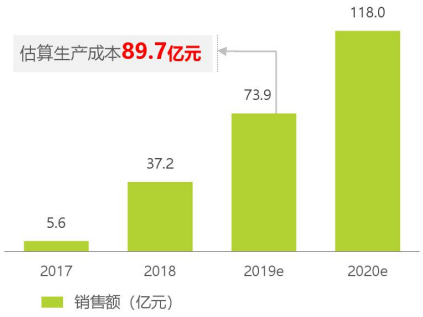
注释：（1）口径：包括天猫精灵、小度音箱、小度在家、小爱音箱、小雅音箱、叮咚音箱的月独立设备数。
（2）月独立设备数：该月使用过该App的设备总数，单个设备重复使用不重复统计。
来源：根据艾瑞Mobile App Index监测数据，加总处理绘制。

国内智能音箱卡位家庭流量迁移（2/3）

为什么说国内智能音箱会成为巨头的市场

2019年，尽管我国智能音箱硬件补贴已进入收缩阶段，补贴额依然达到15.8亿元（产品库存对该数值有一定影响），中小玩家难以支撑大量补贴，因此巨头占据了绝大部分市场。目前智能音箱市场主要由天猫精灵、小度音箱和小度在家、小爱音箱占据，互联网基因使它们在智能音箱产品上复制了互联网玩法——补贴攻城、低价策略、互联网服务运营回血，同时应用开发者的广泛聚拢、产品智能化提升的开发都需要强大的资金和资源支持，使智能音箱市场很难存在群雄并起的格局，智能音箱的流量也相应聚拢在大平台。而在智能音箱的生产成本中，麦克风阵列仍然是最大的部分。

2017-2020年中国智能音箱整机销售额



2019年中国智能音箱成本分布

不算营销、渠道、开发者补贴等，2019年我国智能音箱市场约补贴15.8亿元

AI算法授权费（麦克风阵列算法以外的部分）在总成本中占比约3.1%

注释：（1）此处补贴指年销售额与年出货量生产成本间的差距，不涉及厂商赠送会员服务、对开发者补贴、营销等带来的成本，因此数值受产品库存影响较大。误差会来自对智能音箱产品价格折扣率及总销量中以折扣价销售数量的误差；可能的成本分布误差会来自产品型号与配置的划分精细度不足。仅供参考。
（2）AI算法成本囊括了企业采用自研技术（无需对外支出成本）的情况，因此实际发生在市场中的交易量级应少于2.8亿元数值。

来源：艾瑞根据Canalys出货量、奥维云网销售量等基础数据及艾瑞推算模型，自主研究绘制。

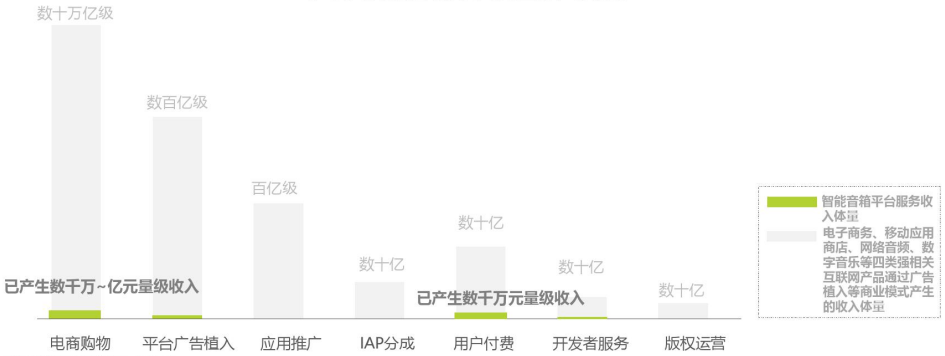
来源：根据专家访谈、市场上主流产品配置统计、不同品类销售情况，结合推算模型，艾瑞自主研究绘制。

国内智能音箱卡位家庭流量迁移（3/3）

流量的变现模式是下一步需要考虑的问题

目前终端设备销售以外的商业化还不是市场主要关注的问题，但已开始有一些尝试。**智能音箱的应用/技能基本是以设备绑定形式存在，因此品牌设备方本身也是平台方**（可以理解为智能音箱的核心预置应用、应用商店、主页、操作系统提供方），这为智能音箱更好地复制互联网变现模式打下了基础，电商购物、平台广告植入、应用推广和应用内购买（IAP）分成、用户增值服务付费、开发者服务等都是可能的变现方式，其中用户增值服务和电商购物已开始抢跑。与传统的互联网产品商业模式相比，由于前述智能音箱在活跃度、应用丰富性、流量质量等尚未取得突破，且口播广告不符合音箱产品使用逻辑、信息流及原生广告有待开发，因此广告形式、应用推广及IAP形式的变现还存在较大瓶颈。

2019年中国智能音箱平台商业模式探索



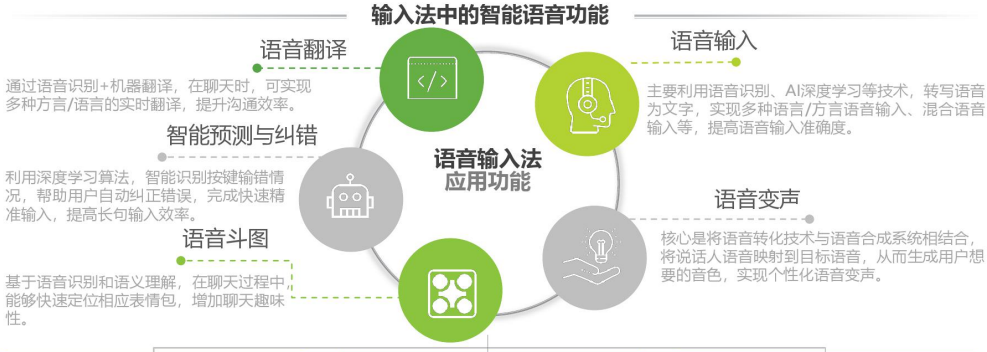
注释：智能音箱用户付费以会员费为主。
开发者服务指智能音箱平台方向开发者提供运营支持、云资源、通用软件功能模块支持及IoT模组等。
IAP分成指用户进行应用内购买增值服务后，平台与应用开发商对收入进行分成，此处不对会员费及电商购物产生的分成做重复统计。
来源：艾瑞根据公开资料和专家访谈自主研究绘制。

