阿 地 力 江 [homepage]

性别: 男
电话: 18699136853
现居城市: 苏州市
E-mail: 1549684025@qq.com
地址:苏州市高新上瑞阁21号
意向: 算法工程师/大模型

教 育 背 景

新疆大学/信息科学与工程学院
2018 - 2021 **信息与通信工程**/硕士
新疆大学/信息科学与工程学院
2014 - 2018 **电子信息工程**/本科

专 业 技 能

编程语言: 熟练使用 Python /Shell/等编程语言; 熟练使用 Pytorch 框架;
系统/平台: 熟练使用 git/GitHub/huggingface; 熟练使用/kaldi/ESPnet/Wenet/fairseq 等工具;
熟练使用 Windows、Linux 操作系统以及常用工具 (sox, vim, sed, awk, grep…);
熟练使用 Photoshop\Office 等办公软件;

工 作 经 历

上海国音智能科技有限公司	算法工程师	**2021/10-2022/12**
苏州君林智能科技有限公司	算法工程师	**2023/2-至今**

项 目 经 历

语音转换模型 **2023/9-至今**

VC 模型：阅读语音转换相关的文献、复现实验，基于 PPG 的模型；该模型都是由多个独立的模型构成，包括 ppg-extractor, speaker-embedding，ppg2mel 和 vocoder; 首先 ppg-extractor 从源音频中提取 ppg 特征，speaker-representing 模型从目标说话人的音频中提取说话人信息并与 ppg 特征拼接送到 ppg2mel 模型生成 mel 倒谱图，最终将已生成的 mel 特征通过 vocoder 生成目标音频(在研究中)。
RVC 模型数据准备：从 b 站收集指定 up 主的所有音频数据(还包括音频时长、点赞评论转发次数、发布时间、标题等)；然后将长音频分别经过 whisperX、语音增强、语音分类模型最终获得指定说话人质量高的音频数据用于模型训练。

语音唤醒模型(KWS) **2023/07-08**

在 MobvoiHotwords 数据集上前后训练，基于 MDTC-MaxpoolingLoss、TCN-CTCLoss 等语音唤醒模型; MobvoiHotwords 数据集包含‘hixiaowen，nihaowenwen’两个关键词和其他负样本；数据集在 MDTC-Maxpooling 模型上训练收敛速度很快，在测试数据集上的真确率 97.8%的高分，但是在训练时容易出现过拟合，从麦克风输入关键词时表现不是很佳，泛化能力差。因此，一是解决以上问题，二十训练出自定义关键词的 KWS 模型，在模型的目标函数上做了一些变化以 maxpooling loss 换成 ctc-loss 训练帧级别的分类模型，具体步骤包括找到目前最好的中文 ASR 模型对数据集打标签,然后建立建模单元(词典 2599 个汉字)，此后训练模型，测试模型性能；由于该模型词典较大，导致训练出来的模型相比 maxploong 的模型大，收敛速度较慢测试结果仅 93%左右。

语音增强模型 **2023/05-07**

ASR 模型虽然在用大量带噪声的数据训练，泛化能力强、健壮性好，但对于语音中出现复杂环境、多说话人、回声等情况，其识别效率大大降低。语音降噪模型同 VAD 模型一样，有助于提升 ASR 模型的识别效率和准确率。因此，输出了以下两款模型：
基于 FRCRN 的模型：该模型是 U-Net 的结构，是由编码器-解码器构成：编码层和解码层以多个 CNN+CFSMN 堆成，其中编码部分利用池化层进行逐渐下采样，解码部分利用反卷积进行逐渐上采样，编码层和解码层由 skip-net、通道 attention 和空间 attention 链接。网络的输入是音频的短时傅里叶值，训练使用 SI-SNR 和 MSE 联合训练，评价指标有 MOS 和 STOI；
基于 CNN+LSTM 的模型：该模型也是 U-Net 的结构，由编码器-解码器构成，编码层和解码层由 skip-net 和 LSTM 链接，网络输入为原始音频，目标函数为 spectral convergence loss 和 magnitude loss 联合训练模型参数。以上两个模型中前者模型复杂度较高但目前是在 SOTA 中排第一，不支持流式。后者模型复杂度不高，输入是原始音频，支持流式但效果不如前者。

基于神经网络的语音端点检测模型 **2023/03-05**

为提高 ASR 模型对长音频的识别效率，输出实时的语音识别模型，研发基于 CNN+LSTM 结构的两款语音端点检测模型；其一是基于 Snicnet+LSTM 的结构，输入是 16k 原始音频输出是每一帧的概率值，用 CELOSS 训练模型的参数，chunk_size 为 8 万个采样点。第二模型是 CNN+LSTM 结构，输入是 MFCC(音频为 16k)特征输出为每一个帧的概率值，用 CELOSS 拟合模型参数，chunk_size 可以选 512，768，1024，2048 等。由于 chunk_size 的不同，输出 Sincnet+lstm 结构的模型为非流式、CNN+LSTM 结构的模型为流式的模型。

端到端语音识别模型以及其序列化 **2023/02-03**

从稳定性、可靠性和可行性出发，调研和统计当前比较主流的端到端语音识别模型，然后在模型结构、模型大小、流式、非流式、编码解码速度、准确率、训练所用的数据规模、语种等方面进行对比。从中选择最适合预期的模型，然后序列化该模型并输出；根据以上步骤，确定 WenetSpeech 模型作为基础模型（该模型用 conformer 模型结构在 1 万多个小时中文语音识别数据上训练在测试数据上获得较低的字符错误率(CER)、泛化能力较强、健壮性好）输出符合业务需求的 ASR 模型，模型支持的解码方式有 greedy search、beam search 和 attention rescore、同时输出 torch script 和 onnx 两种版本。

