# 《方言种类识别 AI 挑战赛》赛题解读

今天小助手为大家来详细解读“方言种类识别 AI 挑战赛”的赛题，希望可以帮助各位大神们更精准的了解赛题，不走弯路哦~

## 一、赛事简介：

**利用开放的方言语音数据集进行模型训练，优化方言种类的识别效果。**

联合国教科文组织统计，世界范围内，每两周就有一种语言消失。语言是文化的载体，中国幅员辽阔， 方言众多，保护方言，刻不容缓。科大讯飞基于“方言保护计划”，面向全球首次开放珍贵的中文方言语音数据集，聚焦方言种类识别问题，向广大人工智能开发者发起挑战，共同推进关于方言的算法研究和保护传承。

**参赛人群：**高校师生、企业单位、科研机构、创业团队、个人开发者等。可以以个人或者团队名义参赛，每支队伍规模不超过 7 人。

（备注：大赛组委会等有机会提前接触赛题和数据的工作人员禁止参赛，科大讯飞集团员工可以参加比赛排名，但无评奖资格，若名次在获奖范围之列，实际获奖团队顺延。）

## 二、赛题详情：

方言种类识别 AI 挑战赛任务为汉语方言语言种类识别，即根据给定语音，判断该语音属于哪个方言。科大讯飞全球首次开放覆盖中国六大方言区，总时长约 60 小时的 10 种汉语方言语音数据集，供参加竞赛的科研单位以及开发者免费使用。

根据测试语音长度，方言识别 AI 挑战赛分为两个不同难度的子任务，即任务一（有效语音长度≤3s ）和任务二（有效语音长度>3s）。结果评价指标为分类正确率 acc：即分类正确的语音条数/所有语音条数。训练集合与开发集合供参加竞赛的科研单位以及开发者调试系统使用，测试集合不开放，最终排名以参赛者提交的系统在线上测试集合上的结果为准，分类正确率越高排名越靠前。

## 三、开放数据：

初赛共有六种方言，分别来源于六大方言区，具体为：长沙话（changsha）、河北话 (hebei) 、南昌话 (nanchang) 、上海话 (shanghai) 、闽南语 (minnan) 和客家话（kejia）。每种方言平均包含 6 小时的朗读风格语音数据，覆盖 40 个说话人。数据由各个型号的智能手机采集，录制环境包含安静环境和噪声环境。数据以采样率 16000Hz，16 比特量化的 PCM 格式存储。

数据集包含训练集、开发集和测试集三个部分。训练集每种方言有 6000 句语音，包含 30 个说话人，其中 15 位男性和 15 位女性，每个说话人 200 句语音；开发集和测试集分别每种方言包含 5 个说话人，其中开发集为 2 名女性和 3 名男性，测试集为 3 名女性和 2 名男性。开发集和测试集的数据根据语音段的时长分为两类，一类是小于等于 3 秒的短时数据（任务一），另一类是大于 3 秒的为长时数据（任务二），分别对应于两个比赛任务，其中每个说话人两类数据各 50 句，共 100 句。训练集、开发集、测试集的说话人均没有重复。数据具体描述见表 1。

为了增加本次比赛技术方案的多样性，每条语音对应文本内容的音素序列标注也将同样提供。

| 初赛数据集 | | | | 训练集 | | | 开发集（≤3s,>3s） | | | 测试集（≤3s,>3s） | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 方言代码 | 方言种类 | 口音区域 | 信道 | 说话人个数 | 每个说话人语音数 | 总语音数 | 说话人个数 | 每个说话人句数 | 总语音数 | 说话人个数 | 每个说话人语音数 | 总语音数 |
| nanchang | 南昌方言 | 南昌及周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |
| shanghai | 上海方言 | 上海及周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |
| hebei | 河北方言 | 石家庄、保定、衡水等及周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |
| changsha | 长沙方言 | 长沙及周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |
| kejia | 客家方言 | 梅县、梅州、惠阳等及周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |
| minnan | 闽南方言 | 厦门、漳州、泉州等周边地区 | 手机 | 30 | 200 | 6000 | 5 | 50 | 250 | 5 | 50 | 250 |

表 1 数据详细描述

## 四、参赛系统：

参赛系统的搭建方法不限，所有机器学习的方法均可以使用，并且参赛系统可以是多种方法以任意形式的结合，比如投票法等等。两个不同的比赛任务可以采用两套完全独立的系统。比赛采用离线测试的方式进行，因此本次比赛对参赛系统的响应时间不做要求。

同时，考虑到复赛和初赛的难度差异，复赛和初赛也可以采用不同的系统。

## 五、评测方式：

本次比赛的测试集是不公开的，因此需要参赛者提交自己的系统，具体操作方式如下：

a)初赛提交系统时，请提交参赛者名称、第一作者、该系统对应的任务、参赛系统（提交方式见下面详述）、训练集和开发集上的分类正确率

b)复赛提交系统时，需要额外提交一份参赛系统的论文或者说明书（最好能够附带提供源代码），详细介绍系统的构成、训练方法和对应的参数

c)如无特殊情况，每天上午 11 点在官方网页上公布各个参赛者在测试集上的分类正确率并对结果进行排序（每个参赛单位的结果以最新提交的为准）

**为了能够正确的进行测试，所有测试均在相同配置的 Linux 64 位服务器上统一采用 CPU 进行测试。因此提交的系统不能是 windows 等其他操作系统下的程序，并且不能和 GPU、FPGA 等其他硬件相关联。同时为了方便参赛者更好的参加比赛，本次比赛制定了详细的参赛系统提交和评估系统，介绍如下：**