语音输入法

提升输入效率，满足个性化表达需求

输入法是智能语音技术在C端的重要落地场景，语音输入（多语种支持）、智能纠错、语音翻译等功能开始成为标配；语音变声、语音斗图等针对年轻群体的创新功能也相继推出。智能语音在输入法上的应用提升了用户的输入效率、更好地满足了用户在个性化表达上的需求，为产品本身增加了吸引力，以第三方输入法的头部产品搜狗输入法为例，个性化语音识别功能上线之后，搜狗输入法登陆率提升10.1%。

输入法中的智能语音功能



搜狗输入法	讯飞输入法	百度输入法
支持个性化语音识别；上线语音变声功能，可模仿明星、动漫发送仿真语音消息，实现中日韩语音随心译；支持多种语言输入、智能预测与纠错等功能；搜狗输入法日均语音请求峰值最高达到8.3亿次。	可识别23种方言、2种民族语言、4门外语，支持中英免切换输入、普通话方言免切换输入、在线离线免切换输入、语音实时翻译、方言语音翻译、中英混输、智能纠错等功能，支持定制个性化结果及语音修改后即时学习。	能够实现高精度的方言自由说、中英自由说，支持语音翻译、智能预测与纠错等功能、语音混输；离线语音识别精准度达到与在线相同水平，准确率超过98%。

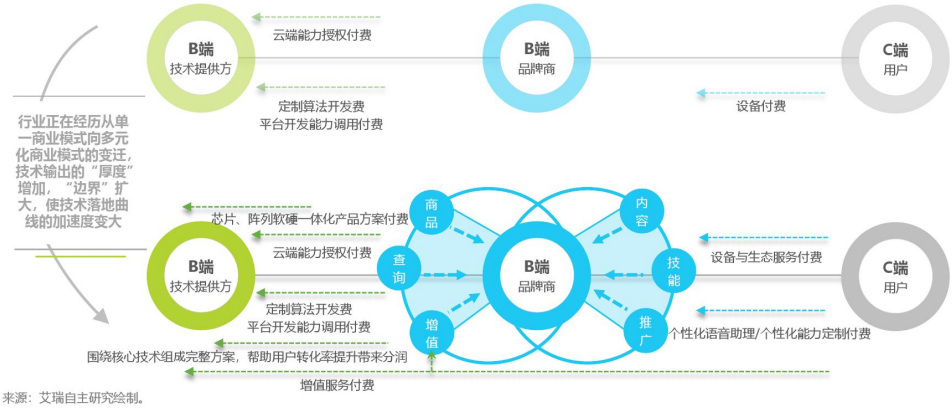
来源：艾瑞根据公开资料自主研究绘制。

本章小结

复盘消费级市场：商业模式多元化与技术落地曲线的加速度

在智能音箱部分我们探讨了品牌设备商如何形成多元化的变现模式，对于消费级市场另一大主力参与者——语音交互技术提供方而言，发展空间也远远不止下游B端品牌设备商在设备开发过程中支付的技术付费。一方面，技术提供方可以通过提供芯片、麦克风阵列解决方案、AI算法的全链方案，增加技术输出的“厚度”，同时实现技术与解决方案的研发中基础环节与模块标准化，降低客户的开发配置门槛；另一方面，强化对应用场景的理解，打磨交互功能和用户体验，给实际问题提供“向前一步”的解决能力，从而获得C端收费的可能。这两类发展空间的实现有赖于两点基础要素：（1）具备全链条语音交互技术能力；（2）有建立用户联系、获取用户体验反馈的场景。

智能语音技术商业模式的多元化变迁



来源：艾瑞自主研究绘制。

子研究（2/3）

企业级与公共级市场

市场画像

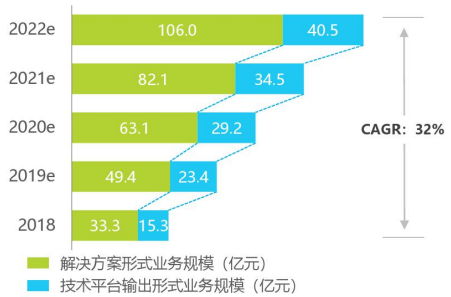
应用场景

企业级与公共级市场画像

平台化技术输出和解决方案两类模式，解决方案业务占比高

智能语音消费者业务主要通过硬件出售及相关互联网增值服务获利，而企业级和公共级业务则主要有两类合作模式：一是技术平台输出模式，将通用技术能力封装为SDK或API，下游客户或生态中的开发者使用时向技术提供方支付一定费用，当然为了促进生态的快速发展，一些平台如华为HiAI、百度语音技术采取面向开发者免费的策略；二是切入传统行业，提供解决方案（含核心设备），这种情形下涉及智能语音企业与传统行业集成商或最终客户进行定制化、深度合作。

2018-2022年中国智能语音企业级和公共级市场规模

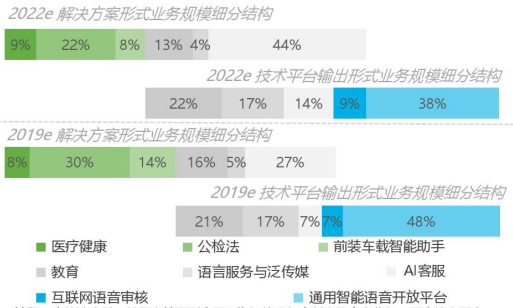


注释：（1）统计口径：未统计金融、社保声纹识别应用和为智能设备定制产品方案业务。解决方案业务指以项目制交付软硬件产品和服务，其中设备仅包括核心产品如翻译机、专用麦克风、专门服务于语音识别与转写的服务器等，不包括同一采购项目中其他终端（如监控设备、电脑）、各类其他服务器与存储设备、安装服务。技术平台输出形式业务指通用型、直接调用的服务，不局限于公有云形态。

