(19)国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 118486333 A
(43)申请公布日 2024.08.13

(21)申请号 202410576381.1

(22)申请日 2024.05.10

(71)申请人 上海声通信息科技股份有限公司
地址 200240 上海市闵行区顾戴路2337号维璟中心G栋
申请人 上海交通大学

(72)发明人 薛广涛 张力允 鲁昱 汤敬华 张宗振 方楠 郑波 丁典 陈奕超

(74)专利代理机构 上海科盛知识产权代理有限公司 31225
专利代理师 廖程

(51)Int.Cl.
G10L 25/63(2013.01)
G10L 15/18(2013.01)
G10L 15/26(2006.01)
G10L 17/04(2013.01)

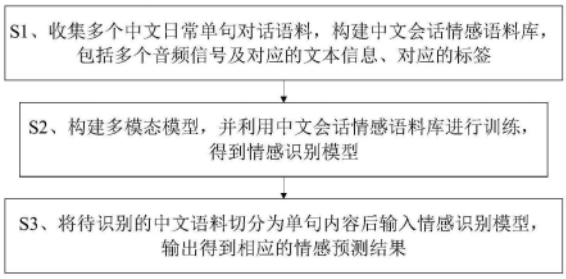
权利要求书1页 说明书7页 附图1页

(54)发明名称

一种基于声纹和文本的多模态情感识别方法

(57)摘要

本发明涉及一种基于声纹和文本的多模态情感识别方法,包括:收集多个中文日常单句对话语料,构建中文会话情感语料库,包括多个音频信号及对应的文本信息、对应的标签;构建多模态模型,并利用中文会话情感语料库进行训练,得到情感识别模型;将待识别的中文语料切分为单句内容后输入情感识别模型,输出得到相应的情感预测结果。与现有技术相比,本发明创建一个全新的中文会话情感语料库,补充了中文语料在情绪识别领域中的不足;并设计多模态情感识别模型,用于处理声学数据和文本数据(包含词嵌入和预训练的BERT嵌入),使用共注意结构进行多模态特征融合,能够有效提高情感识别的准确性和稳定性。



CN 118486333 A

1.一种基于声纹和文本的多模态情感识别方法,其特征在于,包括以下步骤:

S1、收集多个中文日常单句对话语料,构建中文会话情感语料库,包括多个音频信号及对应的文本信息、对应的标签;

S2、构建多模态模型,并利用中文会话情感语料库进行训练,得到情感识别模型;

S3、将待识别的中文语料切分为单句内容后输入情感识别模型,输出得到相应的情感预测结果。

2.根据权利要求1所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述步骤S2中多模态模型包括特征提取模块、融合模块、自注意力模块和预测模块,所述特征提取模块用于从中文语料中提取音频特征、文本特征以及编码特征;

所述融合模块用于将音频特征、文本特征以及编码特征进行融合处理,以得到融合特征;

所述自注意力模块和预测模块用于对融合特征进行全局处理和情感预测。

3.根据权利要求2所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述特征提取模块包括音频特征提取单元、文本特征提取单元和编码单元,所述音频特征提取单元用于从音频信号中提取音频特征;

所述文本特征提取单元用于从音频信号中提取文本特征;

所述编码单元用于提取文本的编码特征。

4.根据权利要求3所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述音频特征提取单元的工作过程为:应用短时傅里叶变换STFT对音频信号进行处理,提取出梅尔频谱图,以提供音频信号在不同频段随时间变化的能量的可视化表示。

5.根据权利要求3所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述文本特征提取单元的工作过程为:利用自动语音识别ASR技术将音频信号转换为文本信息,对文本信息进行处理,使用文本嵌入技术学习文本特征。

6.根据权利要求3所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述编码单元具体是利用预训练的中文BERT模型提取文本的编码特征。

7.根据权利要求2所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述融合模块包括第一融合单元和第二融合单元,所述第一融合单元用于将音频特征与文本特征进行融合,并结合共注意力单元实现模态间的信息交互,输出得到时序特征;

