2022 世界机器人大赛—共融机器人挑战赛

智能人机交互组-语音交互

竞赛手册

“共融机器人基础理论与关键技术研究”重大研究计划指导专家组

2022 年“世界机器人大赛—共融机器人挑战赛”组织委员会

2022 年 5 月

一、 赛事内容

1.比赛目的

由于近些年来人工智能技术的进步，语音识别技术得到了很大的发展。然而，在高噪声和无声等场景下的说话内容识别依然面临巨大挑战。近年来，基于视觉信息和面部肌电的语音识别技术取得了可喜的进步，音视频结合的多模态语音识别和和基于面部肌电信息的新型无声语音识别技术也逐渐得到广泛关注。然而目前视觉信息和肌电等非音频模态的信息可以如何来有效的帮助识别说话内容尚未被很好地探索。

本次竞赛目的是探索如何有效结合语音信息与视觉信息及面部肌电信息来实现鲁棒的语音识别，以期促进语音识别技术的发展，并期望为当下的语音识别领域提供新思路与新方法。根据实际应用场景的需求与任务难度等级，我们分别设置了词级与句子级、闭集与开集识别、面部肌电跨被试的任务，希望通过这些比赛任务可以抛砖引玉，启发今后对于语音识别技术的研究工作。

在本次比赛中，我们不仅期待参赛队伍可以在识别性能上取得优异成绩，也鼓励选手们在方法层面为语音识别这一领域带来新的启发，同时也欢迎未用于参赛方案但是契合比赛主题的有关贡献。

2.语音交互赛项

2.1 任务说明

该赛项包含三个任务，主要从语音识别的对象与识别的范围两个层面分别设立了难度不同的三个任务，分别是基于音视频信息的词级

闭集语音识别、基于视觉信息的句子级无声语音识别以及基于面部肌电的句子级无声语音识别。

参赛团队可以参加其中任意一个或者同时参加多个任务。每个任务都包含性能分和效率分，最终分数是两项分数总和。每个任务按照各自的最终分数独立进行排名。

注：参赛团队需要在三个任务各自的赛项页面，进行报名、提交作品以及参与最终分数的独立排名。各个任务赛项页面链接如下：

- 任务一：基于音视频的词级闭集语音识别

  https://fc.osredm.com/competitions/pn5ikx/home
- 任务二：基于视觉模态的句子级无声语音识别

  https://fc.osredm.com/competitions/whbfk6/home
- 任务三：基于面部肌电的句子级语音识别

  https://fc.osredm.com/competitions/56plot/home

任务一：基于音视频的词级闭集语音识别

任务要求：该任务的主要目标是根据视频片段中说话人的说话音频和与说话过程对应的面部动作序列，实现对单个中文词的分类。该任务的目的是为了检验基于音视频的语音识别在识别目标相对容易但数据具有一定挑战的条件下所能达到的程度。参赛者需根据主办方提供的数据集，构建鲁棒高效的音视频词级唇语识别模型，鲁棒应对不同数据条件下，包括数据中的音频噪声、图像的明暗程度不一、说话人的性别/年龄/妆否/姿态不一等挑战。

数据说明：本项任务所有数据均由主办方提供，包括训练集、初赛测试集和决赛测试集。训练集用于模型训练，初赛测试集用于初赛评分，决赛测试集用于最终阶段的模型评估。训练集、初赛测试集和决赛测试集均源于同一类数据，但为了方便训练和测试，提供的信息稍有不同：训练集中包含大量音频片段及对应的图片序列，标注信息将给出各个词在完整音频片段与图像序列中的起止时间、对应的词语内容等信息，参赛者需根据提供的标注信息提取其中各单个词的样本进行模型训练；初赛测试集和决赛测试集中，每个样本直接对应于目标单词，主要包含三部分内容，分别为该目标单词对应的文本标签、音频片段和图像帧序列，其中文本标签包含该目标中文词的无调汉语拼音和汉字表达两种形式。参赛团队可自选建模对象，音频片段为对应于该目标词的原始 wav 文件，图像帧序列为说话人说该目标词时面部的连续变化的图片序列。**在本任务中，参赛者只能使用提供的训练集和初赛测试集进行模型构建，不允许使用额外数据。**本项任务共涉及 1000 个目标词，均包含在训练集中，不存在集外词(OOW，Out-Of-vocabulary Word)。所提供数据集共包含约 718,018 个样本，汉字字符总计超过 1000,000，覆盖超过 2,000 个说话人，该数据集在说话模式和成像条件等方面都具有多样性，包括每一目标词的样本个数、视频分辨率、光照条件、说话人的性别/年龄/妆否/姿态等。需要注意的是，在数据集中，每个词对应的样本总数不完全等同，存在某些词训练样本较多、某些词训练样本较少的情况，同时每个样本的音频长度和图片帧数也与实际说话速度有关，不完全一致。参赛团

