**证券代码：688787 证券简称：海天瑞声**

**北京海天瑞声科技股份有限公司**

**投资者关系活动记录表**

**编号：2021-002**

|  |  |
| --- | --- |
| **投资者关系活动类别** | √特定对象调研 □分析师会议□媒体采访 □业绩说明会□新闻发布会 □路演活动√现场参观 √电话会议□其他 （请文字说明其他活动内容） |
| **参与单位名称及人员姓名** | 人寿养老 王雁杰 泓澄投资 张弢、徐哲彤 博时基金 于玥 嘉实基金 何鸣晓国寿安保基金 郑楠浙江国恬私募 卫申鹏东方证券 徐宝龙康曼德资本 丁楹金新城集团 梅阳西南证券 叶泽佑高盛证券：Allen Chang、Verena Jeng、Ting Song、Yuhe Wu、Grace ZhaoVisione Asset：Zhang YimanTorq Capital Management：Tina Xu拾贝投资：Li Yang陽光資產管理(香港)有限公司：Tuo Chen-Sunshine AMCPutnam Asset Management：Yan Li柏駿資本管理：Enchi Deng、Leo Liu路博迈集团：Zhu YunhanJ O Hambro Capital Management：Cho-Yu KooiHel ved Capital：Charlie Chai德意志银行资管：Alan Lau山高國際金融控股有限公司：Kai ZHU兴证国际：Bao Haimiao国泰金融控股：Alan Fu |
| **会议时间** | 2021年9月7日2021年9月8日2021年9月27日 |
| **会议地点** | 海天瑞声卢沟晓月会议室腾讯会议 |
| **上市公司接待人员姓名** | 董事会秘书： 吕思遥证券事务代表： 张哲 |
| **投资者关系活动主要内容介绍** | **公司简介：**北京海天瑞声科技股份有限公司是一家从事训练数据的研发设计、生产及销售业务的人工智能训练数据专业提供商。公司通过设计数据集结构、组织数据采集、对取得的原料数据进行加工，最终形成可供AI算法模型训练使用的专业数据集，通过软件形式向客户交付。公司提供的主要产品和服务包括训练数据定制服务、训练数据产品和训练数据相关的应用服务。自2005年成立以来，海天瑞声始终致力于为AI产业链上的各类机构提供算法模型开发训练所需的专业数据集。海天瑞声所提供的训练数据覆盖智能语音（语音识别、语音合成等）、计算机视觉、自然语言等多个核心领域，全面服务于人机交互、智能驾驶、智慧城市等多种创新应用场景。其中智能语音为海天瑞声的优势领域，产品线已包含160余个主要语种及方言，并凭借稳定的质量、优质的服务赢得了阿里巴巴、腾讯、百度、微软、三星、亚马逊、科大讯飞、商汤科技、中国科学院、清华大学等众多大型优质客户的认可，核心技术、产品资源、优质客户等竞争优势逐步凸显。经过多年发展与积累，公司逐步构建起了较为坚实的竞争壁垒，主要体现在：1、技术能力：在多语种的语音语言学基础研究和高效数据处理技术方面积累下多项具备较强专业性、较高技术壁垒的核心技术，覆盖语音语言学基础研究、多语种多模态训练数据设计、数据同步、大数据驱动的高效数据处理、分布式高性能自动校验等方面。2、资源能力：已沉淀下近千个自有知识产权的数据集产品，拥有多场景、多语言覆盖能力，并在全球范围内建立了丰富的多语种语言学家团队资源及稳定合作的供应商、客户等上下游资源体系。从过去五年的发展看，海天瑞声较好的抓住了训练数据自2016年至今行业第一个五年的发展机遇，在人工智能技术开始应用的情况下，使训练数据发挥了其商业价值，为公司过去五年获得了收入约30%、净利润约70%的良好增长。未来一个时期，随着AI应用端的高速发展，智能语音、计算机视觉和自然语言技术已经越来越多地应用到更多的场景中，在以智能驾驶等为代表的领域拓展方面，公司将适度超前布局。同时，随着国家“一带一路”战略的进一步深入推进和我国企业出海布局增多，将带动多语言训练数据需求迎来新一轮增长，公司也将在现有语种积累基础上进一步拓展、升级多语言资源体系。从中长期来看，垂直行业数据、政府数据治理等领域也将是亟待挖掘、拓展的方向，公司也已经开始了这些方面的探索，并计划将该类研究持续深化，逐步扩展到更多的行业领域，提升公司的数据处理能力，赋能人工智能技术与更多行业实体经济的深度融合。**公司数据集产品和数据生产平台演示：*** 语音识别数据集
* 语音合成数据集
* OCR数据集
* 图像数据集
* TTS标注平台
* 3D点云标注平台