（2）统计方法：采用细分垂直领域市场当年释放的需求和主要玩家细分子项业务收入两种方法，具体细分见右图。请读者务必注意数据口径，尤其在引用数据进行二次计算时。

来源：根据基础数据（国家统计局、卫健委、最高法公开数据，公开采购信息，科大讯飞及垂直行业上市公司年报，其他公开资料），结合专家访谈，艾瑞自主搭建模型核算。

2019&2022年中国智能语音企业级和公共级市场细分结构



注释：未统计金融、社保声纹识别应用和为智能设备定制产品方案业务。医疗健康不含家用慢病管理音箱/机器人等设备；教育含在线教育用户付费中应核算为口语评测功能贡献的部分、新高/中考口语人机对话考试系统，不含在线口语教育的整体销售收入、不含智慧课堂、学情分析、教育云、校园多媒体；AI客服包括电话质检和客服机器人，全部客服业务均归属该子项，而不统计入各垂直领域；通用智能语音开放平台指面向开发者提供的通用非定制开发技能服务，已计入垂直领域的开放服务不再重复统计。

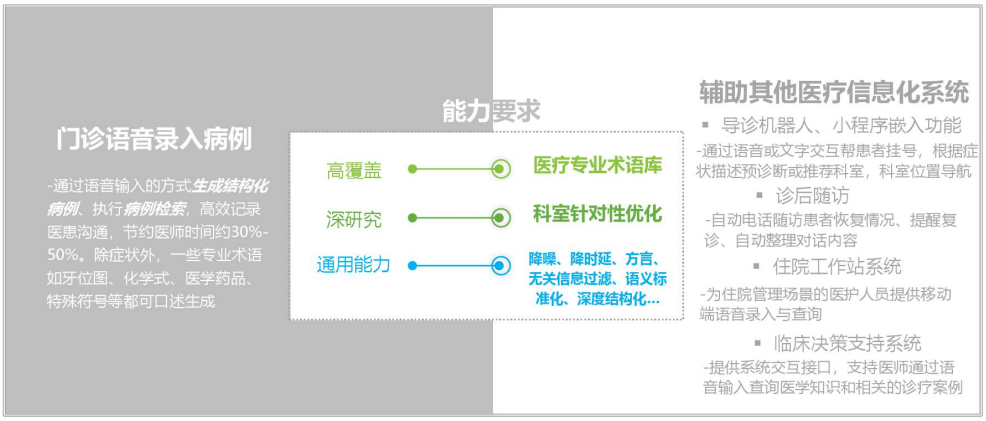
来源：根据基础数据（国家统计局、卫健委、最高法公开数据，公开采购信息，科大讯飞及垂直行业相关上市公司年报及其他公开资料），结合专家访谈，艾瑞自主搭建模型核算。

智能语音与医疗健康（1/2）

核心价值在于提升输入效率和查询效率

医疗领域对于智能语音的需求主要来自电子病历系统上的语音功能，通过语音输入的方式生成结构化病例、执行病例检索，节约医师输入病历的时间，解决方案一般包括ASR/NLU技术和专用医疗麦克风。在导诊机器人、问诊小程序、诊后随访系统、住院病房管理系统、临床决策支持系统（CDSS）中也有应用。在落地过程中，需要重视针对医疗专业术语和各科室专有名词/符号/用药等知识进行模型训练和优化，建立筛选机制以过滤问诊无关信息，并进一步增强病例整理的语义标准化与深度结构化能力，以使系统便捷提取病例主症状、伴随症状、用药等重要特征信息。

智能语音在医疗健康领域的主要应用



来源：艾瑞根据公开资料自主研究绘制。

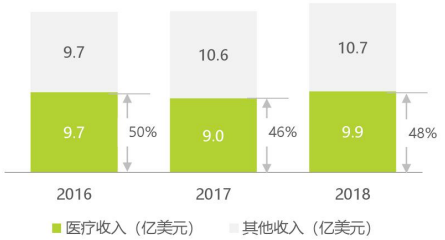
智能语音与医疗健康（2/2）

发展速度受限于我国医疗信息化建设现状

Nuance是全球最大的智能语音公司，2018年其在医疗业务上取得9.9亿美元收入，占公司总收入的48%。相较而言，我国智能语音市场中2018年医疗健康仅占0.7%。这主要是由于美国医疗机构以私立为主，对诊疗服务人性化、医疗信息化关注度更高；我国医疗信息化发展水平相对落后，三级以下医院信息化建设经费有限、专项政策引导力度有待提升、数据孤岛普遍存在，因此目前市场处于单点式推进状态，短期内推进速度比较平稳。不过，智能临床决策支持系统和电子病历语音录入等应用与医疗信息系统打通集成、分级诊疗、医保控费、民生建设等都有直接关系，若相关政策引导加强、医疗数据标准建立和医疗数据跨机构整合推动加速，则有望复制海外市场的医疗业务体量。按照现状估计，预计到2022年，我国电子病历语音输入累计覆盖近1600家三级与二级医院（付费数，渗透率分别为36%和4.5%），180万医生受益。