基于预训练的语音识别技术 **2022/11-12**

了解 wav2vec, wav2vec2.0 的结构，阅读相关的文献；学习基于半监督(预训练)语音识别的技术路线， wav2vec 在 fairseq 中的实现、数据准备、模型训练、调参、结合 ASR 任务微调的过程。

语音识别标点符号恢复 **2022/7-2022/10**

数据处理与词典生成: 用词汇量达70万的语言模型生成词典(word-num), 用bpe算法在6.5GB文本数据上训练bpe模型和词典(subword=1000, subword-num), 此外，以上两个词典中加入，MASK, PAD, SEP, CLS和四个常用的标点符号(period, comma, exclamation, question mark)。
建立基于transformer的分类模型: 分别建立建模单元为word和subword的transformer分类模型，模型以transformer的encoder部分组成，12层网络，输入为词向量和位置向量，class为4(period, comma, exclamation, question mark)。两种模型中subword为建模单元的模型在测试数据集上的F1 score 相对比较好，但还是没到预期其中句号和逗号的recall相对稳定，其余两个类的效果较差。(多线程)
训练基于BERT的预训练模型以及微调(多线程): 由于基于transformer的分类模型没学到语义信息，建立了基于BERT，RoBerta, GPT2 的预训练模型并在它的基础上进行finetune, 得到了较好的标点恢复模型，经过模型裁剪，量化等操作输出给引擎。

基于端到端的语音识别系统 **2021/10-2022/7**

建立测试集_指定方案: 为推进维语ASR，需建立验证模型性能的测试集，根据现有的数据资源以及我们所具备的条件，以场景，语音内容，环境，口音/方言等要素作为建立指标，给出了详细的方案。
数据预处理: 音频数据处理和文本数据清洗脚本，其功能包括音频数据的切割(vad)，合并，文本数据的转换，文本中数字的转换，长句转短句，清理噪点等。
数据生成: 完成调用谷歌翻译接口生成维语文本的脚本，生成7.5万条txt文件(总数据大小为460MB)。调用公司现有的tts系统分两批共生成了100小时左右的asr训练数据，并加到现有的训练数据里。
建立训练数据集和测试集: 用文本处理脚本共建立了四个维语文本数据集，数据大小为6.5GB。用音频处理脚本共建立了3个测试集，分别是带噪访谈，干净访谈，新闻以及其reference，时长为5小时。
模型建立与优化: 用处理好的文本训练数据（四个文本数据集），前后建立了共15个符合业务需求的语言模型，平滑，插值等优化方式，最终生成了词汇量高达70万的4-gram语言模型，在对应的测试集上模型的困惑度(ppl)在190左右，未登录词(oov)降到最低(对于15个语言模型)，并在ASR解码时生成其对应的TLG。
声学模型建立: 在已有的156小时维语ASR数据上分别加入第一批50个小时tts数据和第二批55小时tts数据，训练char，bpe为建模单元的声学模型(conformer)，用 ctc_greedy search，prefix beam search，attention rescore算法解码，端到端模型解码部分加入n-gram语言模型，对实验结果进行分析，加入编辑距离算法第二次打分，提升效果，将模型转成onnx形式输出给引擎。

基于多任务(MTL)学习的低资源语言语音识别系统 **2020/11-2021/6**

针对维吾尔语、哈萨克语等低资源语言都是黏着语，通过词根、词缀可以产生大量词汇，容易出现 OOV 问题，端到端语音识别模型性能较低进行研究，先收集数据建设数据库，在kaldi, espnet上构建语音识别基线系统(GMM/DNN-HMM, CTC，Attention 以及混合CTC-Attention)，在测试数据上进行对比试验，从实验结果中可以看出来，端到端模型对实验数据的依赖比传统模型较为明显，端到端模型解码时引入语言模型能够得出较好的测试结果。

基于 DNN-HMM 和 RNN 的语音识别系统 **2019/9-2020/11**

熟悉与搭建kaldi语音识别平台，学习shell脚本并对THUGY-20数据集进行数据预处理，对音频进行特征提取，生成解码网络HCLG.fst, 建立基于GMM/DNN-HMM的语音识别系统，训练RNN语言模型并在解码时与N-gram语言模型切换，引入区分性训练(discriminative training)，进行对比实验, 其中最小因素错误率(MPE)为WER获得3.66%的relative decrease.

荣誉奖项及成果

英语能力:已通过英语六级，无障碍听读写英文文献。
个人荣誉: 2014学年荣获新疆大学“三好学生”荣誉称号 ； 2014-2015学年荣获国家励志奖学金 ； 2015 学年第十一届“挑战杯”大学生课外科技作品竞赛中荣获“优秀奖”(队长) ； 2015-2016学年荣获 国家励志奖学金 ； 2015至2016学年荣获新疆大学“三好学生”荣誉称号； 2017年第10届中国大 学生计算机设计大赛软件服务外包竞赛中荣获“三等奖;2018年新疆大学信息科学工程学院优秀毕业 生，保送研究生；2020年自治区研究生奖学金；

文章：基于DNN-HMM和RNN的维吾尔语语音识别(期刊论文，录用)
A. Abudubiyaz, M. Ablimit and A. Hamdulla, "The Acoustical and Language Modeling Issues on Uyghur Speech Recognition," 2020 13th International Conference on Intelligent Computation Technology and Automation (ICICTA), 2020, pp. 366-369, doi: 10.1109/ICICTA51737.2020.00084. (EI检索)