队在做此项任务时需要考虑对这些因素的处理。

基准模型： 本任务提供基准模型参考，链接如下：https://github.com/VIPL-Audio-Visual-Speech-Understanding/learn-an-effective-lip-reading-model-without-pains 。参赛队伍可以此基准模型为出发点，构建自己的识别模型。

评分规则：本项任务采用识别准确率(Accuracy)作为评判标准。每个目标词完全识别正确则作为识别正确，否则认为识别错误。此外，参赛者需要提供自己的模型代码、权重、可执行文件以及可执行文件的运行方式说明文件，无法正常运行代码将扣除一定分数。

任务二：基于视觉模态的句子级无声语音识别

任务要求：该任务的目标是根据视频中说话人的面部动作信息，来分析和识别出对应的说话内容。该任务的目的是为了模拟真实场景下，通常难以保证待识别内容一定包含在训练集中的情况，以评测基于视觉模态的语音识别模型的鲁棒性。本任务的初赛测试集和决赛测试集包含多样的视觉与语言条件，例如光照、说话姿态、年龄、性别、妆容、背景噪音、方言等。**该任务允许参赛者引入主办方提供的训练数据之外的任意数据，但需要在提交结果时明确说明用的额外数据的情况，也可以使用预训练模型、添加额外语言模型、采用多模型融合等方式。**模型的最终性能在主办方提供的测试集上评测。

数据说明：主办方提供训练集、初赛测试集与决赛测试集，分别用于初赛阶段和最终决赛阶段的评测。集合中将包含两类数据：一类

是与所提供的训练视频同源的测试数据，另一类是与所提供的训练数据不同源的测试数据，分别用于验证模型在同源与不同源条件下所能达到的性能。初赛测试集与决赛测试集数据类型相同，但决赛数据的难度与规模会高于初赛数据。数据集将提供包含唇部区域的图像序列和对应于说话内容的文本标注，其中标注文本将以中文汉字与汉语拼音两种形式给出，可由参赛队伍自行选择建模单元。

评分规则：该任务采用识别的字错误率（Character Error Rate，CER）与句错误率（Sentence Error Rate，SER）两项作为综合评测指标。此外，参赛者需要提供自己的模型代码、权重、可执行文件以及可执行文件的运行方式说明文件，无法正常运行代码将扣除一定分数。

字错误率(CER)的计算方式是预测结果与真实文本之间的汉字级的编辑距离和真实文本汉字个数的比值。编辑距离指将预测输出通过插入、替换、删除三种编辑操作来得到真实标注的汉字文本句子所需的操作次数。句错误率(SER)计算方式为预测文本是否与真实文本完全匹配，若完全匹配则认为该句识别正确，否则认为该句识别错误。该任务的最终测试分是每个测试样本所对应测试分数的平均值。

任务三：基于面部肌电的句子级语音识别

任务背景：目前，随着社会老龄化的情况日趋严重，人们所面临的中风、脑梗塞等其他疾病的几率也逐渐变高。受疾病因素影响，言语困难和突发言语障碍的情况使得传统基于音频信号的语音交互方式信道受阻，老年人难以在日常和突发疾病时有效利用语音交互进行

辅助交流。相似的，患有咽喉疾病或进行喉部切除术的病人也常伴有言语困难和障碍，影响日常生活。考虑到老年人和患有特殊疾病的患者日常生活及康复治疗的辅助交互需求，设计和实现面向养老助残特殊应用场景的无声语音识别系统。通过采集和处理使用者的面部、颈部与发音相关的肌肉位置处表面肌电信号，进行指令内容识别。

任务要求：基于面部肌电的句子级语音识别赛项计划设置面向养老助残应用场景的无声语音交互技术研究。根据老人年及患有特殊发音疾病病患的日常生活辅助需求和紧急场景（如突然失声）的交互辅助需求，要求参赛团队据此应用场景和数据集设计算法解决该场景下语音交互问题，通过指令识别测试验证无声语音识别技术在实际应用中跨被试识别的有效性、实用性及适应性。

数据说明：基于马斯洛需求层次理论设计 100 个中文指令，每个中文指令包括三到五个不同的中文汉字。根据此应用场景和指令集内容同步采集不同说话人的面部和颈部 6 个发音运动相关的肌肉表面肌电信号，通过数据采集程序得到 6 通道，时长 2s 的表面肌电信号。训练集数据包括 30 名被试，100 类指令以及一个额外的空指令，每类指令每个被试采集了 10 条表面肌电数据。通过数据清洗和筛选，最终每类指令的样本数以数据集实际提供为准。主办方提供的训练数据集包括六通道的表面肌电信号以及对应的指令标签，参赛团队可自行划分训练、验证和测试样本用于模型和识别系统搭建，训练和验证。在本任务中，参赛者只能使用提供的数据及标签进行模型构建，不允许使用任何额外数据。

