## 语音交互竞品分析

（语音交互的业界发展情况如何？为什么要分成语音交互、麦阵、语音技术分析？以及在商用服务机器人上的应用情况，尽可能的对比技术方案、技术参数等）

### **1.2.1 语音交互**

（为什么对比猎户和擎朗）

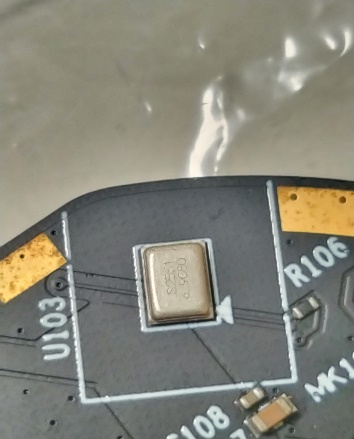
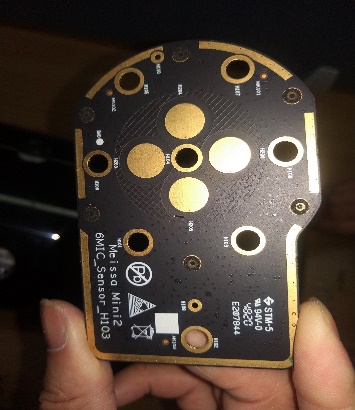
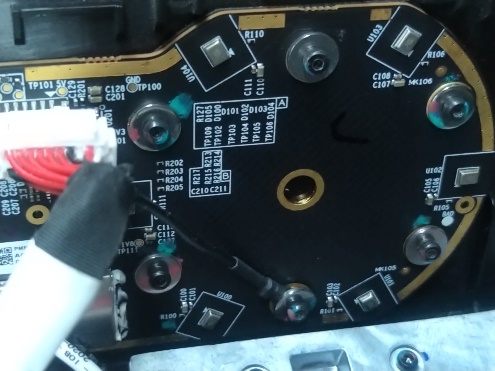
|  |  |  |  |
| --- | --- | --- | --- |
|  | **普渡** | **猎户** | **擎朗** |
| **名称** | **(葫芦/机器猫等)** | **(豹小秘/招财豹等)** | **(花生/T系列)** |
| **语音功能模块**  **（信号处理、语音识别）** | 多语种识别 | 自定义播报话术  5m内360°声源定位  75db噪声环境下高唤醒率  主动迎宾（10m内字幕播报） | 智能手环召唤机器人 |
| **招揽与主动打招呼** | 自定义话术播报  交互营销 | 自定义话术播报  主动上迎 | - |
| **行业问答 自定义问答**  **知识图谱 推荐系统** | 行业场景大数据分析  商家自定义问答和关键词问答 | 行业场景大数据分析  商家自定义问答和关键词问答  用户按结构化模板构建图谱 菜品推荐 | - |
| **技能**  **送餐领位、日历天气、娱乐服务，语音指令等**  **多轮对话** | 语音/UI界面送餐领位  语音日历天气等技能交互 非信息闲聊交互  支持40多种语音指令（内测阶段） | UI界面送餐领位  语音日历天气等技能交互 非信息闲聊交互  支持停止、首页、跳舞等语音指令 | UI界面送餐领位 |
| **多模态交互** | CV人脸-激光人腿-NLP多模态交互  红外感应托盘物品 | CV人脸-NLP多模态交互 | 红外感应托盘物品  (酒店场景) |
| **NLP多语言支持** | 支持中、英两种语言交互 | 支持中、英两种语言交互 | 支持多种语言UI交互 |
| **待改进** | 前端负载大、噪声环境识别效果差 | 不支持语音免唤醒 | - |

