

小语种识别方案开发须知

概述：

为了给客户带来更多有价值的方案，开发多个小语种模型，方便方案商在无小语种语料的情况下，能进行初步的模型制作和样品测试；

本文档主要用于指引方案商快速了解相关开发及注意细节；

一、方案注意事项：

- 1、模块选型：**CI-03T（推荐）**、CI-33T
- 2、芯片选型：**CI1302（推荐）**、CI1303
- 3、词条数量： 建议<30 条；词条多了，一方面出现相似的词概率多，另一方面验收时间和优化反复时间也会多；
- 4、词条要求：命令词建议至少 4 个音节，尤其是唤醒词，必须保证至少 4 个音节，**命令词请选择发音有明显差异的词条**，发音相近的词容易误识别
- 5、测试注意：声音平稳，语速正常，不要刻意太慢或者太快

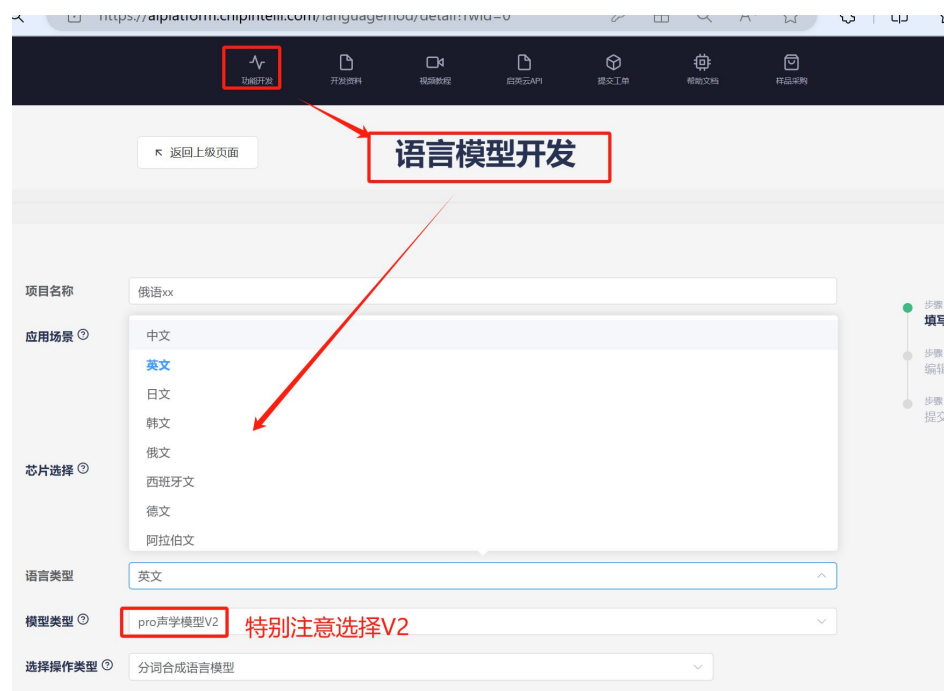
【特别注意】

- 1、选词好坏对项目落地的影响很大
- 2、可能存在客户的词条在原有的训练集中很少的情况，该情况下需要客户提供至少 50 人的语料进行迭代训练【a、3 周左右出模型 b、模型出来后单独开放给提供语料的客户 c、收到模型的客户，可自行更改逻辑后再次调试，或二次开发固件调试】
- 3、截止 24 年 08-17 支持语言：中、英、日、韩、西班牙、德、俄、泰、越南、阿拉伯语，这些模型可以直接提交文本联系客服制作模型（注意不支持阿拉伯数字）； 【其他语言需要确定商务条件，提供语料后，周期约 3 个月】

二、小语种方案开发流程

- 1、[注册智能公元平台账号](#)，熟悉中文固件开发流程（智能公元的文档中心）
- 2、与客户确定好词条，请客户提供至少 4 个人的测试语料（语料应尽量干净无底噪，2 男 2 女）

3、使用中文开发好固件后，再替换为对应小语种的语言模型、声学模型、和 cmd_info 文件



三、客户录音训练集的文件格式：

如需录音，请按下面格式进行录音


XX language
training voice rec

备注：

- 1、如果要录制语料，可先录制一个人的语料给我们看下是否格式正确，如无问题，在进行后续录音
- 2、录制后的声音必须严格按照一个声音一个文件按压缩包进行校对处理好，如未清洗好，会导致识别效果不稳定；

四、验收标准见表格：

小语种产品识别验收标准				
测试条件 \ 数据量		50 人	200 人	500 人
安静环境（35~45db），人声（60~65db）	1 米距离	90%	93%	95%
安静环境（35~45db），人声（60~65db）	3 米距离	90%	93%	93%
安静环境（35~45db），人声（60~65db）	5 米距离	85%	87%	90%
稳态噪声（55db），人声（65~70db）	1 米距离	85%	87%	90%
稳态噪声（55db），人声（65~70db）	3 米距离	80%	85%	87%
稳态噪声（55db），人声（65~70db）	5 米距离	75%	80%	85%
误唤醒： 测试噪音集合：8 小时安静，8 小时人声，8 小时音乐，噪音源不包含命令词及与命令词相似的声音； 测试方法:噪音与MIC 距离 1.5 米;音源数量：1 个;采用音响播放测试集；	误唤醒 2 次/24 小时			
备注： 1、数据采用文本中的录音方法，数据格式、环境要求进行录制 2、录音发音人尽量标准，清晰，录音男女比例均衡分布 3、命令词识别和误唤醒情况与词条文本强相关，建议使用 4-8 个音节内词条，相互之间发音尽量区别尽量大；【特别注意：前期先录制一个声音试听无问题，再开展录音】 4、识别率：所有词条的平均正确识别率 5、命令词：相似度尽量小，否则容易误识				

PS：如**命令词条区分度高**，则误识会大幅度减少，项目更容易落地量产；