尼日英语数据采集需求

一、需求细节

1. ASR 语料

ASR录音文本语料需求：

* 语料量满足录音有效时长500小时
* 单句语料录制的音频长度控制在15s左右以内
* 语料内容：通用域语料可通过爬虫爬取，语音助手域语料通过人工设计获得，占比为通用域350小时，语音助手领域150小时。指令类这部分语料要求每个意图的说法的单个pattern生成的数据条数不超过50条。语料需要增加常用的**尼日利亚**pidgin English，并且需要覆盖具有当地特色的人名、地名等实体类名词，pidgin采集的地区可以以lagos50%左右，warri/sapele50%左右来，pidgin English主要针对语音助手域的说法采集。语料需覆盖热门高频实体名如排行榜榜首的歌曲名/歌手名、热点新闻等加入录音语料中。
* 语料设计需考虑音素覆盖完全

录音要求：

* 录音对象需要在**尼日利亚**，需覆盖约鲁巴族、豪萨族、伊博族三大种族英语口音，剩余的种族也可占一定比例
* 男女比例3:2，年龄分布在15-60岁，18-35岁为主（80%），成正态分布
* 单人采集时长在1小时左右，采集500人
* 同一句语料不超过2个人录音(VA指令类除外)，VA控制指令主要针对传音列的意图列表设计。单人手机VA部分可以增加未来有可能扩展的功能，一步到位。对于使用频率最高的场景指令（如目前的指令列表），同语料可相应增加录音人数，范围从10人-100人均可。
* 使用手机录制，模拟用户使用语音助手的方式（30cm左右以内的距离）

录音环境：

* 相对安静的正常室内环境。

录音设备：

* 使用安卓手机录制，传音手机优先。

交付形式：

* 对每一条录音数据，音频（16k采样率wav文件）、对应文本文件、说话人信息文件（包括：ID、性别、年龄、地区、种族）

关键时间节点：

* 7月中旬之前完成文本数据准备，开始采集数据，第一周每天检查采集数据，之后每录制完每50小可批量提交验收，七月20日前完成100小时数据交付，8月底完成所有500小时尼日口音英语数据采集交付

通用域为：体育、音乐、视频、闹钟、通讯（电话、短信）、社交（whatsapp,facebook,Instagram，vskit）、新闻（scooper），金融，保险，物流，教育，天气，玩笑，导航，国际节日，手机故障维修类，非洲TOP500 应用名等，通用域也可适量放到两人对话里。

1. NLU 语料

语料要求：

* 以定义的场景、意图、槽点为依据，每种场景、意图、槽点都需要覆盖每种意图的表达方式需要全面、有效(每种意图40+种表达方式，对于不同的意图可根据具体情况调整,相同说法、不同槽点值的表达算一种表达方式)，每种意图的数据尽量均匀，槽点覆盖尽量均匀，数据标注准确，包括意图和槽点的标注，系统命令控制、音、视频领域的数据可多采集
* 实体列表（槽点）需要覆盖的类型包括有歌曲，歌手，专辑，播放列表，歌曲类型，app名，视频名， 地点，时间（日期），人名，数量词（如音量、亮度），商户名等等

NLU语料交付形式（以英文为例）：

* 意图交付样例：

## intent: Music.Play.Certain.Singer.Album 音乐播放意图  
- play [taylor swift's](Entity.Singer) music in the album of [never worn white](Entity.Album)  
- play [taylor swift's](Entity.Singer) album of [never worn white](Entity.Album)  
  
## intent: Navigation.Start 导航意图  
Navigate to [Shanghai](Dst.Location) and pass [Beijing](Mid.Location) in the middle  
  
## intent: Weather.Forecast天气查询意图  
- What's the weather like in [madrid](Location) [this afternoon](Datetime)

[ ]是槽点位置标注，里面是实体名，跟在之后的( )内是实体类型标注

* 另外需交付实体列表

关键时间节点：

七月中旬完成交付

二、验收标准

1. ASR验收与返工
2. 第一步验收语料，语料设计需满足需求所规定的通用域、语音助手域的占比为350小时、150小时的有效时长，豪萨语中音素、音连需全覆盖，并且音素尽量均衡分布，对于覆盖不全、不够均衡的情况需要对文本做调整
3. 发音人的地区、性别、年龄分布成正态分布，需覆盖约鲁巴族、豪萨族、伊博族三大种族英语口音，对于覆盖不全、不均匀的情况会要求补录
4. 录音与文本对应的句准确率达97%以上，由传音随机采样抽检准确率，对准确率不达标的情况会要求返工（每50小时的批次为单元）
5. 时间点要求：ASR采集开始后第一周每天交付，验收定位数据采集中的问题，之后每50小时交付，七月底前交付前200小时数据，8月底完成全部数据交付，每批次数据验收和返工需要在7日内完成
6. NLU验收标准与返工
7. 数据有效性：意图表达准确、槽点标识准确，准确率大于99%，对于表达不准确，槽点标识有问题的情况需要返工修改  
   2）数据多样性：每种意图40+表达方式，常用表达方式100%覆盖，对于没有覆盖完全的需要补录
8. 时间点要求在七月中旬完成所有采集、验收返工工作

时间节点规划（待定）

|  |  |
| --- | --- |
| 尼日英语ASR 500小时 | |
| 2020年7月2日 | 指定指令语料，NLU槽位模板，实体词编写方案制定  编写及运行ASR英语语料爬虫脚本 |
| 2020年7月4日 | 完成ASR英语通用语料爬取、review |
| 2020年7月6日 | 完成指令语料，NLU槽位模板， 实体词编写  完成ASR英语指令语料（含pidgin）review  客户确认语料  正式开始ASR录音 |
| 2020年7月10日 | 第一批50个小时ASR数据交付 |
| 2020年7月20日 | 交付100小时ASR录音 |
| 2020年7月30日 | 交付100小时 ASR录音 |
| 2020年8月10日 | 交付100小时 ASR 录音 |
| 2020年8月20日 | 交付150小时ASR录音 |