在模型搭建完成的基础上，参赛团队可自行下载测试数据集。测试数据集由相同的数据采集系统和流程得到六通道的表面肌电数据。与训练集不同的是，测试集的数据样本来自 20 名新被试的数据采集实验。由于表面肌电信号是一种典型的生理电信号，具有较强的被试依赖性，以新被试的肌电数据考察参赛团队语音识别系统跨被试的识别性能。主办方提供的测试数据仅包括六通道的表面肌电数据，其对应的指令内容与训练集相同，需要参赛团队进行模型分类识别测试，得到分类识别的结果。

评分规则：本项任务采用识别准确率(Accuracy)作为评判标准。每个目标指令内容完全识别正确则作为识别正确，否则认为识别错误。由于训练集提供了指令标签，此项识别准确率的最终计算以标签识别结果为基础。具体计算公式为：

准确率 = 测试集标签分类正确的样本个数/测试集总样本个数

参赛团队在完成无声语音识别系统设计的基础上，自行下载跨被试的测试数据集，通过模型测试和分类识别得到识别结果。参赛者需要提供自己的模型代码、权重、可执行文件、基于测试数据集的模型识别结果文件以及提交材料的说明文件。参赛团队提供的程序无法运行、结果文件保存错误导致的识别准确率计算偏差或无法判定等问题均会导致最终成绩的扣分。

2.2 评分说明

比赛评分包括任务分和效率分两方面。

任务分：我们根据比赛任务完成情况进行评分，初赛阶段各任务

的总分为800分，决赛阶段该项对应总分为600分。任务一的任务分计算公式为：总分*准确率；任务二的任务分计算公式为：总分*[1-0.5*(CER+SER)]，该分数小于0时按0分计，任务分为0分的队伍不参与排名；任务三的任务分计算公式为：总分*准确率。比如，某支参赛队伍初赛时在任务二中的CER为21%，SER为59%，那么该队伍的任务分为800*[1-0.5*(0.21+0.59)]=480分。

效率分：为鼓励兼具创新性与实用性的比赛方案，充分考虑到不同参赛组之间所能使用的硬件资源差异等因素，我们将从模型文件大小、模型推理速度、模型训练使用的数据量、算力资源消耗等方面综合评判参赛方案的效率分。其中初赛阶段该条目为200分，决赛阶段该条目为400分，期望选手设计方案时能兼顾性能与实用性，而非单纯通过强大的硬件资源和庞大的数据量来提高模型的性能却忽略了模型的实际应用性。三个任务的效率分评分标准相同，**初赛阶段以模型文件大小为度量标准，比重为1.0，决赛阶段同时包括运行时间和提交文件的大小两个因素，各占比重0.5。**

对于任务一和任务二，①运行时间以主办方统计的时间为准，比如选手提交的可执行脚本为“run.sh”，主办方通过“time ./run.sh”统计该脚本的运行时长，进而计算运行时长对应分数；文件大小以参赛队伍提交的压缩包大小为准。以M表示效率分满分时的分数，r表示运行时间比重(决赛阶段r=0.5,初赛阶段r=0)，则运行时长分数统计的规则为，脚本执行时间超过900s时得分为0,大于300s小于900s时得分为M*r*60%，小于300s时得分为M*r*60%+M*r*40%*(1-运行时

长/300)，时长均以 s 为单位。② 模型文件大小分数标准为，大于1024M 时得分为 0，大于 512M 小于 1024M 时得分为 M*(1-r)*60%，小于 512M 时得分为 M*(1-r)*60%+M*0.5*40%*(1-文件大小/512)，文件大小以 M 为单位。比如，某只参赛队伍决赛时的运行时长为 420s，文件大小为 200M，M 为 200 分，那么该队伍的运行时长分数为 200*0.5*60%=60 分，文件大小分数为 200*0.5*60%+200*0.5*40%*(1-200/512)=84.375，效率分总分为 60+84.375=144.375 分。

对于任务三，①运行时间以主办方统计的时间为准，比如选手提交的可执行脚本为“run.sh”，主办方通过“time ./run.sh”统计该脚本的运行时长，进而计算运行时长对应分数；文件大小以参赛队伍提交的压缩包大小为准。以 M 表示效率分满分时的分数，r 表示运行时间比重(决赛阶段 r=0.5,初赛阶段 r=0)，运行时长分数统计的规则为，脚本执行时间超过 1500s 时得分为 0，大于 1000s 小于 1500s 时得分为 M*r*60%，小于 1000s 时得分为 M*r*60%+M*r*40%*(1-运行时长/1000)，时长均以 s 为单位。② 文件大小分数标准为，大于 150M 时得分为 0，大于 80M 小于 150M 时得分为 M*(1-r)*60%，小于 80M 时得分为 M*(1-r)*60%+M*(1-r)*40%*(1-文件大小/80)，文件大小以 M 为单位。

