苏州方言标注项目招聘

一、公司介绍：

澳鹏 Appen 拥有 20 多年采集和处理多种数据的经验、包括语音、文本、图像和视频等数据类型。全球 100 多万名经过严格验证的众包资源群体聚集在澳鹏，采集和处理的数据覆盖 180 多种语言。澳鹏为全球的高科技、汽车、消费电子、电子商务和金融服务 等企业提供服务，帮助他们开发、改进基于自然语言理解和机器学习的产品和技术。全球 10 大领军科技企业中的 8 家是澳鹏的合作伙伴。

二、项目说明和系统平台

项目说明： 为持续优化苏州话转写模型识别效果，需按照规范对采集到的自由交谈方言音频进行转写，并依照标准化程序对标注成品进行验收。

系统平台： http://ainote.iflytek.com/open/default#/portal/default。

三、标注文档

数据标注方案

1. 标签规范

标签	解析
不清晰（DEAF）	说话人可能语速较快或者为声音低，口音太重，无法听清。
非语音（NOISE）	非人发出的声音，比如音乐声、配音的旁白等。
混度（OVERLAP）	相同时间段里，多人同时说话。

2. 音频分段

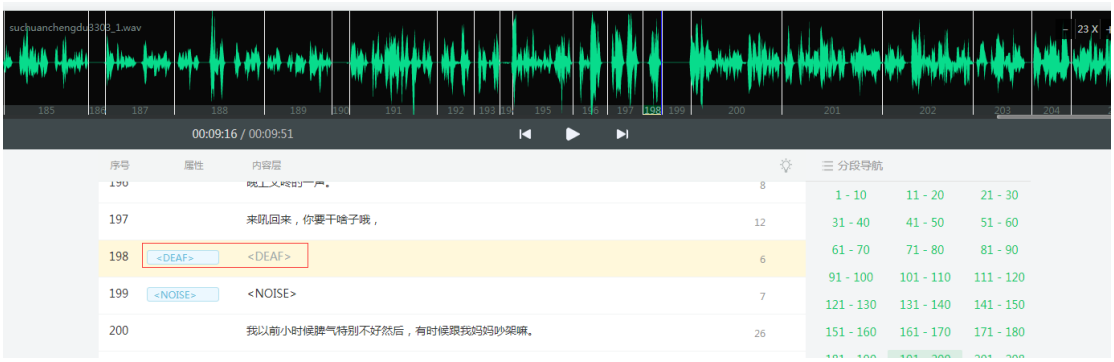
音频情况	操作
大于等于 0.5s 的静噪音段落（即无人说话的段落）； 唱歌、笑、咳嗽声等属于噪音（不是正常说话的声音）	单独切分，自成一段，标记为 noise （注意静噪音的前后边界线不要离语音说话人太远）
不清晰（音频不清晰或者听不懂，不能用文字表示出来）	说话人栏选择“不清晰<DEAF>”
混读（多人同时说话的段落）	1. 两个人同时说话，都听不清，说话人及内容层标注 OVERLAP； 2. 两个人同时说话，有一方说话音量大且清晰，那就标注那一方的语音；

有文字的内容层段落	保证每个分段小于五十字

3. 转写规范

1)基本规则

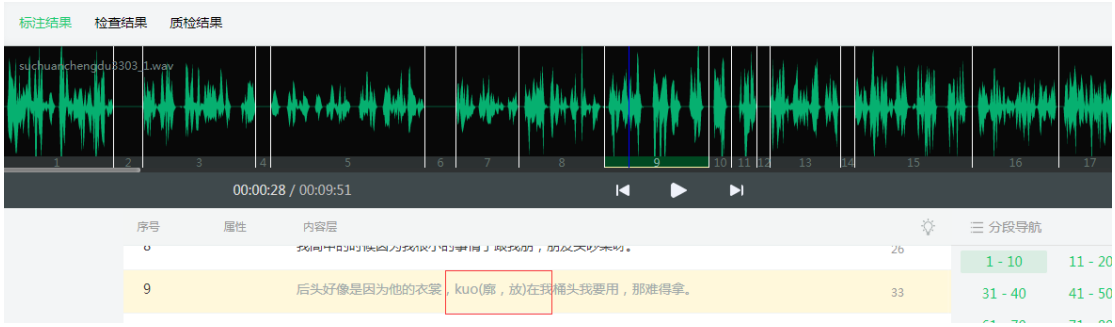
（1）文字基于音频进行转写，语音什么内容转写什么内容，不可多字，漏字，错字；
（2）最终结果要求语音与文字一一对应，切忌不可将方言词句翻译成普通话；
（3）听不懂或听不清的内容请多方确认，如仍然听不懂或听不清楚请单独将该部分语音切出来，请在说话人栏选择“不清晰<DEAF>”；
（4） 咳嗽声、笑声及其他无意义的声音、大于 0.5s 的无声段，请单独切出来并标记为 noise。