**一、评测系统：**

1.评测系统目录结构

/dataset……开发集目录，用于系统提交后的正确性验证

/inference……评测代码及资源目录，系统运行的当前路径

/result……请将评测结果以 result.txt 命名，存放在此目录

result.txt 请按照 result.txt 中的格式。

2.本地开发调试

a)使用开源深度学习训练框架(推荐)

请从公开镜像仓库下载对应版本的深度学习镜像 CPU 版本，编写本地程序进行评测。以 tensor flow 工具为例：

I.下载镜像，docker pull tensorflow/tensorflow: 1.7.0

II.下载开发集到 /dataset 目录，将评测代码 inference.py 及评测所需资源复制到 /inference 目录

III.运行镜像，docker run –it –v /dataset:/dataset –v /inference:/inference –v /result:/result tensorflow/tensorflow:1.7.0 /inference/inference.py

IV.查看输出结果，并检查该输出结果的正确性

其他框架与此类似。

b)使用非开源深度学习训练框架

首先，要将使用的深度学习训练框架制作成 docker 镜像，上传至公开镜像仓库（推荐使用国内稳定镜像仓库服务，如 UCloud），制作详情可参考 docker 官方文档，具体操作如下：

I.在 hub.docker.com 注册账号，并创建仓库

II.本地执行 docker tag your_demo your_account/your_demo:latest

III.本地执行 docker push account/your_demo:latest，等待命令执行成功后，即可在 hub.docker.com 网页上，看到新提交的镜像信息

镜像提交完成后，参考使用开源深度学习训练框架中的步骤，进行本地开发和调试，具体的为：

IV. 下载上传的镜像，docker pull yourtoolname

V.下载开发集到 /dataset 目录，将评测代码 inference.py 及评测所需资源复制到 /inference 目录

VI.运行镜像，docker run –it –v /dataset:/dataset –v /inference:/inference –v /result:/result yourtoolname /inference/inference.py

VII.查看输出结果，并检查该输出结果的正确性

3.提交评测系统

a)将/inference 目录打包成 tar 文件，tar –cvf inference.tar inference/

b)在比赛官网中评测系统提交页面进行上传

4.线上验证及评测

a)配置系统所需的镜像仓库地址，镜像入口及验证参数(/dataset 目录由系统自动将开发集挂载到镜像内)

b)点击”运行”，等待评测结果

c)如评测报错或效果异常，请排查/inference 目录结构、镜像等配置信息

d)确定评测结果无误后，点击”提交”。此时/dataset 内将替换成为非公开的测试集，并记录系统效果

**二、基线系统介绍：**

官方针对汉语方言分类比赛任务搭建了基于 i-vector(Identity Vector)的基线系统，本基线系统的搭建目的并不是提供一个有竞争力的系统，而仅仅是提供一个效果参考。i-vector 基线系统基于 TV（Total variability, TV）模型，其中语音特征采用的是 56 维的 SDC（Shifted delta cepstral, SDC）特征；UBM(Universal Background Model, UBM)模型包含 256 高斯；i-vector 维度为 300；i-vector 之后采用 LDA(Linear Discriminant Analysis, LDA)提取方言信息，LDA 后的方言表征矢量维度为 5。每种方言的表征矢量由该方言训练集所有语音的表征矢量取平均得到。在测试时，待测试语音与所有方言表征矢量求取余弦距离，取距离最大的方言为判定的方言种类。

| System | Acc(dev set > 3s) | Acc(dev set ≤3s) | Acc(dev set > 3s) | Acc(dev set ≤3s) |
| --- | --- | --- | --- | --- |
| Baseline | 71.93 | 57.40 | 66.20 | 54.53 |

**三、限制条件：**

为了保证比赛的公平性，本次比赛仅允许使用官方发布的数据和标注，否则比赛成绩将被视为无效。不符合规定的情况包括以下几种：

a)参赛系统搭建过程中有任何一个环节（包括数据加噪、模型初始化等）用到了官方发布的训练数据集之外的其他数据

b)人工对发布数据集的音素序列标注进行矫正或改动

c)其他对发布数据集的人工处理，比如人工对数据集进行语音端点检测等

以下情况是允许的：

a)仅利用官方发布的训练数据集进行数据的机器仿真和加噪

b)利用官方发布数据集中已公布的所有信息，包括性别、说话人等

**四、数据集：**

注：完整的初赛数据集请在报名成功后前往个人中心-我的比赛，进入方言识别比赛专题页面进行下载

**五、FAQ:**

**1、我如何了解自己的参赛状态以及提交作品?**

请前往“个人中心”（通过大赛官网的菜单栏中“个人中心”进入）查看自己已报名的比赛，点击已报名的比赛可进入相应赛题的专题页面查看赛题详情以及提交作品查看成绩等。

**2、如何和组委会取得联系？**

您可以通过以下三种方法联系大赛组委会：

（1）发送邮件至：aicompetition@iflytek.com

（2）加入官方微信群：请添加 AI 大赛助手微信号——iFLYTEKAI（不区分大小写），AI 大赛助手会邀请您进入 AI 大赛官方微信群

（3）前往大赛论坛

今天的干货就先分享到这里啦~

小助手在这里恭候各位大神的作品哦！