2016-2018年美国智能语音巨头Nuance

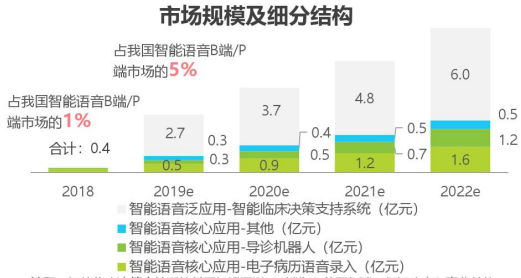
营业收入结构



注释：Nuance的医疗业务起家于为临床专业人士提供语音导航文件系统和应用程序，目前还包括临床文档改良（CDI）、临床语音识别、智能影像诊断、实时听写、计算机辅助编码、医疗质量质量把控、移动云计算、放射科精准报告等业务。
来源：艾瑞根据动脉网公开资料整理绘制。

2018-2022年中国智能语音在医疗健康领域

市场规模及细分结构



注释：智能临床决策支持系统基于知识图谱；P端指公共服务端，包括政府和事业单位。若读者希望进一步了解医疗影像辅助诊断AI可阅读艾瑞《2019年中国人工智能产业研究报告》医疗健康部分。
来源：智能临床决策支持系统以各主要参与公司的公开披露信息、卫健委公布的电子病历系统功能应用水平分级评价高级别医院情况为基础，结合专家访谈，艾瑞自主搭建模型推算。电子病历语音录入以卫健委和CHIMA关于EMR的统计数据为基础，结合专家访谈，艾瑞自主搭建模型推算。导诊机器人由艾瑞根据公开中标信息推算。

智能语音与公检法

帮助公检法系统实现便捷办公和战法突破

智能语音在公检法领域的主要应用

智能庭审

采用多语种多方言语音识别、语音合成等技术，结合针对法律业务的专门优化，实现庭审纪律自动播报、**庭审笔录自动生成**、庭审笔录音频即时回听及快速检索等功能。艾瑞预计到2022年，全国累计有**70%**的法院将上线智能庭审系统；发展方向上，使机器在长文字转写中提取关键信息、做结构化整理应得到更多重视

电信网络反欺诈

电信诈骗指编造虚假信息，设置骗局，大部分电信诈骗是通过电话进行的，声纹识别电信网络反欺诈系统会**自动提取声纹并与黑名单做比对**，提示重点人员可疑行为，对语音内容关键词识别动态预警，提示可疑案件与犯罪意图

虚拟法官

通过语音合成和虚拟形象生成，在互联网诉讼平台上，以**虚拟AI形象**同当事人进行初步沟通，协助真人法官完成线上诉讼接待等重复性的基础工作

声纹研判战法

声纹鉴定与语音分析系统多由市级及以上公安单位建设，基层办案单位则会配备声纹采集设备用于刑侦办案。声纹鉴定与语音分析系统能够进行**语音片段检索、语音自动检测分离和声纹模拟画像**，可协助鉴定人员自动在音频数据中检索出特定的音素及相似的音素、进行语音自动比对，缩小侦查范围，快速确定犯罪嫌疑人身份

智能接警

- 窗口报警自助录入：系统**转写报警人叙述的警情信息**，并导入笔录系统，记录警情处理结果，提高接警效率
- 电话警情自动记录：系统**智能提取**对话中出现的报警人姓名、联系方式、案件类别等要素信息，生成警情记录表单，存储对话内容
- 警情数据汇总分析服务

警务语音语言服务

针对公安领域专业词汇做专门优化，提供**警用语音输入法**和机器翻译等服务

来源：艾瑞根据科大讯飞、搜狗科技等企业官网，及其他公开资料自主研究绘制。

智能语音与教育

应用于教、管、测、考等环节

智能教育领域，AI课堂的建设进入快车道，强调两点：一是解决家校之间、线上线下之间学习资源互通的问题，二是通过多模态识别收集课堂学情信息并做数据精准分析，因此通过语音转录、语音识别等技术实现授课语音转录为文字、利用多模态识别进行课堂质量监测不可或缺。另一方面，在线教育竞争呈白热化态势，用技术解决教育资源的复用、增加学习交互体验感等诉求也促进了智能语音技术在线上口语测评、虚拟教师等领域的应用。考试赛道方面，北京、上海、江苏、广东等省市近年推行在新中考、新高考英语考试中以机考形式进行口语测试，因此人机对话技术和智能语音评测技术开始应用于考试场景，以提升口语考试的效率。

智能语音在教育领域的主要应用

01

语音转录丰富教学模式

通过语音识别实时转写教师讲课的语音为文字，可在授课视频嵌入字幕，并进行关键词和知识点的快速定位，应用于直播课、小班课、互动课堂。

02

语音算法助力课堂质量监测

利用静音检测、语速检测，结合计算机视觉等多模态算法，自动化监测上课互动情况和教学质量。

03

虚拟教师互动教学

通过语音合成+VR技术，可以打造虚拟的名师形象，通过亲切的语音、动作、文字等方式与学生互动。

04

口语测评

涵盖中文（普通话、古诗词）测评和英文测评，可对语音的完整性、韵律节奏及语义、语法进行评测等综合打分，有些产品涉及发音纠正功能，中文测评还可覆盖轻音、儿化音等汉语语音特征，可用于日常口语学习及新中/高考口语机考。口语机考方面，截至2019年年底约已累计建设**150万**个机位。

