doi:10.19562/j.chinasae.qcgc.2025.03.003

基于ChatGLM2大模型的座舱多模态拒识模型研究*

张 强[1,4],石 琴[1,2,3],程 腾[1,2,3],倪 昊[1,2,3]
(1. 合肥工业大学汽车与交通工程学院,合肥 230009; 2. 自动驾驶汽车安全技术安徽省重点实验室,合肥 230009;
3. 安徽省智慧交通车路协同工程研究中心,合肥 230009; 4. 奇瑞汽车股份有限公司,芜湖 241000)

[摘要] 在智能网联汽车领域,车载系统在复杂环境下对非指令性语音输入的识别精度(系统正确识别语音输入的比例)具有重要意义。针对这一挑战,本文提出了一种多模态拒识模型。该模型基于开源的ChatGLM2-6B大型语言模型,并针对车机交互场景进行了专属的拒识数据集构建和模型微调。拒识数据集采集自真实的驾驶场景,综合了语音信息与驾驶员的面部朝向、情绪等非语言信号,以提供更为丰富的交互信息,有效克服了纯语言识别机制在复杂环境中的局限性。通过实验发现,多模态拒识模型相较于纯语言拒识模型,在测试集上展现出更高的识别准确率ACC和更低的误识别率FRR。

关键词:智能网联汽车;车载语音交互;拒识;大模型

Research on Multimodal Rejection Model of Cockpit Based on ChatGLM2 Large Model

Zhang Qiang[1,4], Shi Qin[1,2,3], Cheng Teng[1,2,3] & Ni Hao[1,2,3]
1. *School of Automotive and Transportation Engineering, Hefei University of Technology, Hefei* 230009;
2. *Key Laboratory for Automated Vehicle Safety Technology of Anhui Province, Hefei* 230009;
3. *Engineering Research Center for Intelligent Transportation and Cooperative Vehicle-Infrastructure of Anhui Province, Hefei* 230009;
4. *Chery Automobile Co., Ltd., Wuhu* 241000

[**Abstract**] In the field of intelligent connected vehicles, the recognition accuracy of in-car systems for non-command voice input in complex environment (the proportion of correct voice input recognition by the system) is of great significance. To address this challenge, in this paper a multimodal rejection model is proposed. The model is based on the open-source ChatGLM2-6B large language model and has undergone exclusive rejection dataset construction and model fine-tuning for the in-vehicle interaction scenario. The rejection dataset is collected from real driving scenarios, integrating voice information with the driver's facial orientation, gestures, and emotion, and other non-verbal signals to provide richer interaction information, effectively overcoming the limitation of pure language recognition mechanisms in complex environment. Through experiments, it is found that the multimodal rejection model shows higher recognition accuracy (ACC) and lower false rejection rate (FRR) on the test set compared to the pure language rejection model.

Keywords: intelligent connected vehicles; in-vehicle speech interaction; rejection; large model

* 安徽省自然科学基金(2208085MF171)、中央高校基本科研业务费专项资金(JZ2023YQTD0073, PA2023GDSK0112)和安徽省重点研究与开发计划项目(202304A05020087)资助。
原稿收到日期为 2024 年 05 月 20 日,修改稿收到日期为 2024 年 07 月 23 日。
通信作者:程腾,副教授,博士,E-mail:cht616@hfut.edu.cn。

前言

在智能网联汽车领域的迅猛发展中,智能座舱和车机交互助手的广泛应用标志着车辆功能、安全性和便捷性的显著提升。然而,随之而来的挑战也日渐凸显,尤其是在车载语音交互系统领域。用户对车机语音交互的需求日益增长,他们不仅期望系统能够理解和执行基本命令,还希望它能够在各种复杂环境下准确无误地响应,这使得提高车载系统在处理非指令性语音输入时的拒识能力变得尤为重要。

