豪萨语数据采集需求

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1000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700小时，语音助手领域300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
* 语料书写体系为latin-based boko

录音要求：

* 录音对象需要在**尼日利亚**，口音覆盖豪萨族人口占比较多的主要城市
* 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1000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）
* 需要发音字典

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，8月底完成1000小时交付

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

1. TTS 语料

TTS录音文本语料要求：

* 总量10000句，其中2000句语音助手交互应答领域文本（涵盖传音提供的语音助手固定话术语句数百条），其余为通用域口语化文本。（通用域口语化文本是指各个领域口语表达的文本，如果量不够多的话，可以在确保覆盖高频常用的口语化语料的前提下使用普通通用域文本）
* 语料需要对音节音子的数量、类型、音调、音连以及韵律等进行覆盖。
* 单句文本语料间不得有重复冗余表达的情况，对于已有文本基础上变化不大的不同文本需要筛除。
* 句平均词数为15词
* 语料书写体系为latin-based boko

发音人要求：

1. 音色选定：

发音人音色需要试听样音后确认，并且音色的选取需要产品等多方面参与决定，举例可以以对候选发音人试听投票的形式来选定。

2）发音人录音状态：

要求发音人录音时自然流畅，吐字清晰，发音与文本内容严格一一对应，不得自由发挥，不得有添加语气词或者拖音等等问题，也不能有不自然的长段停顿静音情况。

对选定发音人的录音状态也需要试听决定，需要要求该发音人就自己的情绪、语速、音量、音高等不同录音状态下录制不同的样音， 使用这些样音区分确定最终需要的录音状态，并要求发音人在整个录音期间保持这种录音状态一致性。

录音环境要求：

* 标准录音棚环境，需要符合专业音库录制标准，并保持录制期间设备和环境始终不变，录音环境信噪比不低于35dB。

录制工具：

* 专业录音设备及录音软件。

音频采样格式：

* 无压缩 wav格式，采样率为48kHz, 16bit。

数据交付形式：

* 录音文本文件，文本文件中包含10000条文本数据和对应数据编号。
* 录音文件，和录音文本中一一对应的10000条音频文件，文件以文本中的编号命名，录音首尾静音段保留在0.5秒左右。

关键时间节点：

* 2020年七月中之前旬完成文本验收，刚启动第一周得到每一两天结果试听沟通反馈发音人状态，之后每录制完一千句提交验收，八月底之前完成TTS全部采集采集。

1. 验收标准
2. ASR验收与返工
3. 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700小时、300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
4. 发音人的地区、性别、年龄分布成正态分布，对分布不
5. 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）
6. 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20日前交付前200小时数据，8月底完成全部数据交付，每批数据验收和返工需在七日内完成
7. NLU验收标准与返工
8. 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

3）时间点要求在七月中旬完成所有采集、验收返工工作

1. TTS验收与返工
2. 第一步需要验收录音文本，对于文本的音节音子的数量、类型、音调、音连以及韵律等覆盖情况进行评估，对于冗余文本进行筛除，对于发音覆盖不全以及冗余的情况会要求修改录音文本,文本期望在七月5日前完成验收。
3. 录音刚启动时第一周希望得到每一两天的结果试听，便于沟通反馈录音人状态有没有问题。
4. 之后每录制完一千句分阶段验收。
5. 对于有发音状态不好、有噪音、大段静音、读错文本等问题的音频需要返工重新录制，字准确率需要100%，八月底完成所有采集、验收返工工作。

时间节点规划（待定）

|  |  |
| --- | --- |
| 豪萨语ASR 1000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  编写及运行豪萨语语料爬虫脚本 |
| 2020年7月5日 | 完成豪萨语ASR通用语料爬取、review |
| 2020年7月8日 | 完成指令语料，NLU槽位模板， 实体词编写  完成ASR豪萨语指令语料review  客户确认语料  正式开始ASR录音 |
| 2020年7月10日 | 第一批10个小时ASR数据交付 |
| 2020年7月20日 | 交付190小时ASR录音 |
| 2020年7月30日 | 交付200小时 ASR录音 |
| 2020年8月10日 | 交付200小时 ASR录音 |
| 2020年8月20日 | 交付200小时 ASR录音 |
| 2020年8月30日 | 交付200小时 ASR录音 |

|  |  |
| --- | --- |
| 豪萨语TTS 1000句 | |
| 2020年7月8日 | 语料确认 |
| 2020年7月10日 | 发音人音色试听选择、客户确认  正式录音开始  初期每一两天反馈数据，确认录音状态 |
| 2020年8月25日 | 录音完成 |
| 2020年8月30日 | 质检完成并交付 |