**问题回答：**1. 采集环节的技术门槛反映在什么方面

主要反映在两个方面，一是采集方案的设计，包括采集场景、被采集人的画像、采集平台和采集设备的研发等，这些都决定了采集的生数据是否能够被加工成供机器学习的高质量数据；二是采集资源的充分性，尤其涉及全球采集的时候，对当地法律的理解、当地资源的迅速获取和组织都是采集规模化和合规的门槛，如何在全球各地招募、组织、实施这些采集活动，也是一家公司资源能力、项目管理能力的体现。1. 中报的数据是下滑的，原因是什么？公司对下半年业绩的预期如何

今年总体受海外疫情的严重反复的影响，导致海外客户的需求调整、或海外采集暂停，均对海外收入有不确定性影响。国内业务从疫情中恢复的较好，预期将恢复到增长的态势。1. 定制服务业务开展过程中数据集所有权的归属如何，客户需求对于公司研发的意义如何？

训练数据定制服务涉及的原料数据和加工后的训练数据在交付给客户并完成验收后，所有权完全转移给客户，是不能用于自身产品建设的，这点是公司始终遵循的知识产权要求。在提供训练数据定制服务过程中，公司会接触到各种类型的数据，该等数据会在公司的采集加工平台上进行处理，对公司的算法提升和平台的工程化能力都提供了很好的资源，是研发迭代的较强助力。1. 训练数据定制服务、产品收入的占比结构；销售模式（重复销售、一次性销售）；客户数量、年活客户数量

从过去几年的历史数据看，训练数据定制服务和训练数据产品的收入贡献在6:4左右。定制服务属于一次性销售，产品数据一次开发多次授权销售。公司目前累计合作的客户数量超过500个，每年活跃的客户在20%-30%左右，每个客户可能在某些领域有合作，在合作的深度和合作的广度上都还有很大的空间挖掘。1. 产品数据集的更新迭代是什么样的？

产品库的完善、更新主要针对加工层面会进行一定的改良，部分数据集会做一定的扩建，但考虑到数据集的整体平衡性，单一数据集产品的完善和更新的需求不会太多。从整体产品体系上看，公司会根据不同的维度，例如规模、语言、设备、场景等，结合技术趋势和市场需求，对产品体系进行不断的完善和扩建。1. 业务领域是只有语音吗？是否有消费品公司找公司针对超市/大卖场的货架摆放进行数据采集、处理？

公司的业务范围包括智能语音、计算机视觉、自然语言三大领域，智能语音对收入贡献占比较高。公司进行过针对零售领域的视觉类训练数据定制服务业务。1. 国内、海外的收入占比结构

一般是6:4的比例，2021年受海外疫情反复的影响，国内占比会有提高的趋势。1. 客户自建数据团队对公司的业务来说是否构成冲击

公司认为需求方自建团队的模式是有存在的合理性的，会长期存在，但不会对公司的业务带来明显的负面冲击。主要由于：1.人工智能产业的高速发展使训练数据服务行业的整体容量很大，需求方自身解决部分数据需要，不会对公司的市场需求带来太大影响；2.从AI对训练数据的需求逻辑出发，即使需求方消化了一些需求，也不影响他们继续购买公司生产的数据集，可以充实更多的数据类型；3.需求方自己做数据，是不会分享给其他客户的，只能自用，缺少了分享机制，是无法成为一种有效商业模式的，因此整体空间有限，而公司为所有下游客户提供训练数据，可以借鉴不同客户的需求提升所有数据的广度和质量，是一种正向促进的循环，和更为效率的产业链专业化分工安排。1. 算法开源趋势下，通用算法模型越来越多，对公司产品库的影响？