语音拒识技术旨在识别用户的语音是否针对语音助手,最早由 Amazon 的 Mallidi 等[1]在 2018 年提出,并应用于语音助手产品 Alexa。Alexa 利用 ASR(automatic speech recognition)解码器特征、声学特征和词汇特征,通过深度神经网络进行设备指向性分类。然而,首次唤醒后的语音通常具有高度指向性,这一点在连续对话中尤为明显。Huang 等[2]通过实验验证并优化了这一现象,降低了 Alexa 的相等错误率。Gillespie 等[3]进一步结合语义和声学特征,引入上下文信息,显著提升了拒识性能。尽管基于特征级的语音拒识研究众多,但受限于 ASR 系统的依赖和数据隐私问题,研究进展存在挑战。相比之下,模态级的语音拒识通过文本和语音模态的结合,减少了对 ASR 系统的依赖[4-5]。Shriberg 等[6]尝试使用 ASR 转录的文本进行拒识,而 Lee 等[7]针对多人多领域场景[8-9]提出了一种有效的拒识方式。在音频模态方面,Norouzian 等[10]和 Tong 等[11]分别提出了基于卷积神经网络和长期记忆神经网络的方法,以适应动态语音输入和提高拒识准确性。

然而,当前的研究还面临着两大挑战:(1)传统的深度学习拒识模型的训练需要大量真实且多样化的人机交互数据集,但是获取高质量高覆盖的数据集是一项极具挑战性的任务。它不仅需要大量的时间和资源,还须解决隐私保护和数据安全等问题。(2)现有的主要依靠纯语言输入的识别机制在复杂环境中面临诸多局限性。纯语言的识别机制往往难以保持高精度,导致拒识率不尽如人意。这不仅影响用户的交互体验,更在关键时刻可能影响车辆的安全操作。

针对这些问题,本研究提出了一种新的解决方案:采用预训练大模型结合车机交互专属拒识数据集进行微调的策略,如图 1 所示。该微调的本质在于在训练过程中固定大部分预训练模型的权重,仅调整少量与特定任务相关的提示参数[12],该方案的核心在于以下两个关键方面。

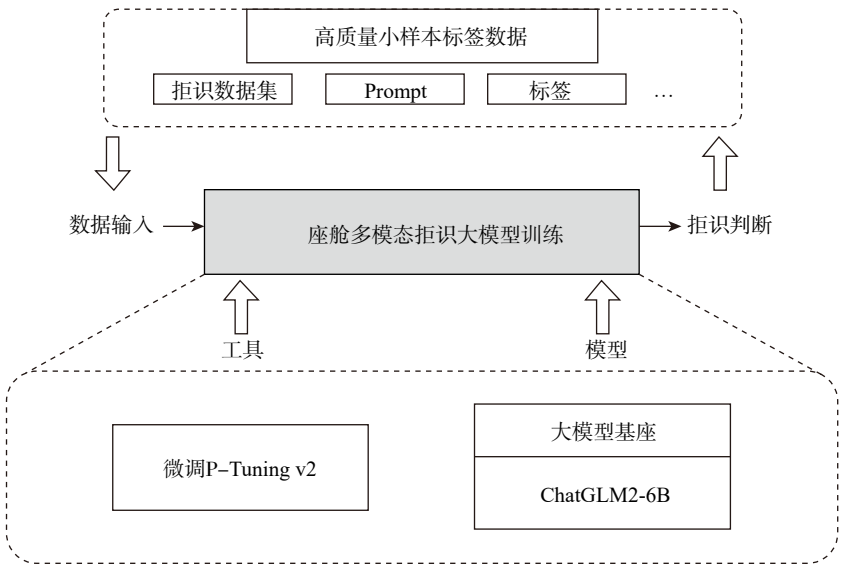

图 1　拒识大模型研究总体框架图

(1)通过人工标注实车采集的真实人机交互场景,构建了一个专属的车机交互拒识数据集。该数据集不仅包含语音信息,还融合了驾驶员的面部朝向、情绪等非语言信息,以提供更全面的交互信息。这种多模态输入使得系统能够更准确地捕捉用户的真实意图,从而在各种嘈杂和非标准语言环境下保持高效的交互性能。