趋势：主动迎宾 + 免唤醒 + 多技能 + 多模态 + 多语种支持

### 1.2.2 麦阵

（麦阵的具体含义，以及为什么要拿豹小秘和小花生对比，技术方案部分可以做进一步详细阐述）

|  |  |  |  |
| --- | --- | --- | --- |
|  | **方式** | **猎户星空豹小秘** | **普渡小花生** |
| **1** | **麦阵固定方式** | 全贴合专业声腔设计 | 插座+打胶 |
| **2** | **泡沫棉降噪** | 有 | 无 |
| **3** | **音频输出接口** | PCM | USB |
| **4** | **mic全向/单向** | 未知 | 全向 |
| **5** | **mic 类型** | 硅mic | 驻极体 |
| **6** | **音频处理方式** | 骁龙d845(2.8GHz,8核)  直接处理数字mic | 通过模数转换后送入rk3399(1.8GHz，6核) |
| **7** | **喇叭** | 有高中低音  专业防尘罩  音腔完全独立 | 音腔未封闭 |
| **8** | **麦阵有无阻挡物** | 无 | 头部凹陷，急停按钮等 |
| **9** | **回路使用相同时钟源** | 未知 | 否 |



**策略：**

需考虑如何优化麦克风阵列和本体声学设计

并进行声学评测，提升声学链路整体性能

**计划：**

1. 建立PUDU自主声学实验室，并联合讯飞优化麦克风阵列硬件。

2. 设计新的麦克风阵列模块，解决上诉问题1-9问题。提高拾音SNR到73db

3. 优化PUDU产品流程，从初始EVT阶段，建立声学设计和评测标准流程

### 1.2.2 语音技术分析

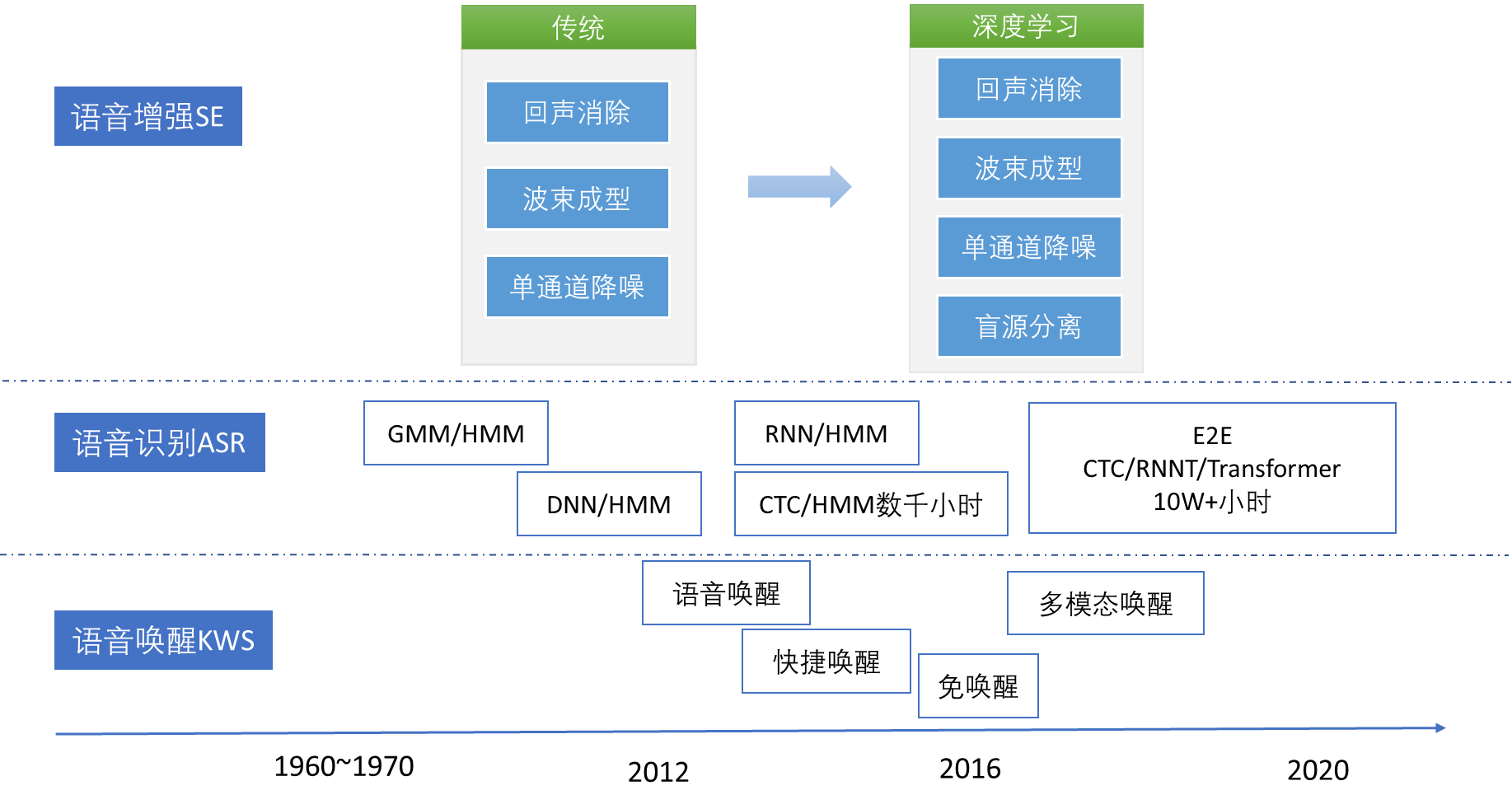
（为什么竞争对手为什么主要看猎户，为什么从语音增强、语音识别、语音链路三方面对比？可以增加技术方案对比）

