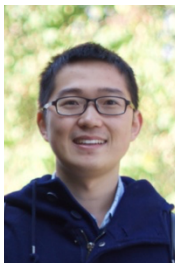
杜志浩 (男，31)

阿里巴巴-通义语音实验室-算法专家

个人主页：https://zhihaodu.github.io　　联系方式：15600609952

邮　　箱：duzhihao.china@gmail.com



个人简介

杜志浩，阿里巴巴通义语音实验室算法专家，负责语音基础算法研究，研究兴趣包括语音生成、音频多模态大模型、语音量化、语音识别以及多说话人语音分析等。在TASLP、AAAI、EMNLP、ICASSP、INTERSPEECH等期刊和会议上发表学术论文20余篇，担任ICASSP、INTERSPEECH、IJCNN等会议审稿人；是CosyVoice 1.0&2.0、FunCodec等工作第一作者、全双工语音多模态大模型MinMo的核心开发人员，其中CosyVoice系列模型落地阿里云、淘宝直播、书旗小说、雷鸟等业务场景，开源模型获得10k+ star，受到业界广泛关注。2015年毕业于内蒙古大学获学士学位，导师张学良教授；后保研直博到哈尔滨工业大学，2021年获得博士学位，导师韩纪庆教授，期间研究课题包括语音增强、语音分离以及语音合成。

项目经历

CosyVoice 1.0 零样本语音生成大模型

【角色定位】第一作者、最核心的开发人员。负责整体方案设计、模型选型、原型实现，带领团队协作方完成语音生成大模型CosyVoice 1.0的大规模数据验证以及最终推动模型上线和开源。

【项目介绍】CosyVoice 1.0是业内首个引入有监督语音token、结合LM和FM的零样本语音生成大模型，实现了阿里内部语音合成大模型从0到1的突破，显著提升了合成语音的韵律自然度、说话人相似度等技术指标。

【业务价值】应用于阿里云TTS合成服务、奥运云小宝、通义APP等云集团内产品，还被夸克、书旗、淘宝直播等团队采用，显著提高了产品的用户体验。

【技术影响力】GitHub Star达到7.3k，技术直播同时在线听众超过2k，被创维、雷鸟、中移动、理想汽车、阶跃星辰等外部公司采用，在SuperCLUE-TTS榜单全球排名第三，在语音领域形成了较强的影响力。

代码库：https://github.com/FunAudioLLM/CosyVoice

技术报告：https://arxiv.org/abs/2407.05407

相关PR：https://mp.weixin.qq.com/s/oO5Gy-MwDNVl_j5Zt9vkyg

https://www.douyin.com/video/7393295198534585610

CosyVoice 2.0 双向流式语音生成大模型

【角色定位】第一作者、最核心的开发人员。负责方案设计、代码实现、大规模数据验证，带领团队协作方完成CosyVoice 2.0的升级迭代，推动模型上线及开源。

【项目介绍】业内首个开源离线/流式一体化语音合成大模型方案，在几乎不损失效果的前提下，实现双向流式合成，首包仅需150ms。通过对LM和FM的优化，相较于1.0合成音频错误率相对降低30%～50%，音色一致性明显提升，MOS分从5.4提升到5.53。

【业务价值】对基于1.0的云产品进行迭代升级，同时对创维、雷鸟等重要客户进行服务升级。

【技术影响力】在语音领域受到广泛关注，开源项目Github Star增长至10k+。

代码库：https://github.com/FunAudioLLM/CosyVoice

技术报告：https://arxiv.org/abs/2412.10117

相关PR：https://mp.weixin.qq.com/s/BgRxjrbnFOhPL99rzf3lYw

https://mp.weixin.qq.com/s/BsRM1OpDzVx__EzDVhpObA

MinMo 多模态语音对话大模型

【角色定位】核心开发人员，负责设计和实现结合文本基模型的无缝流式语音生成方案、大规模数据验证，参与方案选型、模型验证、数据生成以及技术难点讨论等。

【项目介绍】MinMo是阿里首个无缝多模态语音对话大模型，通过丰富多样的数据生成以及多阶段对齐训练框架，在音频理解、语音生成、语音对话等多个方面达到SOTA性能。相较于ASR+LLM+TTS的级联模型，能够感知更多用户输入信息，包括情感、口音、性别、声学事件等，响应延迟低于700ms，具备全双工对话打断能力；相较于端到端原生多模态大模型而言，对文本基模几乎无损，保留更多的任务泛化性。

【业务价值】针对Speech-to-Text Translation任务特化的MinMo模型已落地阿里云、听悟等产品，提供语音到文本的流式转写服务。

FunASR 语音识别开源代码库

【角色定位】主要开发人员，参与早期代码库建设，负责开发相关模型的训练Recipe、推理Pipeline等。

【项目介绍】FunASR是阿里巴巴开源的全栈语音识别解决方案，包括模型训练、推理、导出及部署服务等全套流程。除了语音识别、还包括语音端点检测、文本正则化、逆文本正则化、说话人识别、说话人日志等语音理解和分析的全栈模型，致力于推动语音识别行业基础模型效果。

【业务价值】其旗舰模型Paraformer累计下载超过两千万，开源代码库Github Star达到7.7k，在业内具备广泛的影响力。

其他项目经历

【端到端语音文本多模态模型LauraGPT】第一作者，负责方案设计、早期的实验验证以及技术报告撰写。LauraGPT通过有机结合音频的连续表示和声学离散token，在理解类任务上达到了与专用模型相当或更好的效果，同时原生支持语音流式音频输出，为当前基于无缝连接的语音多模态大模型提供了坚实的研究基础，起到了重要的推动作用。

【FunCodec音频量化开源库】第一作者，负责项目开发、模型训练、实验验证以及最终技术报告撰写和项目开源。FunCodec是一个面向研究的、基础的、可复现的语音量化开源库，其中实现了多个当前流行的语音量化方法，包括SoundStream、EnCodec等。同时提出了基于频域的量化方法，以及语义信息增广的量化方法，在重构质量方面超越了SoundStream、EnCodec、DAC等量化模型。此外，FunCodec的频域量化模型更具扩展性，对音乐、歌唱等音频信号更加鲁棒。

【说话人日志模型SOND和TOLD】SOND是首个提出通过幂集编码将有重叠的说话人日志建模为单标签分类的神经网络模型，在重叠度较高的会议分析场景中超越TSVAD等多标签分类模型，获得了Alimeeting等数据集上的SOTA效果。随后，将该思想引入到端到端说话人日志模型中，提出了重叠敏感的EEND-OLA模型，进而将两者结合提出两阶段说话人日志模型TOLD，在Callhome Benchmark上获得了SOTA效果，为建模说话人重叠提供了新的技术思路。

【AIR实习生项目】先后指导6名实习生在目标说话人语音识别、说话人日志、个性化语音识别、多模态音色控制、语音识别大模型等方向完成基础算法研究，在AAAI、ICASSP、INTERSPEECH等学术会议发表论文9篇、在投论文3篇。