所述第二融合单元用于将时序特征与编码特征进行融合,输出得到融合特征。

8.根据权利要求7所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述第一融合单元采用元素相加的方式针对音频特征与文本特征进行融合处理。

9.根据权利要求7所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述第二融合单元采用拼接的方式针对时序特征与编码特征进行融合处理。

10.根据权利要求2~10任一所述的一种基于声纹和文本的多模态情感识别方法,其特征在于,所述预测模块包括两层多层感知机。

一种基于声纹和文本的多模态情感识别方法

技术领域

[0001] 本发明涉及人工智能情绪识别技术领域,尤其是涉及一种基于声纹和文本的多模态情感识别方法。

背景技术

[0002] 音频数据不仅包含着词汇和语言,还蕴含着人类经历和情感的错综复杂的图景,音频数据展现出惊人多样的应用,基于声纹的情感识别对于与人沟通以及人机交互应用至关重要。比如在呼叫中心,可通过及时反馈客户情绪来选择合适的策略,或依赖客户情绪水平进行业务评估,同时,软件应用可通过实时监控用户情绪来调整和增强用户体验,以实施适当的行为[1]。

[0003] 在情感识别领域,现有的方法[2,3,4]大多利用神经网络来完成任务。这些神经网络模型主要在高质量的英文数据集上展现了它们的有效性(例如,IEMOCAP[5],MELD[6],RAVDESS[7])。然而,在情感识别的背景下,中文语料库数据的可用量相对有限,并且现有的中文数据集[8,9]存在以下问题:

[0004] 1).缺乏多样性和普适性:先前数据集中声音信息的收集通常依赖于少数特定的专业读者或专业演员。因此,一方面,数据集中只包含了少数人的发音信息和声纹特征。另一方面,其专业化的发音特点与日常生活不够接近。这可能导致使用这些数据集训练的模型缺乏泛化能力。

[0005] 2).文本数量有限:以CASIA[8]为例,该数据集是通过不同的读者朗读相同的句子编制而成。因此,数据集中没有足够的文本数据支持模型对文本特征提取模块的训练。

[0006] 3).无法进行多模态融合:专业读者可以以不同的情感朗读相同的句子,以获取与不同情感相对应的语音数据。相同的文本内容将具有不同的情感标签,导致无法从文本信息中挖掘情感数据。

[0007] 此外,语音情感识别在机器学习和语音社区中已经被广泛研究。与当前的研究方法一致,学者们从音频数据中提取特征见解,随后在一系列分类器中运用这些见解,包括:隐马尔可夫模型[10]、卷积递归网络[11]、支持向量机(SVM)[12]、层次二叉决策树[13]、高斯混合[14]、神经网络[15]。

[0008] 比如徐等人[16]提出了一种基于注意力的网络,用于对齐文本和音频信息,并进行特征提取。Yoon[17,18]提出了一种具有突破性的深度双重循环编码器模型,无缝融合文本数据和音频信号,该模型采用一对递归神经网络(RNNs)全面编码信息。Delbrouck等[19]提出了一种基于Transformer的联合编码模型,称为UMNOS,用于单句情感识别和情感分析。但上述工作依赖上下文来提供额外信息,以纠正和推断从数据中提取的情感内容,从单句音频数据中挖掘和分析情感信息可能面临更大的挑战,难以将来自外部世界的知识融入到情感识别任务中,无法全面深入地整合不同模态的信息,不利于准确稳定地实现情感识别。

[0009] 现有参考文献如下:

[0010] [1]Mingmin Zhao,Fadel Adib,and Dina Katabi,“Emotion recognition using

wireless signals,"in Proceedings of the 22nd annual international conference on mobile computing and networking,2016,pp.95-108.

[0011] [2]Sreyan Ghosh,Utkarsh Tyagi,S Ramaneswaran,Harshvardhan Srivastava, and Dinesh Manocha,"Mmer:Multimodal multi-task learning for speech emotion recognition,"arXiv preprint arXiv:2203.16794,2022.