|  |  |  |
| --- | --- | --- |
| **重点方向** | **竞争对手技术分析（以猎户为主）** | **策略或计划** |
| **语音增强** | 噪音场景识别率高，   1. 麦阵设计优良 2. 自研降噪算法 3. 通过视觉跟踪，辅助定向降噪 | 1. 算力优化，全系统拉通 2. VAD算法改进，提升准确率 3. 自研核心定向降噪算法。 4. 视觉+语音联合唤醒 5. 环形6麦，360°±10°定位，5米收音。回声消除-40db，噪音抑制-25db 6. 人力需求：   回声消除，波束成型，单通道降噪，盲源分离，引擎和集成开发各1人。共5人。 |
| **语音识别** | 数百万小时的6mic远场语音数据积累（官网）  通用单通道语料积累（数万小时）  10W+小时规模训练，20-40卡大规模训练集群（估计）  餐饮真实场景噪音数据（数千小时规模）（估计）  餐饮真实文本预料积累（未知） | 1. 已累集开源数据数千小时，采购和标注线上数据，2022年内希望达到1-2w小时。3年内累积达到8-10w小时（机器人领域积累较慢） 2. 采购和积累一些噪音数据（数量TBD） 3. 搭建大规模机器学习训练集群，提升公司机器学习模型训练能力，计划本年度构建最大24卡3090训练集群 4. 领域识别率达到82%-85%(噪音),93%-96%(安静)。 5. 人力需求：   需2名深度学习语音声学模型专家，负责AM模型，优化。同时负责声学唤醒模型，VAD模型训练。需1名/专家，深度学习语言模型专家，负责LM模型训练、优化。需2名/专家，ASR引擎开发，唤醒引擎开发。1名外包负责数据管理，质检，预处理。共6人 |
| **语音链路** | 1. 自研全链技术栈。  6 Mic阵列硬件 + 麦阵驱动 +  语音增强+语音唤醒+语音识别+NLP+TTS  + PA驱动 + EQ调节 + Speaker硬件  2. 支持中英文  3. 延时（待测试） | 1. 讯飞CAE+SDK+ASR, 声希只有ASR 2. 打造自研语音链路   自研麦阵 + 自研语音增强 +  +语音识别（中英文自研，海外采购）  + 自研NLP  + TTS（先采购，后期自研）  E2E（ASR，NLP，TTS），响应时间<=2s |