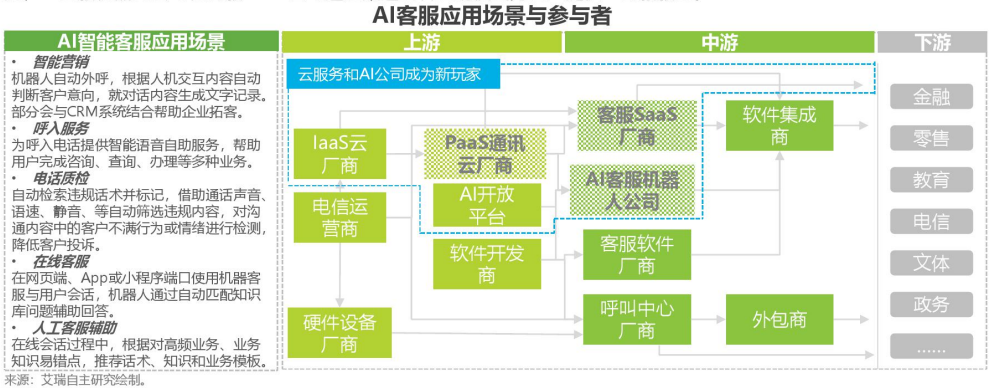
来源：艾瑞根据公开资料自主研究绘制。

智能语音与客服

从人机对话辐射到营销管理和通话服务质检

相对于前文所述的医疗健康、公检法、教育领域，客服领域的行业开放性相对较高，对AI应用迫切性强，参与者众多，未来一段时期内业务体量较大。目前AI客服可以为IVR、APP、小程序、网页等各端口提供自动对话功能，应用场景包括智能营销、呼入服务应答、电话质检、在线客服及辅助人工服务，在一定程度上满足了减轻传统客服中心一线人员工作负担、减少用户等待应答、低成本增加企业营销曝光等需求，应用渗透率较高。但目前AI客服营销转化率低、呼入服务应答转人工率高、业务场景适应性对话系统的建设成本与效果性价比较低、真实场景中对话异常处理灵活度不够等问题依然是行业痛点。传统客服产业由客服软件开发商、呼叫中心厂商、硬件设备厂商、电信运营商和软件集成商组成，AI客服则涉及多种类型的企业：近年来通讯云厂商一定程度上取代了传统呼叫中心，其呼叫中心和云客服业务可以集成提供客服机器人能力，AI客服机器人公司和客服SaaS也可通过渠道或者直销模式为客户提供AI客服服务。

AI客服应用场景与参与者



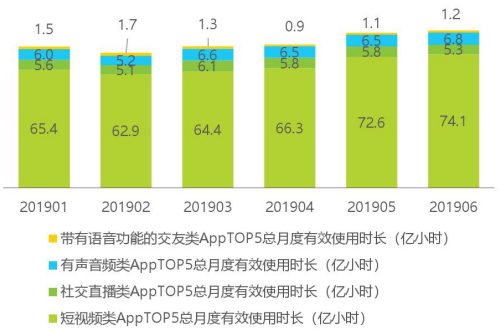
来源：艾瑞自主研究绘制。

智能语音与互联网内容审核监管

特定声音检测和语音识别技术帮助净化网络环境

互联网的变迁使音视频内容的传播得以快速发展，经统计，我国部分头部娱乐社交类App月度总有效使用时长达到每月70亿-90亿小时，用户每天将从娱乐社交平台接触大量信息。这也带来了垃圾信息审核难题，2019年上半年，多款语音社交App因通过语音传播违规信息下架。粗略估计娱乐社交类App背后产生的音视频内容时长接近每月4700万小时，这一数字如果用年来表示，相当于5400年，显然通过人工审核音视频的方式净化网络环境是不可能实现的，而如果依靠举报再人工审核的方式也只如沧海一粟、大量问题语音将被漏查。目前除使用图像识别技术审核图片和视频帧外，以依图科技为代表的AI公司开始通过特定声音检测和语音识别技术赋能实时语音流及音频文件的内容审核，弥补之前针对互联网语音内容的审核空白，提高审核效率与准确度。

2019H1中国部分娱乐社交头部App月度使用时长



注释：排名维度是月度有效使用时长。
来源：根据艾瑞 UserTracker监测数据处理绘制。

互联网音视频中的AI语音审核应用场景

01 **特定违规声音检测**
识别声音特征，拦截喊麦、娇喘、呻吟、ASMR等违法违规音频。

02 **违规语音内容检测**
识别语音内容，过滤与拦截涉黄、辱骂、恐怖主义、违禁内容等违法违规音频。

03 **垃圾广告检测**
识别利用微信号、手机号、QQ等开展的违法垃圾广告内容并进行相应拦截。

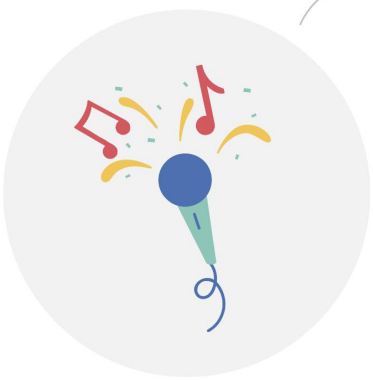
来源：艾瑞根据公开资料自主研究绘制。

智能语音与泛传媒

增加媒体产能，丰富传播形式

智能语音在泛传媒领域的应用主要包括合成主播自动播发稿件，将外语音视频新闻或节目自动翻译、根据画面同步匹配字幕，及为新闻稿件文字自动合成语音等。以自动播发稿件为例，2019年全国两会期间，新华社AI合成主播共播发稿件236条，为资讯内容的生产提供了新的方式；而音频与文字之间的转换则丰富了媒体的传播形式，使用户能够按需、按喜好获取资讯服务。

智能语音在泛传媒领域的应用场景及应用的AI技术



- 合成主播**自动播发稿件**
  - 语音合成
  - 三维人脸重建
  - 视频合成
- 为音视频**自动翻译匹配字幕**
  - 神经网络机器翻译
  - 语音识别
  - 时间轴自动匹配
- 为新闻稿件提供**自动合成的语音**
  - 语音合成