(2)通过采用开源的预训练大语言模型,系统能

够吸收和学习丰富的特征表示和深层次的语义信息。这种模型由于其在大量的文本数据上的预训练,为系统提供了强大的基础能力。再用上述专属车机交互拒识数据集对大模型进行微调,系统能够更好地适应特定的任务需求和场景。这种微调不仅提高了系统在特定场景下的识别精度,也优化了系统的响应速度和准确性。

1 车机交互拒识数据集

在本节中,将逐一介绍用于实验的车机交互专属拒识数据集的采集、筛选、标定和划分过程,如图 2 所示。主要涵盖数据集的数据来源以及构造方法。

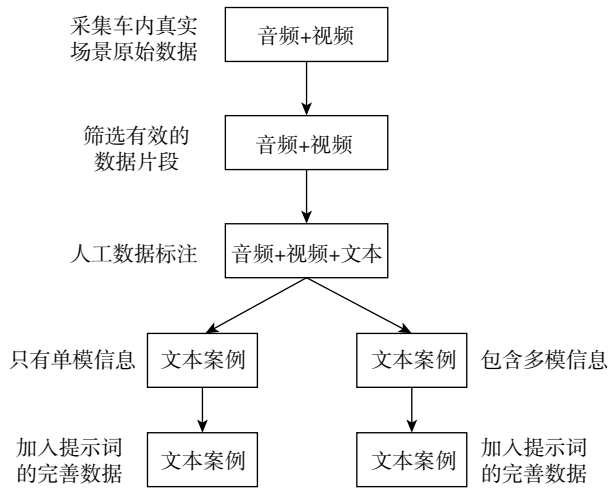

图 2 数据集构建流程

1.1 数据采集

本文研究的是智能座舱内汽车智能语音助手对驾驶员的拒识能力,为此构建了一套专用数据集。数据来源于课题合作方奇瑞提供的丰富的实车测试数据集,采集自实车测试,涵盖了车内视频和驾驶员与汽车智能语音助手的交互语录。数据包含多种复杂的驾驶场景,包括城市道路、高速公路和乡村道路等,确保了数据集的多样性和全面性。为了保护隐私和数据安全,所有语音和视频数据均以加密形式存储于奇瑞汽车的内网环境,并且所有相关实验操作均在企业内网环境下进行。

在数据收集方面,数据收集车内安装有高分辨率摄像头记录车内情况,摄像头位于驾驶室内,专注于捕捉驾驶员的面部表情以及整个车内的环境,如图 3 所示。此外,车内还部署高质量的传声器阵列,用于清晰地记录驾驶员与智能语音助手之间的对话。为了保证数据的一致性和同步性,车上也同时安装了先进的数据收集器,这些收集器能够同时连接车内的视频和音频设备,确保所有数据都有相同的时间戳。这种同步机制对于后续的数据分析至关重要,它能够准确地关联视觉信息和音频信息,从而深入分析驾驶员与智能助手交互的动态过程。

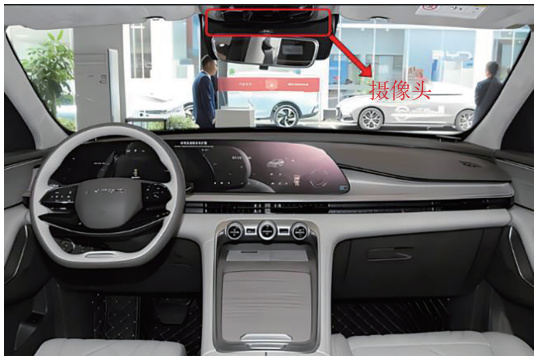

图 3 车内传感器位置图

视频和音频数据是对原始数据进行了精细的预处理得到的,包括视频和音频的去噪处理、时间戳的校正以及数据的格式化。因此这些数据具有高度的真实性和实用性。通过在真实的驾驶场景下收集数据,这些数据集为后续的研究提供了丰富的材料。