2.3 排行榜

每个任务单独评测。参赛队伍可以选择只参加其中一个任务赛道，也可同时参加多个任务赛道，排行榜每周更新 1 次，以每周最后一次

提交为准。

训练/初赛测试集：参赛队伍可在初赛测试集发布后提交结果，初赛测试集排行榜在最终的测试集发布之前会不断更新，每支队伍在每一赛道每周至多可以提交 5 次。

决赛测试集：决赛测试集上的最终结果即为参赛队伍的最终成绩，用于决定比赛名次等，每支队伍在每一赛道每周至多可以提交 5 次。

每个任务的最终成绩和分数计算方式同“评分说明”

二、 比赛流程

比赛分为两个阶段，分别为初赛和决赛。报名截止后，所有参赛团队将进行初赛，在初赛中，我们的分数分配方式为任务分800，效率分200，共计1000分对参赛队伍进行打分，通过分数排名，选拔出符合参赛条件的参赛团队进入决赛，由大赛组委会秘书处下发参赛通知，进行决赛。在决赛中，我们的分数分配方式为任务分600，效率分400，共计1000分对参赛队伍进行打分，最终根据决赛的得分排名，评选比赛等次。

三、 报名要求

1.报名截止时间：请参考各比赛组别竞赛规则规程说明

2.报名“2022世界机器人大赛—共融机器人挑战赛” 将通过统一报名系统注册并填报信息进行参赛，具体报名入口及开放时间请关注官网实时更新。

四、 提交作品说明

参赛队伍需将可执行脚本以及说明文档打包成 zip 文件，文件在

比赛平台提交。在提交参赛作品时，请以“语音任务X-赛队名称”的形式进行对打包文件命名，如“语音任务2-飞鸟队.zip”。

统一命名格式：为方便计算模型大小和匹配预测结果，zip 内的代码、模型、预测结果、说明文档四种类型应分别存放，**代码与模型存放的文件夹各自统一命名为“code”、“model”，预测结果统一命名为“result.txt”，说明文件统一命名为“ReadMe”。若模型中涉及多个子模型，则应将所有子模型一起打包到“model.zip”。**

预测结果文本文件的格式要求如下：

任务一：每一行为一个样本预测的标签序号，结果按照测试数据的顺序提供，其中标签序号的排序方式应当同主办方提供的词汇表中词语的排序方式保持一致。比如，结果文件中第11行的内容为“90”，则表示对应于测试数据中的第11个样本的预测结果是主办方提供的词汇表中的第90个词（标签序号从0开始编号）。

任务二：每一行为一个样本的预测结果，结果按照测试数据的顺序提供，比如结果文件中的第12行对应的内容为“早上好”，则表示测试数据中的第12个样本的预测结果为“早上好”。

任务三：每一行为一个样本预测的标签序号。请将基于测试数据集的模型标签识别结果**按测试数据的顺序**以整数形式保存在文本文件中，便于主办方根据数据指令标签真值进行识别准确率计算和判定。

注意，提交的代码无法运行、提交格式不符合要求、未评测所有样本等将扣除一定分数。

注意：

（1）比赛数据将以邮件的形式发送获取方式，请注意邮件查收。比赛数据严禁随意传播或用于其它用途，一经发现，必将追究相关责任。

（2）参赛队伍如有作弊行为将直接取消参赛资格。

五、 奖项说明

1. 获奖团队在未来申请某部委项目时，同等情况下予以优先考虑。

2. 语音交互技术组三个任务分别设置一等奖 1 名、二等奖 2 名、三等奖 3 名，奖金根据比赛赞助总金额分配。

比赛任务	奖项
基于音视频的闭集词级语音识别	一等奖 1 名
	二等奖 2 名
	三等奖 3 名
基于音视频的句子级开集语音识别	一等奖 1 名
	二等奖 2 名
	三等奖 3 名
基于面部肌电的句子级语音识别	一等奖 1 名
	二等奖 2 名
	三等奖 3 名

3. 比赛遵循公开、公平、公正的原则，对比赛获胜及优秀团队颁发相应荣誉证书。

六、 赛事联系人

总联系人：吴沁蕾

联系电话：010-68600682，18811067454

联系邮箱：wrcc_office@163.com，ciewuqinlei@163.com ，

wuqinlei@cie-info.org.cn

语音交互联系人：冯大蕗（任务一、任务二），吴竞寒（任务三）

联系电话：010-62600520（任务一、任务二），

13920317102（任务三）

联系邮箱：lipreading@ict.ac.cn（任务一、任务二），

wujinghan@tju.edu.cn（任务三）

请优先邮箱咨询。