随着算法开源，算法难度和准入门槛均随之降低，那么其实增加了海天的潜在客户数量，会有更多的中小型公司开始某一专门方向的算法研究，需要相对应的数据采购，同时传统企业也会开展自身的算法研究，这些都会给公司的客群带来变化。同时，应对这样的变化，海天也会更加提升整体数据解决方案的友好性、便利性，满足更多类型客户的训练需求1. 今年到明年的招聘计划？主要新增哪些岗位？

公司的人员扩张主要针对研发团队和技术团队，其中，研发团队主要针对算法研究（CV算法、预识别算法）、工具/平台开发（工程化能力提升、产品化水平提升）；技术团队主要针对产品研发团队人员扩张支撑公司的产品扩建计划。此外，公司可能在销售团队上有所扩张，尤其针对海外市场的商务布局。1. 公司在自动/智能驾驶领域的布局？

在过去1-2年的时间里，AI+驾驶开始迅猛发展，公司也开始在智能驾驶专项领域加大储备，为业务发力奠定基础：在能力储备方面，公司在研发上开始在智能驾驶领域开始了专项投入，包括逐步建立了车载环境采集麦克风阵列、3D点云标注、2D-3D联合标注平台/工具研发团队等，并开始建立专门针对智能驾驶数据集的算法团队；在自有知识产权产品数据集的开发方面，已经建设、积累了一些这方面的数据集产品。从市场需求方面，公司在过去2-3年的时间里，陆续收到包括智能座舱等在内的语音类智能驾驶领域的数据需求，从去年开始，视觉方面的智能驾驶数据处理需求开始较多出现，目前业务当中所覆盖的应用类型涵盖了智能座舱、道路识别、道路规划等。当前，公司已经与国内几大造车新势力厂商、以及一些传统车企建立了合作。未来，公司将深度聚焦智能驾驶领域，加大研发投入，努力将其培养为一个新的主要营收增长点。1. 一般来讲，公司的项目周期有多长？

公司的项目实施周期可以分为以下2个类别：（1）产品类数据库的周期：因为其在销售时已经是成品状态，通常1-2个月能实现收入确认；（2）定制类服务的项目周期：根据项目实际，生产过程可能涵盖设计、采集、处理、质检等环节，平均周期在6个月左右。如果仅为数据加工服务，则根据客户提供的数据量和难度，周期在3-9个月不等。1. 智能驾驶业务的毛利率，跟其他行业的毛利率相比，有什么区别？

考虑到各个应用领域在数据类型、研发投入等因素方面的差异，根据业务实际，公司目前是从主要应用领域的角度将毛利率拆分为智能语音、计算机视觉、自然语言这些方向来进行分析。可以看到，由于产品数据集的贡献，智能语音类的业务毛利率通常高于其他两类业务，未来随着智能驾驶业务的进一步拓展，公司会考虑新增应用领域维度进行专项分析，并据此关注智能驾驶业务的业务变化和盈利情况。由于该类业务较为新颖，在技术和应用角度也未形成较为统一的标准，因此业务上量初期，盈利能力可能随数据集结构的不同、客户的不同或公司已具有的平台、算法能力的不同而有所浮动。但随着技术的完善和应用的落地，标准化的能力将会提升。1. 多语言拓展近期的发展情况如何？跟竞争对手相比如何？

一般来讲，多语言数据资源体系的研发能力建设和资源厚度，须经过多年积累、沉淀。以发音词典为例，行业内并非每家企业都具备研究、开发发音词典的能力；同时，从语言覆盖广度来看，海天瑞声目前已经覆盖了全球160+语种/方言的研究能力，国内主要竞争对手在此方面的积累是在“十/几十”这个量级。但我们也清醒地看到，国际主要竞争对手Appen的语种/方言覆盖能力达到了180+这个量级。但同时，与Appen相比，海天瑞声的成本优势比较明显。在新冠疫情爆发之前，公司国外客户在多语言方面的需求还是比较可观的，疫情开始后国外客户的多语言业务拓展受阻，对公司造成了一定影响，相信随着疫情减缓此方面的需求将会重新释放。在今年海外疫情反复的不利情况下，公司仍然努力进行多地区多语言的拓展，预计今年，公司仍然会在多语种拓展方面增厚积累，包括进行了较大规模的多语种自有产权产品库的开发、与国内部分客户在多语种拓展上形成了新的合作等，为未来奠定增量基础。1. 科大讯飞与海天是竞争对手？还是客户？