1.2 数据筛选

在数据采集完成后,面临的一个主要挑战是从大量采集的视频和音频材料中筛选出包含有效驾驶员与智能助手交互的数据片段。考虑到在多数时间里驾驶员可能未与智能助手进行交互,文中实施了一套精确的初步识别和数据分割流程。

首先,采用了基于声音检测的算法来识别数据中存在有效语音交流的时间片段。该算法通过分析音频信号中的声音活动,高效地判断并标记了包含有效交流的时段。利用这些标记,能够精确地切割出有用的交互数据片段,为后续分析奠定了基础。

完成自动算法处理后,还引入了关键的人工审核环节。该环节的目的是验证并确保所有自动选定的数据片段确实包含了有效的驾驶员与智能助手间的交互。这一步骤在整个数据处理流程中发挥着至关重要的作用,因为人工审核可以有效识别并纠正自动算法可能遗漏或误判的细节,从而显著提升了最终数据集的质量和可靠性。

1.3 数据标注

本研究中,数据标注是一个关键环节,旨在确保

分析和研究的准确性。标注过程主要由多个经验丰富的内部供应商标注人员负责,他们遵循严格的标准对每条数据进行详细的人工标注。标注的核心目的是区分音频中的人声属于对智能助手的交流(人机交流)还是车内人员的对话(人人交流)。具体来说,如果一段音频包含的人声是针对智能助手的,则该片段被标记为“人机交流”。相反,如果人声非针对智能助手,则被标记为“人人交流”。这种区分对于后续分析驾驶员与智能助手之间的交互模式至关重要。

除了音频内容的标注外,还注重对驾驶员的非语言行为进行记录,包括其语言输入、情绪表达以及面部朝向等,如图4所示。此外,车内人数也被纳入标注范围,以提供更全面的交互环境信息。为了保证标注结果的一致性和准确性,每条数据都由4位专业的算法工程师进行独立标注,只保留所有标注人员一致同意的结果,以确保数据的高度可靠性。对于标注结果出现差异的情况,采取反复校对的方式,包括但不限于审查当前对话的上下文信息,以此来解决可能的歧义和确保标注的准确性。

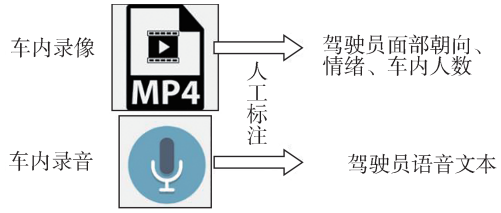

图4　标注过程

构建的数据集中,每条数据都是以纯文本形式呈现,包含两个主要部分:内容和标签。内容部分详细记录了驾驶员的语言交流和其他相关信息,而标签则用于表示这些语言交流的类型。举例来说,典型的数据如表1所示。

表1　构建数据示例

示例	内容				标签
	语音输入	车内人数	面部朝向	情绪	
1	请打开导航去北京路	1	正向前方	中性	人机指令
2	那家餐厅的菜真不错	2	朝向副驾驶	快乐	人人交流

示例1中的内容部分清楚地记录了驾驶员向智能助手发出的指令,而标签“人机指令”则表明这是一条驾驶员对智能助手的直接命令。示例2中,内容部分展示了驾驶员的一段言论,而标签“人人交流”则表明这段话是驾驶员间的普通对话,而非对智能助手的指令。

1.4　数据划分

构建的车内驾驶员意图判断数据集,共包含3 110条训练数据和1 055条验证数据,如表2所示。这些数据涵盖了广泛的车载智能助手交互意图,如导航、音乐播放、闲聊等场景,同时也包括了众多车内人员间的日常对话。训练集包含了多种类型的交互实例,覆盖了从人机指令到人人交流的各类交互场景和驾驶员行为,旨在确保模型能够在多变且复杂的真实世界环境中进行高效学习。