来源：艾瑞根据公开资料自主研究绘制。

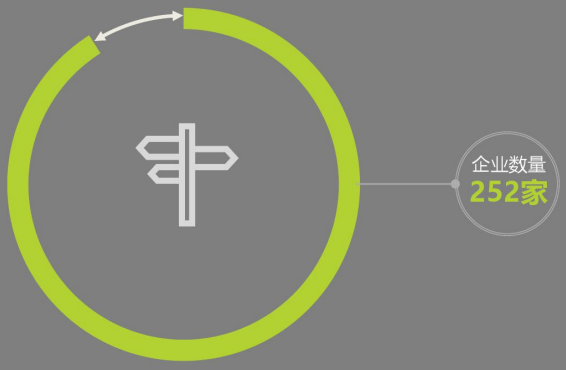
子研究（3/3）

市场参与者

行业图谱

行业热度

典型企业



中国智能语音行业图谱

2019年中国智能语音行业图谱

来源：艾瑞根据公开资料自主研究绘制。以企业主营业务为准，图标排序主要考虑排布整齐，无实际含义。

中国智能语音行业热度

市场较为理性，入局企业数量252家

据统计，我国人工智能创业项目中处于语音识别和语义分析赛道的共有252家，占比10.6%。同时，根据国家工业信息安全发展研究中心数据，截至2018年底，我国人工智能领域合计申请专利44.4万件，而语音识别与自然语言处理技术则合计申请专利6.1万件，占比达到13.6%，反映出智能语音领域单位技术产出情况高于行业整体，且发展也更倚重技术要素。

中国人工智能投资数据概览

	创业项目数量	B轮后项目数量	投资事件数量	投资机构数量
人工智能整体	2,371	751	3,600	1,822
语音识别赛道	180	23	228	252
语义分析赛道	72	11	104	140

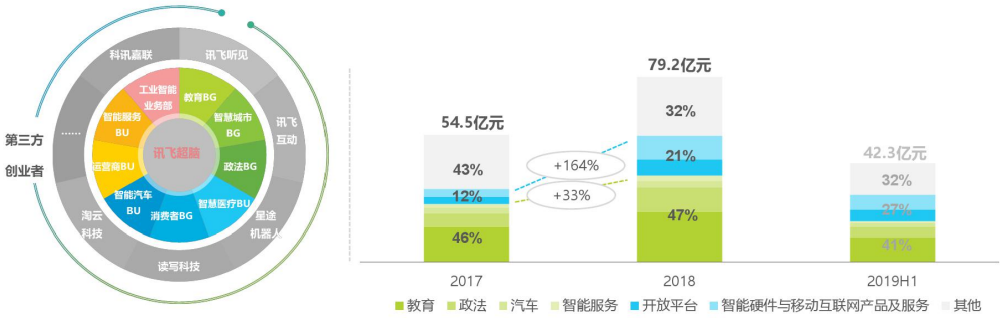
注释：数据截至2019年12月。数据包含已上市企业和巨头企业旗下品牌。
来源：艾瑞根据炜牛数据整理绘制。

代表性企业案例——科大讯飞

以AI解决社会刚需，三个同心圆造就行业头部玩家

科大讯飞股份有限公司成立于1999年，是亚太地区知名的智能语音和人工智能上市企业。目前科大讯飞的人工智能产业生态已经形成三个同心圆：第一层是核心层。围绕“讯飞超脑”，科大讯飞的教育BG、智慧城市BG、消费者BG、智慧政法BG、智慧医疗BU、智能服务BU、智能汽车BU、运营商BU、工业智能业务部等共同构筑科大讯飞人工智能产业生态的核心层。第二层是探索层。在探索性方向，科大讯飞鼓励内部实施创业机制和战略合作机制，通过资本纽带的形式推动人工智能产业化。第三层是开发层。围绕人工智能核心开发平台，科大讯飞为创新创业者提供技术和数据支持，帮助创新创业者在各应用领域进行业务创新探索，将自身源头核心技术提供给平台伙伴，推动整个产业生态构建，截至2019年12月31日，讯飞开放平台已聚集超过 112W开发者团队，总应用数超过73W，累计覆盖终端用户数 26.3亿+，A.I.大学学员总量达到33.3W+，以科大讯飞为中心的人工智能产业生态持续构建。

科大讯飞人工智能产业生态与近期业务结构



注释：“其他”指智慧城市行业应用、信息工程、电信增值产品运营、运营商大数据及其他业务。
来源：艾瑞根据科大讯飞年报、半年报、2019年开发者大会公开披露数据整理。

代表性企业案例——搜狗

为语言理解而生：让AI使人机交互更简单

搜狗成立于2003年，是中国搜索行业的挑战者，AI领域的创新者。搜狗CEO王小川认为，随着AI的发展和应用，搜索和输入法的未来将走向自动问答，从而形成前台的自然交互与后台的知识计算相结合的人工智能结构，搜狗是为语言理解而生的公司，在AI的探索上将以语言为核心。基于“让AI使人机交互更简单”的追求，2012年搜狗输入法和地图上线语音输入功能，2016年上线以语音交互技术为核心的知音人工智能平台，推出面向智能设备的“知音OS”，同年在第三届世界互联网大会上，搜狗第一次把已有的语音技术和基于神经网络的实时机器翻译技术结合在一起，进行现场AI同传，至今AI同传已服务数百场会议；2017年，推出语音实时变文字的速记工具“搜狗听写”，帮助用户实现高效记录和信息输入等服务；2018年，智能硬件翻译宝与翻译笔推出，为用户的出行带来便利，也为搜狗带来了有用户反馈闭环的场景，以便进一步打磨技术，2019年搜狗录音笔C1上市，首发当日销量突破了2万台，而降噪技术、听感优化、语音转写切分说话人、针对连读/发音模糊等细节的优化也在不断完善；同时，AI合成主播也于2019年迭代升级，基于AI分身技术的突破实现站播，姿态和动作更自然。

搜狗智能语音探索历程

来源：艾瑞根据公开资料研究绘制。

代表性企业案例——搜狗

互联网产品、智能硬件和知音平台相辅相成

基于输入法用户大数据的沉淀与积累，搜狗提升输入法与听写服务中针对用户个性化特色词句的识别准确率，提升用户日常生活中表达、传递信息的效率；另一方面搜狗注重技术打磨，其表征学习能力可通过小数据快速定制合成用户个性化语音，自研的Smart Voice麦克风阵列算法则能对噪声和混响进行多重深度优化，确保人声的高保真还原，对技术细节优化的关注使搜狗得以打造出明星产品；而智能硬件又为搜狗带来了向最终用户输出服务的通道，不仅组成了商业闭环，也形成了获取用户体验反馈、进一步打磨技术的服务闭环。