[0012] [3]Itai Gat,Hagai Aronowitz,Weizhong Zhu,Edmilson Morais,and Ron Hoory,"Speaker normalization for self-supervised speech emotion recognition," in ICASSP 2022.IEEE,2022,pp.7342-7346.

[0013] [4]Yingzhi Wang,Abdelmoumene Boumadane,and Abdelwahab Heba,"Afine-tuned wav2vec 2.0/Hubert benchmark for speech emotion recognition,speaker verification and spoken language understanding,"arXiv preprint

[0014] arXiv:2111.02735,2021.

[0015] [5]Amir Zadeh,Paul Pu Liang,Soujanya Poria,Prateek Vij,Erik Cambria, and Louis-Philippe Morency,"Multiattention recurrent network for human communication comprehension,"in AAAI,2018,vol.32.

[0016] [6]Soujanya Poria,Devamanyu Hazarika,Navonil Majumder,Gautam Naik, Erik Cambria,and Rada Mihalcea,"Meld:A multimodal multi-party dataset for emotion recognition in conversations,"arXiv preprint arXiv:1810.02508,2018.

[0017] [7]Steven R Livingstone and Frank A Russo,"The ryerson audio-visual database of emotional speech and song(ravdess):A dynamic,multimodal set of facial and vocal expressions in north american english,"PloS one,vol.13,no.5, pp.e0196391,2018.

[0018] [8]CAS,"Casia:chinese affective corpus,"2005.

[0019] [9]Yixiong Pan,Peipei Shen,and Liping Shen,"Speech emotion recognition using support vector machine,"2012.

[0020] [10]Bjorn Schuller,Gerhard Rigoll,and Manfred Lang,"Hidden markov model-based speech emotion recognition,"in 2003 IEEE International Conference on Acoustics,Speech,and Signal Processing,2003.Proceedings.(ICASSP'03).Ieee, 2003,vol.2,pp.II-1.

[0021] [11]George Trigeorgis,Fabien Ringeval,Raymond Brueckner,Erik Marchi, Mihalis A Nicolaou,Bjorn Schuller,¨and Stefanos Zafeiriou,"Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network,"in 2016 ICASSP.IEEE,2016,pp.5200-5204.

[0022] [12]Thapanee Seehapoch and Sartra Wongthanavasu,"Speech emotion recognition using support vector machines,"in 2013 KST.IEEE,2013,pp.86-91.

[0023] [13]Chi-Chun Lee,Emily Mower,Carlos Busso,Sungbok Lee,and Shrikanth Narayanan,"Emotion recognition using a hierarchical binary decision tree approach,"Speech Communication,vol.53,no.9-10,pp.1162-1171,2011.

[0024] [14]Moataz MH El Ayadi,Mohamed S Kamel,and Fakhri Karray,"Speech

emotion recognition using gaussian mixture vector autoregressive models,"in 2007ICASSP.IEEE,2007,vol.4,pp.IV-957.

[0025] [15]Andre Stuhlsatz,Christine Meyer,Florian Eyben,Thomas Zielke,Gunter Meier,and Bjorn Schuller,"Deep neural networks for acoustic emotion recognition:Raising the benchmarks,"in 2011 IEEE international conference on acoustics,speech and signal processing(ICASSP).IEEE,2011,pp.5688-5691.

[0026] [16]Haiyang Xu,Hui Zhang,Kun Han,Yun Wang,Yiping Peng,and Xiangang Li,"Learning alignment for multimodal emotion recognition from speech," Proc.Interspeech 2019,pp.3569-3573,2019.

[0027] [17]Seunghyun Yoon,Seokhyun Byun,and Kyomin Jung,"Multimodal speech emotion recognition using audio

[0028] and text,"in 2018SLT.IEEE,2018,pp.112-118.