表2　拒识标注数据集分布情况

数据	人人交流(拒识)		人机指令(非拒识)		总计
	数量	占比	数量	占比	
训练数据	1 763	56.7%	1 347	43.3%	3 110
验证数据	583	55.3%	472	45.7%	1 055

而对于验证集,文中采取了更加严格的数据标注标准,以保障数据的一致性和准确性达到最高标准。这部分数据的主要目的是对模型在处理未曾接触过的数据时的表现进行全面评估,从而验证其在各种情境下的泛化能力和预测精确度。通过这样的划分,数据集不仅为模型提供了丰富的学习材料,同时也为评估模型的实际应用效果提供了强有力的测试平台。

2　模型及实验

本节将对实验所用的模型及微调模型所使用的方法进行介绍,然后介绍对数据集的重新编写设计,并对微调后的模型进行验证。

2.1　基础模型

随着计算能力进步和大语言模型的快速发展,全球范围内涌现出众多创新的语言模型。特别是,由清华大学提出的通用语言模型(GLM),现已在国内的多个机构和企业中得到广泛应用。GLM预训练框架通过以下4个方面巧妙地融合了多种技术:其一它采用自编码思想,在输入文本中随机删除连续的tokens;其二它融入自回归思想,通过顺序重建连续tokens,模型在预测缺失tokens时,可以同时访问已损坏的文本和之前预测的文本片段;其三它结合了span shuffling和二维位置编码技术;其四通过调整缺失spans的数量和长度,该模型的自回归空格填充任务可以用于条件生成和非条件生成任务的预训练。

本研究采用的ChatGLM2-6B是一款基于GLM框架开发的开源中英双语对话模型。它不仅继承了其前代模型在对话流畅性和低部署门槛方面的优势,而且还引入了更加强大的性能和更高效的推理能力。ChatGLM2-6B采用GLM的混合目标函数,并经过了1.4万亿中英语言标识符的大规模预训练,以及针对人类偏好的对齐训练,使其在对话生成领域表现出色。这个拥有百亿级参数的模型不仅在逻辑推理方面展现出准确性,而且在多方面达到了接近人类认知水平的表现。

2.2 微调方法

本文采用了"P-Tuning v2"作为微调方法。在自然语言处理领域,传统的模型训练方法如微调Fine-tuning一直是主流。然而,随着模型规模的增大,传统方法在资源消耗方面的局限性日益明显。为了解决这一问题,研究者们提出了基于提示的优化方法,即提示调整Prompt Tuning。这种方法的核心在于在训练过程中固定大部分预训练模型的权重,仅调整少量与特定任务相关的提示参数。这种方法的优势在于显著降低了模型微调时的资源消耗,同时保持了不错的任务性能。由Liu等[13]在2021年提出的"P-Tuning v2"在与其他微调方法保持相似性能的同时,大幅减少了所需调整的参数量,仅为0.1%~3%。

2.3 实验参数

本文的实验均在系统Ubuntu18.04、显卡Tesla V100S的服务器上通过Python3.10版本完成。在微调预训练模型的实验中,具体的参数设置根据数据语料的特点、模型的选择和进行实验的多次调参对比所得。其中学习率采用默认学习率,输入文本的最大长度设置为200,故训练的批次大小(batch size)设置为4,最大训练轮次(epochs)为3 000轮,每1 000轮次保存一次模型参数,以便后续分析和模型恢复。

2.4 实验设计与结果分析

2.4.1 提示Prompt设计

在大型预训练语言模型的训练和微调中,提示(Prompt)至关重要。众多的开源预训练的大模型,如GPT系列、BERT等,都是在广泛的数据集上训练,积累了大量的语言知识,具有强大的语言基础理解能力。Prompt的作用在于将这些模型的通用知识桥接到特定的任务上。通过精心设计的Prompt,模型能够将其广泛的预训练知识应用于具体的NLP任务,如文本分类、情感分析、问答系统等。Prompt提供了一种引导模型理解任务需求的方式,使模型即使在较少数据支持的情况下也能作出较为准确的判断。