搜狗智能语音发展优势与业务矩阵

来源：艾瑞根据公开资料自主研究绘制。

代表性企业案例——思必驰

聚焦智能终端和企业智能对话服务，围绕“云+芯”重点布局

思必驰是国内专业的对话式人工智能平台公司，创立于2007年，专注于为企业和开发者提供自然语言交互解决方案，2018年公司入选国家发改委“互联网+”重大工程和人工智能创新发展工程项目。思必驰于2014年起确定专注赋能终端的业务方向，在智能车载与家用智能终端方面业务高速发展，并于2017年推出面向大规模个性化定制需求的全链路智能对话定制平台DUI；2018年切入企业服务市场，发布针对企业智慧服务的解决方案会话精灵，为企业提供智能客服和知识机器人等服务；2019年，思必驰携手中芯国际推出首款毫秒级语音AI专用芯片TH1520，打通智能语音交互全链路的最后一环，从芯片层推进语音语言技术的应用规模化，同时较上一年实现了近3倍业绩增长。

思必驰发展重心

01

业务 – **智能终端与企业信息服务两大领域**

- **智能终端**
  - 针对车载、白电、音箱、机器人、耳机等智能终端，提供**一站式产品方案**，除对话操作系统，还覆盖麦克风阵列、软硬一体化解决方案、芯片及整机定制方案，方案内置全链路智能对话技术和丰富的AI技能，并配以海量**第三方资源库**和专业用户**运维平台服务**；
  - 树立了包括天猫精灵、美的空调、海信电视、OPPO手机、长城哈弗、小鹏汽车、荣威等诸多案例。
- **企业智能信息服务**
  - 搭建面向全渠道的企业信息服务平台“会话精灵”，提供BOT定制能力，可用于呼叫中心、虚拟机器人、智能会议设备等，以**启发式对话**和**复杂知识结构管理**为特色
  - 已服务于顺丰、重庆农村商业银行、碧桂园等。

02

思路 –**云+芯带来全链路与大规模定制化服务能力**

- **DUI开放平台**
  - 一站式对话定制开发平台，提供**从信号处理到对话管理**等全链路对话技术与开发件。平台注重为开发者提供超高度定制服务，除通用的唤醒词、合成音的定制外，还支持每一轮交互逻辑、知识图谱、产品技能、图形用户界面的**深度定制**。
- **AI语音芯片**
  - 推出聚焦于语音应用场景下的AI专用芯片，主要面向智能家居、智能终端、车载、手机、可穿戴设备等各类终端设备。解决方案包含算法+芯片，具有**完整语音交互**功能，可实现语音处理、语音识别、语音播报等功能，支持离线语音交互。

思必驰核心技术优势 – *拥有近400语音语言研究人员，各项知识产权800余项，专利514项*

来源：艾瑞根据公开资料自主研究绘制。

代表性企业案例——思必驰

专注人性化的智能语音交互技术

思必驰不仅关注语音控制，更关注人机对话式交互，多轮交互、打断纠错等技术业内领先，为产品提供专业深化的场景解决方案，为企业提供启发式对话的智能服务，同时开放DUI全链路智能对话定制平台，推进语音语言技术的应用规模化。

思必驰智能交互技术

来源：艾瑞根据公开资料自主研究绘制。

代表性企业案例——依图科技

由语音识别、NLP、声纹识别三架算法马车拉动的“跨界者”

依图科技创立于2012年，是人工智能独角兽公司，担纲科技部“视觉计算国家新一代人工智能开放创新平台”。依图的愿景是参与人工智能领域的基础性科学研究，全面解决机器看、听、理解的根本问题，建设更加安全、健康、便利的世界，因此，除已取得重大成就的计算机视觉领域之外，依图在语音识别与自然语言理解领域也厚积薄发：

- 2018年底公司首次对外公布语音识别能力，甫一公布便取得亮眼成绩，语音识别算法大幅刷新全球最大开源中文数据库AISHELL-2上的字错率记录，字错率仅3.71%，比过去领先水平还进一步提升约20%；
- 2019年依图NLP成果荣登《Nature Medicine》，这是该期刊全球首次刊发中文NLP在临床智能诊断的研究成果；
- 2019年荣获国际权威声纹识别竞赛VoxSRC冠军，并首次将等错误率（EER）记录刷新至1%以内；
- 2019年，在由网信办、工信部及公安部三部委指导的首届中国人工智能多媒体信息识别竞赛中，依图于11个任务中斩获10个A级，为所有参赛者中最多，其中包含语种相关关键词和声纹识别两项。

在语音识别、NLP、声纹识别等技术的基础上，依图结合自身对企业级和公共级市场的服务经验，已将智能语音相关技术与多应用场景结合，包括语音内容审核、智能会议系统、语音开放平台等。

依图科技智能语音应用领域（部分）

来源：艾瑞根据公开资料自主研究绘制。

代表性企业案例——依图科技

内容安全领域的一匹黑马

互联网平台的内容监管日益严格，特别是《网络音视频信息服务管理规定》的发布，对服务提供者在内容安全管理上的责任做出了明确要求。然而，相较于成熟的图片审核，直播语音因为背景嘈杂、违规内容变化多端，对机器识别违规内容造成了极大的挑战。在这样的背景下，依图入局，公司语音审核的召回率（查全率）和准确率（查准率）居于行业前列，已成功服务十余个互联网音视频客户。