2)方言特殊词标注

（1）请参考资源部提供的基础词表，音频中出现的词汇如存在于基础词表之内的，必须选择与基础词表内一致的写法。

如上，“勒”表示“这”，词典中存在这个词，必须按照词典提供的写法转写。
（2）普通话与方言都有的词，按普通话写法转写。
方言和普通话共有的词汇，他们可能存在读音上的差异，但文字形式一样。如“学习”“去”，虽然方言发音为“xuo xi”“ke”，但依然是普通话和方言共有的词汇。转写时直接转写为“学习”“去”。
（3）方言中特有的词汇。
方言中特有的词汇是指存在于方言中但不存在于普通话中的词汇，或者在方言中的含义与在普通话中的不一致的词汇。这类词如果基础词表里有，请按基础词表转写。如不存在基础词表中，**请选择“汉语拼音（同音字,同义词或释义）”**的形式转写。



如上，**词典中没有 kuo 对应的词条，因此标注时写为“kuo(廓,放)”,”kuo”为拼音，“廓”表示方言中的同音字，“放”是该词在普通话中的同义词，同音字与同义词写在括号内。**

3)英文与数字的标注

（1）英文标注

英文字母大写，单词小写，字母与字母之间空格，单词与单词之间空格，字母与单词之间也需要空格；字母/单词与汉字之间不需要空格，如“A B C D” “M S N” “Q Q” “ok ok” “thank you”等等；英文术语，发音人说成什么就写成什么，如果说是全称不要写成简写，如“A and B”不要写成“A&B”等。如遇网址，则说成什么就写成什么。

（2）数字标注

听到数字请转写成汉字。如听到说话人说“1（yao）31（yao）5894”时标注“幺三幺五八九四”。听到“1（yi）83 斤”时标注“一百八十三斤”。

4)标点符号标注

（1） 可标注的标点符号

中文输入状态下的逗号，句号。顿号、问号？感叹号！

（2）人物转换时，句尾都加句号或者问号；

（3）感叹号只有在感情非常强烈的句子中才使用，谨慎使用；

（4）转写不完整的句子,句末加句号；

（5）标点标注不能受音频切断的影响（不要每行标注末尾都打标点），标点标注主要的依据是语义，其次是说话人的停顿；

（6）遇到语气词、口头禅，根据句子意思添加标点；

（7）标注时不能受说话人停顿的影响，有些停顿是说话人个人的习惯；

（8）遇到“但是（呢）、然后（呢）、所以（呢）、同时（呢）、那么、比如说等等”此类词语，后面酌情加逗号；

（9） 说话人说错的句子不要加标点隔开，如：关关羽那那那么厉害；

（10） 标点标注需要结合上下文的意思，合理断句，加上适当的标点符号；

（11） 如果文本本身有标点，请仔细检查一遍，根据自己的理解，将标注错误的改正过来。

四、数据验收

文本字正确率（含标点符号）达到 95%以上视为合格。
字正确率（含标点符号）计算公式“标注正确的字数-多出的字数/总字数”

五、工资结算：

计件工资，每正确标注 1 小时有效音频，基本薪酬为 220 元（折算时薪约 16 元/小时），数据验收达标后，每 2 周结算一次并付款。对质量良好且产量稳定的标注员予以一定激励，如下，每天稳定任务 4 小时，综合时薪为 22 元以上。

苏州方言标注每日奖励	当天标注音频文件数量达到（个）		
	3	4~5	>=6
当日奖励金额(元)	25	40	70

1. 本方案自 2020/7/30 生效，至 8/31 结束。
2. 统计周期为周四~下周三。
3. 统计周期内“打回文件数+放弃文件数”<=5
4. 每周五前统计上周四~本周三数据，统计数据以后台导出为准。
5. 当日奖励不叠加。
6. 奖励发放时间为验收通过后的结算日。

项目周期为 8/1~8/31。由中国 Appen Global 进行微信支付。

六、人员要求：

- 苏州方言听、说流利，出生并成长于苏州为佳。
- 细致认真、学习能力强、态度端正、工作积极。
- 每天至少有 3 个小时花在项目上。
- 有电脑（台式或笔记本电脑）即可快速上手，需耳机。
- 需有**常用**邮箱接收每日质量报告邮件，与项目组每天互动。
- 注册 Appen Global China 平台兼职账号，便于工资结算：
  https://global.appen.com.cn/
  注册好后，邮箱会收到确认邮件。

六、人员要求：

- 苏州方言听、说流利，出生并成长于苏州为佳。
- 细致认真、学习能力强、态度端正、工作积极。
- 每天至少有 3 个小时花在项目上。
- 有电脑（台式或笔记本电脑）即可快速上手，需耳机。
- 需有**常用**邮箱接收每日质量报告邮件，与项目组每天互动。
- 注册 A9 平台兼职账号，便于工资结算：
  注册好后，邮箱会收到确认邮件。

七、项目申请端：

https://ui.appen.com.cn/worker-job/80199a20-792c-4c45-9323-f5911dc6d7bc?version=v1

八、项目名称：苏州方言标注