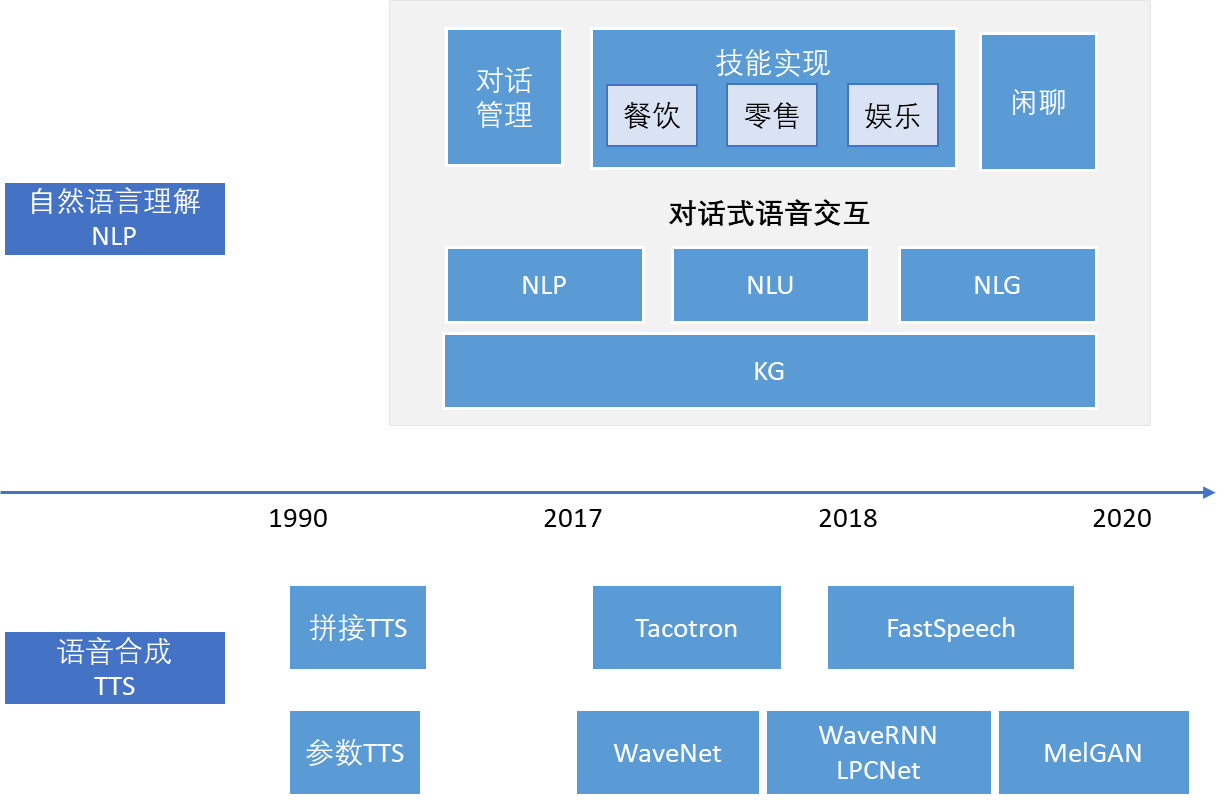
缺少内容总结

## 2.3 人机交互技术趋势

（2-3句话介绍人机交互是什么、在机器人上的应用，简单介绍技术发展）【张智超、程文龙】



* 语音增强是提升远场交互的重难点。传统技术已达天花板，2019年以来深度学习方向是研究热点，并已部分落地；
* 本体声学设计优化+自研场景优化的语音增强是我们提升高噪声场景用户体感的核心手段。目前尚未有相关人员储备。



* 在自然语言理解方面，2018年以来，Bert,GPT3等超大规模预训练模型发布，大幅提升下游任务finetune效果，开源NLP工具包等发布，降低了构建成本。Neo4J等开源知识图谱工具发布，帮助我们构建积累KG提供了基础。后继会重点优化对话管理和垂直领域技能实现。减少无意义闲聊。
* 在语音合成方面，E2E深度声学模型和声码器模型大幅提升了音质和自然度，也降低了数据需求。非自回归的模型演化，试图降低推理成本，提高并发。多情感多音色多发音人的合成技术的热点，基于虚拟形象的数字人是元宇宙的入门技术。

## 人机交互

（增加2-3句人机交互整体规划的语句）【张智超、程文龙】