公司的客户类型分为大型科技公司、AI技术企业、科研机构。科大讯飞是公司多年以来的优质客户，属于是AI技术企业这个方阵中的典型代表，双方不存在竞争关系。多年来，公司在多语言数据等方面持续为科大讯飞提供数据产品/服务。1. 公司在向垂直行业、政府数据治理拓展过程中，如何确保数据安全、合规？

公司在2019年底引入了由中央网信办直接管理的中国互联网投资基金，通过战略协同，公司参与了国家相关法律法规的制定过程，确保自身在数据安全、合规方面的实践与国家的要求完全一致。此外，从公司初创历程看，由于长期与国际性科技企业合作，对数据安全和合规的重视是深入到公司基因当中的。近年来，公司在数据安全、合规方面持续投入，不断升级信息安全系统，目前已经实现了可追溯至每一个终端授权人的授权流程线上化，以及处理环节的业务数据与个人信息相分离，确保授权、采集、存储、隐私保护等重要方面的安全、合规。同时，行业数据、政府数据的业务需求方大多都要求数据服务企业做到“数据源不出服务器、数据源可用不可见”，公司目前在技术研发、项目能力方面已经能够满足此类要求，并将继续加大这方面投入力度。同时，在政府数据治理的商业模式探索方面，公司也在积极的参与研讨，紧跟趋势。众所周知，数据安全和合规是需要投入较高的成本建设的，在近期日益完善的法律环境下，这方面的投入为公司带来了潜在壁垒，为公司未来在垂直行业和政府业务延展积累较强的壁垒。1. 在人机协作方面，与其他厂商相比，海天领先多少？未来有哪些点可以提升自动化率？

机器并不能代替人类对训练数据进行判断，因此人机协作追求的是机器作为人的“助理”，提高人工处理的效果和效率，降低数据处理成本。在人机协作方面，公司拥有的算法能力和工程化能力及二者之间的结合，决定了协作产生的效果。公司相比于竞争对手最大的优势是拥有足够多和足够多元化的数据，其中产品数据集可以用于公司自身的算法训练，定制数据集在生产过程中也能反哺公司的算法提升和平台优化。同时，标注员与算法之间的合作是非常紧密的，能够实时反馈结果给算法，形成快速高效的算法迭代循环。目前，公司在主要业务方向上均布局了底层算法，能够由算法执行助理标注和质检的工作。由于数据生产流程中，工序较多，人机协作的效果不能一概而论，也并不是机器处理量大就一定代表协作效果好，因为数据质量是一切的前提。公司会在算法和工程化能力上投入较大研发力量，训练更多具有专项适配能力的模型，使算法覆盖率、数据处理速度等都能有较大的提升，得到最佳人机协作效果。1. 关于数据集产品的价格走势，是呈逐年下降？上升？

一般来讲，数据集产品在生产出来后，为了保证数据集自身的平衡程度，通常不会做大规模的更新、拓展，而是在标注类型方面做一些完善或者进行小规模的扩建调整。如果判断了新的市场趋势，公司会研发新的数据集去覆盖。因此，每个数据集都是有生命周期的。在这个周期之内，价格通常会随时间推移而呈现下降的趋势，而在这个过程中，那些竞争力较强的数据集会在相当长的时间保持价格的竞争力，其价格下降趋势会较为缓慢、甚至在一定时间内几乎不下降。因此，价格的调整是根据数据集的前瞻性、稀缺性、市场供需关系等各方面因素决定的。1. 数据集产品的开发计划大概是什么样？

本次IPO募集资金的绝大部分都将用于建设大规模的自有产权数据集产品，因此我们说今年到明年将会是数据集产品开发的大年，也是公司为未来几年的发展奠定一个良好的基础。在这个过程中，通过研判近年来的市场趋势，公司会将开发重点放在几个方向上，如大型数据集、多元化特色数据集（例如，多语种、多音色、多模态等）、专有场景数据集（如智能驾驶）等。 |
| **附件清单（如有）** |  |
| **日期** | 2021年9月30日 |