设计Prompt是提高模型在特定任务上表现的关键。Prompt设计主要分为两种类型:生成式和抽取式。生成式Prompt的核心思想是让模型直接生成任务的输出,这种方法的优势在于其灵活性和创造性,特别适用于开放式问题、内容创作或任何需要模型展现创造力的场景。抽取式Prompt将答案直接嵌入到输入文本中,要求待生成的部分被置于输入文本的末尾,这种方法的优势在于它通常能提供更精确和具体的答案,尤其适用于信息检索或具体事实的查找。

根据上述描述,本研究对前文所述构建的数据集进行了编写,示例如表3所示。

表3 各式Prompt设计

序号	Prompt设计	类型
1	{"content": "请分析以下语音输入是否为对车辆系统的直接指令;<数据内容>", "summary": "是"}	抽取式
2	{"content": "请分析以下语音输入是否为与其他人的交流内容;<数据内容>", "summary": "否"}	抽取式
3	{"content": "根据以下语境,判断驾驶员的语音输入属于哪种类型;<数据内容>", "summary": "人机指令"}	生成式
4	{"content": "分析语音输入是人机指令和人人交流;<数据内容>", "summary": "人机指令"}	抽取式
5	{"content": "假设你是一个智能车机交互助手,判断驾驶员的语音输入是属于人机指令还是人人交流;<数据内容>", "summary": "人机指令"}	抽取式
6	{"content": "假设你是一个智能车机交互助手,判断用户说的话是属于人机指令还是人人交流。人机指令是驾驶员对智能车机助手的指令或询问,而人人交流是车内人员的交流对话;<数据内容>", "summary": "人机指令"}	抽取式

上述的各种Prompt包含生成式和抽取式,且在Prompt的长短上也做了适当的调整,由于使用的P-Tuning v2的微调技术,不同任务对提示长度有着不同的需求,指令的长短也会影响到模型训练的效果和测试的效果。

2.4.2 数据模态设计

在当前的车载智能系统中,纯语言输入的识别机制尽管便利,但在复杂环境和多变交互场景下存在局限性。特别是在复杂的车内环境,或是驾驶员意图分辨困难的情况下,纯语言的识别机制往往难以维持高精度,导致误识率的增加,影响交互体验和车辆安全操作。

这种局限性不仅降低了用户的交互体验,使得驾驶员可能需要重复或修正指令,而且在某些关键时刻可能对车辆的安全操作构成影响。因此,弥补纯语言识别机制的这些不足显得尤为重要。通过引入多模态数据,如面部表情、情绪状态等非语言信息,可以有效地提升识别机制的准确性和鲁棒性。其中多模相对于单模系统,ACC和F1均提升6%左右,而FRR降低了7%。这些多维度信息的结合为模型提供了更全面的上下文,可以让大模型更加充分地理解分析语音输入所处的车内情景,使其能够在复杂环境中更准确地理解驾驶员的真实意图,从而提高交互的精准度,确保车辆操作的安全性。

因此在上述提示Prompt设计的基础上,对数据内容进行了更改,内容划分如表4所示。

表4　数据内容划分

序号	类型	数据内容
1	多模	语音输入:请打开导航去北京路。车内人数:1人。面部朝向:正向前方。情绪:平静
	单模	语音输入:请打开导航去北京路
2	多模	语音输入:我们快到目的地了吗。车内人数:2人。面部朝向:朝向副驾驶。情绪:高兴
	单模	语音输入:我们快到目的地了吗
3	多模	语音输入:查询明天的天气预报。车内人数:1人。面部朝向:正向前方。情绪:平静
	单模	语音输入:查询明天的天气预报

本文主要使用准确率(ACC)、F1值和误识别率(FRR)对多模态拒识任务进行评估,图5和表5展示了在对数据集不同的处理情况下对大模型进行微调后,在测试集上的各项指标。Baseline是没有微调的大模型ChatGLM2的实验结果,准确性远低于微调之后的,且输出的结果也是杂乱无章的,不利于座舱系统的整体运行。其中多模数据有着较大的提升,提示Prompt设计中抽取式的结果也较好于生成式的。结果也验证了在模态类型对比上,多模数据取得了更好的结果,提示Prompt设计也对模型的训练产生了正面的影响。