依图科技语音内容审核服务

7年公共安全服务，行业理解深刻

依图在公共安全领域起家，多年的行业经验和技术积累，使其熟知内容安全监管规范和保障方法，帮助企业抵御违规风险。

全栈自研AI，带来业界领先的识别水平

依图的语音识别、自然语言理解和声纹算法均为自研，且在行业内的各项赛事上屡获殊荣。在实际端到端审核应用中表现一马当先。

实时追踪最新违规态势

依图基于公网进行大数据挖掘，及时发现最新及潜在违规内容，第一时间更新算法模型，解决了违规内容更新不及时的行业痛点。

应用场景

来源：艾瑞根据公开资料自主研究绘制。

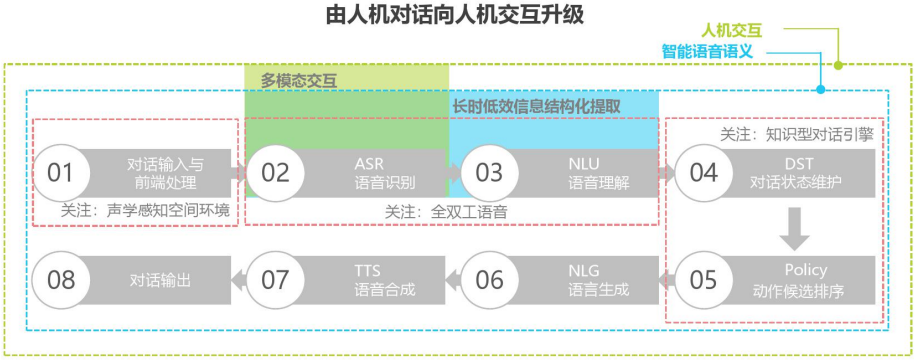
写在最后

AI助理的真正形态：向多模态高密度交互升级

在5G快速发展的背景下，高带宽和低时延特性使多模态识别开始普及，未来支持多模态识别的AI芯片、支持多模态识别的物联网操作系统以及AI算法将受益。多模态识别的主要应用场景包括车载（第三空间）、智能机器人、身份鉴定，具体将会通过语音识别、人脸识别、表情分析、唇动状态、眼球跟踪、手势识别、触觉监控等智能人机交互手段综合识别人的情绪、疲劳状态、复核验证人的身份，对于更加精准、主动和个性化地提供人机交互方式十分重要。

另一方面，语音转写已经成为智能语音技术落地的重要场景，目前在短时间、对话人数少的场景下应用效果较好，但在企业级和公共级场景下往往面临对话时间很长的情况，仅做语音转文字和简单的结构化，不能甄别有效信息、语义结构分类不理想等将是限制语音转写规模化落地的最大问题，行业的高速发展有赖于准确地按照需求提取长时语音消息的有效内容。

由人机对话向人机交互升级



来源：艾瑞自主研究绘制。

写在最后

各类企业行动方向

智能语音行业参与方行动建议

互联网•ICT巨头

以建立人机交互大生态为核心目的，最终将服务覆盖到全产业、全场景、全网用户，基于数据的流转和计算提升业务效率和用户体验，例如通过基础云服务+物理连接+数据分析实现个性推荐、不同情境下的精准产品后服务。这一策略的成本在于前期投入资源理解场景、打磨适应场景的平台化能力，策略的实施效果取决于内部对于*试错的包容性和投入支持的连续性*。

品牌设备商

一方面，设备仍然是触达最终用户的核心端口，在智能语音特别是消费级AIoT领域居于主导地位，然而另一方面大部分AIoT设备领域格局比较分散，用户黏性较弱，设备商必须保持对智能语音与交互技术变革的充分重视，与技术提供方、内容方开诚合作，共同吸引用户*做大设备后服务*，以保持自身在市场竞争和产业链条中的优势。

技术提供方

一方面做好技术与解决方案的研发中基础环节与模块标准化，*降低客户的开发配置门槛*，做强服务消费级AIoT市场的能力，另一方面，关注技术平台架构/服务面向企业级、公共级市场的开发优化，*配置细分市场商务团队*，在企业级、公共级市场爆发前抢滩，占据一席之地。

集成商

企业级、公共级市场集成商的客户资源壁垒很高，短期看将维持发展优势。远期，需要正视行业价值链扩展的必然趋势，挖掘智能语音带来的新应用市场，实现“软” 实力的升级，从注重系统集成、工程建设扩展到*注重IT服务与软件研发能力建设，避免做薄自身价值*、低效消耗客户资源。

来源：艾瑞自主研究绘制。

关于艾瑞

在艾瑞 我们相信数据的力量，专注驱动大数据洞察为企业赋能。

在艾瑞 我们提供专业的数据、信息和咨询服务，让您更容易、更快捷的洞察市场、预见未来。

在艾瑞 我们重视人才培养，Keep Learning，坚信只有专业的团队，才能更好的为您服务。

在艾瑞 我们专注创新和变革，打破行业边界，探索更多可能。

在艾瑞 我们秉承汇聚智慧、成就价值理念为您赋能。

我们是艾瑞，我们致敬匠心 始终坚信“工匠精神，持之以恒”，致力于成为您专属的商业决策智囊。

扫描二维码
读懂全行业

海量的数据 专业的报告

400-026-2099 ask@iresearch.com.cn

法律声明

版权声明

免责条款

本报告中行业数据及相关市场预测主要为公司研究员采用桌面研究、行业访谈、市场调查及其他研究方法，并且结合艾瑞监测产品数据，通过艾瑞统计预测模型估算获得；企业数据主要为访谈获得，仅供参考。本报告中发布的调研数据采用样本调研方法，其数据结果受到样本的影响。由于调研方法及样本的限制，调查资料收集范围的限制，该数据仅代表调研时间和人群的基本状况，仅服务于当前的调研目的，为市场和客户提供基本参考。受研究方法和数据获取资源的限制，本报告只提供给用户作为市场参考资料，本公司对该报告的数据和观点不承担法律责任。

为商业决策赋能

EMPOWER BUSINESS DECISIONS

iResearch

艾 瑞 咨 询