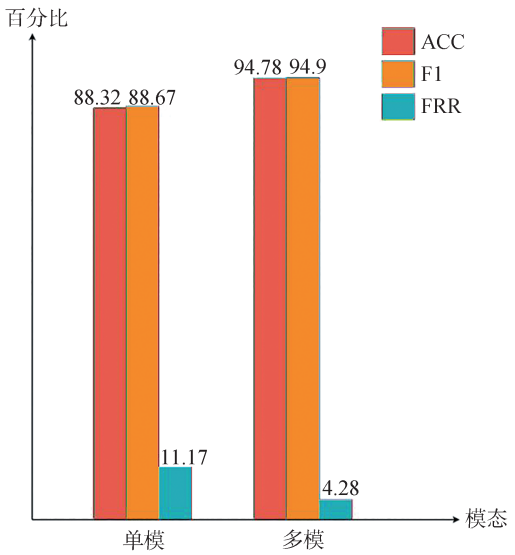

图5　模态平均效果对比

表5　大模型微调的相关对比实验结果

模态类型	Prompt设计序号	驾驶员语音输入拒识/%		
		ACC	F1	FRR
单模(未微调)		48.5		
多模(未微调)		56.0		
单模	1	86,6	86.9	12.1
	2	87.0	87.6	12.1
	3	87.3	87.9	11.6
	4	88.3	88.8	10.8
	5	89.5	90.4	10.2
	6	89.5	90.4	10.2
多模	1	93.8	93.7	6.2
	2	94.2	94.1	5.8
	3	94.6	94.8	3.8
	4	94.9	94.9	3.9
	5	**95.6**	**96.0**	**2.8**
	6	95.6	95.9	3.2

3　结论

本文提出了一种基于开源大模型和车机交互专属拒识数据集的多模拒识模型。首先,针对汽车座舱场景下的人机交互语音拒识进行了需求分析,并简要介绍了使用专属车机交互拒识数据集微调开源大模型的方法。接着,详细叙述了专属拒识数据集的制作过程,并对数据集的案例进行了介绍。最后,介绍了大模型的基础模型、微调的基本技术,以及如何对数据集进行处理以提高大模型的判断效果,包括多组对照实验的实验设置及其结果分析。实验数据表明,本文所构建的基于开源大模型和车机交互专属拒识数据集的多模拒识模型,在性能上相较于单模具有显著优势。

参考文献

[1] MALLIDI S H, MAAS R, GOEHNER K, et al. Device-directed utterance detection[J].ArXiv preprint arXiv:1808.02504,2018.

(下转第429页)

ate and cooperative model predictive control for energy-efficient truck platooning of heterogeneous fleets[J]. IEEE Transactions on Vehicular Technology, 2022, 71(6): 5755-5769.

[21] ZHANG Yu, BAI Yu, WANG Meng, et al. Cooperative adaptive cruise control with robustness against communication delay: an approach in the space domain[C]. American Control Conference, 2022.

[22] XU Liwei, ZHUANG Weichao, YIN Guodong, et al. Modeling and robust control of heterogeneous vehicle platoon on curved road subject to disturbances and delays[J]. IEEE Transactions on Vehicular Technology, 2019, 68(99): 11551-11564.

[23] XU Liwei, ZHUANG Weichao, YIN Guodong, et al. Stable longitudinal control of heterogeneous vehicular platoon with disturbances and information delays[J]. IEEE Access, 2018, 6: 69794-69806.

[24] YU Guokuan, WONG Pak Kin, HUANG Wei, et al. Distributed adaptive consensus protocol for connected vehicle platoon with heterogeneous time-varying delays and switching topologies[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(10): 17620-17631.

[25] TIAN Bin, WANG Guanqun, XU Zhigang, et al. Communication delay compensation for string stability of CACC system using LSTM prediction [J]. Vehicular Communications, 2021, 29: 100333.

[26] 王雪彤,罗禹贡,江发潮,等.纯电动商用车异质队列的多目标控制[J].汽车工程, 2020, 42(4): 505-512,559.
WANG Xuetong, LUO Yugong, JIANG Fachao, et al. Multi-target control for heterogeneous platoon of battery electric commercial vehicle[J]. Automotive Engineering, 2020, 42 (4): 505-512,559.

[27] XU Hao, TU Ran, LI Tiezhu, et al. Interpretable bus energy consumption model with minimal input variable considering powertrain types [J]. Transportation Res Part D: Transportation Environment, 2023, 119: 103742.

[28] DUNBAR W B, CAVENEY D S. Distributed receding horizon control of vehicle platoons: stability and string stability [J]. IEEE Transactions on Automatic Control, 2012, 57 (3): 620-633.

[29] MA Hao, CHU Liang, GUO Jianhua, et al. Cooperative adaptive cruise control strategy optimization for electric vehicles based on SA-PSO with model predictive control [J]. IEEE Access, 2020: 3043370.

[30] LEFEVRE S, VASQUEZ D, LAUGIER C. A survey on motion prediction and risk assessment for intelligent vehicles[J]. Robomech Journal, 2014, 1: 1-14.

(上接第417页)

[2] HUANG C W, MAAS R, MALLIDI S H, et al.A study for improving device-directed speech detection toward frictionless human-machine interaction[C]. INTERSPEECH. [S.l.:s.n.],2019: 3342-3346.

[3] GILLESPIE K, KONSTANTAKOPOULOS I C, GUO X, et al.Improving device directedness classification of utterances with semantic lexical features[C].ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).[S.l.:s.n.],2020:7859-7863.

[4] YAMAGATA T, TAKIGUCHI T, ARIKI Y.System request detection in human conversation based on multi-resolution gabor wavelet features [C]. Tenth Annual Conference of the International Speech Communication Association.[S.l.:s.n.], 2009.

[5] REICH D, PUTZE F, HEGER D, et al.A real-time speech command detector for a smart control room[C].Twelfth Annual Conference of the International Speech Communication Association.[S.l.:s.n.],2011.

[6] SHRIBERG E, STOLCKE A,HAKKANI-TÜR D, et al.Learning when to listen:detecting system-addressed speech in human-human-computer dialog[C].Thirteenth Annual Conference of the International Speech Communication Association. [S.l.:s.n.],2012.

[7] LEE H, STOLCKE A, SHRIBERG E. Using out-of-domain data for lexical addressee detection in human-human-computer dialog [C].Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies.[S.l.:s.n.],2013:221-229.

[8] PAEK T, HORVITZ E, RINGGER E K.Continuous listening for unconstrained spoken dialog [C]. INTERSPEECH. [S. l.: s. n.] , 2000:138-141.

[9] DOWDING J,ALENA R,CLANCEY W J,et al.Are you talking to me? dialogue systems supporting mixed teams of humans and robots[C].AAAI Fall Symposium: Aurally Informed Performance. [S.l.:s.n.],2006:22-27.

[10] NOROUZIAN A, MAZOURE B, CONNOLLY D, et al. Exploring attention mechanism for acoustic-based classification of speech utterances into system-directed and non-system-directed [C]. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP).[S.l.:s.n.], 2019:7310- 7314.

[11] TONG X, HUANG C W, MALLIDI S H, et al. Streaming Res LSTM with causal mean aggregation for device-directed utterance detection[C].2021 IEEE Spoken Language Technology Workshop (SLT).[S.l.:s.n.],2021:659-664.

[12] HOWARD A, ZHU M.Fine-tuning large pre-trained models for language understanding[J]. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019.

[13] LIU Xiao, JI Kaixuan. P-tuning v2: prompt tuning can be comparable to fine-tuning universally across scales and tasks[J]. arXiv preprint arXiv:2110.07602, 2021.