[0029] [18]Seunghyun Yoon,Seokhyun Byun,Subhadeep Dey,and Kyomin Jung,"Speech emotion recognition using multi-hop attention mechanism,"in ICASSP 2019.IEEE,2019,pp.2822-2826.

[0030] [19]Jean-Benoit Delbrouck,Noe Tits,Mathilde Brousmiche,and Stephane Dupont,"A transformer-based joint-encoding for emotion recognition and sentiment analysis,"ACL 2020,p.1,2020.

发明内容

[0031] 本发明的目的就是为了克服上述现有技术存在的缺陷而提供一种基于声纹和文本的多模态情感识别方法,通过创建一个全新的中文会话情感语料库,并设计多模态情感识别模型,能够有效提高情感识别的准确性和稳定性。

[0032] 本发明的目的可以通过以下技术方案来实现:一种基于声纹和文本的多模态情感识别方法,包括以下步骤:

[0033] S1、收集多个中文日常单句对话语料,构建中文会话情感语料库,包括多个音频信号及对应的文本信息、对应的标签;

[0034] S2、构建多模态模型,并利用中文会话情感语料库进行训练,得到情感识别模型;

[0035] S3、将待识别的中文语料切分为单句内容后输入情感识别模型,输出得到相应的情感预测结果。

[0036] 进一步地,所述步骤S2中多模态模型包括特征提取模块、融合模块、自注意力模块和预测模块,所述特征提取模块用于从中文语料中提取音频特征、文本特征以及编码特征;

[0037] 所述融合模块用于将音频特征、文本特征以及编码特征进行融合处理,以得到融合特征;

[0038] 所述自注意力模块和预测模块用于对融合特征进行全局处理和情感预测。

[0039] 进一步地,所述特征提取模块包括音频特征提取单元、文本特征提取单元和编码单元,所述音频特征提取单元用于从音频信号中提取音频特征;

[0040] 所述文本特征提取单元用于从音频信号中提取文本特征;

[0041] 所述编码单元用于提取文本的编码特征。

[0042] 进一步地,所述音频特征提取单元的工作过程为:应用短时傅里叶变换(short-time Fourier transform,STFT)对音频信号进行处理,提取出梅尔频谱图,以提供音频信号在不同频段随时间变化的能量的可视化表示。

[0043] 进一步地,所述文本特征提取单元的工作过程为:利用自动语音识别(Automatic Speech Recognition,ASR)技术将音频信号转换为文本信息,对文本信息进行处理,使用文本嵌入技术学习文本特征。

[0044] 进一步地,所述编码单元具体是利用预训练的中文BERT(Bidirectional Encoder Representations from Transformers,来自变换器的双向编码器表征量)模型提取文本的编码特征。

[0045] 进一步地,所述融合模块包括第一融合单元和第二融合单元,所述第一融合单元用于将音频特征与文本特征进行融合,并结合共注意力单元实现模态间的信息交互,输出得到时序特征;

[0046] 所述第二融合单元用于将时序特征与编码特征进行融合,输出得到融合特征。

[0047] 进一步地,所述第一融合单元采用元素相加的方式针对音频特征与文本特征进行融合处理。

[0048] 进一步地,所述第二融合单元采用拼接的方式针对时序特征与编码特征进行融合处理。

[0049] 进一步地,所述预测模块包括两层多层感知机(Multilayer Perceptron,MLP)。

[0050] 与现有技术相比,本发明具有以下优点:

[0051] 本发明通过收集多个中文日常单句对话语料,以构建包括多个音频信号及对应文本信息、对应标签的中文会话情感语料库,以专门用于中文情感识别任务,该中文会话情感语料库具有丰富的声纹信息和文本信息,能够适应多模态融合处理过程,利用该多样性的中文数据集,能够增强后续情感识别模型的泛化能力,从而在日常对话等实际场景中更为有效。

[0052] 本发明搭建一个多模态模型,并结合中文会话情感语料库训练得到情感识别模型,该多模态模型包括特征提取模块、融合模块、自注意力模块和预测模块,利用特征提取模块从中文语料中提取音频特征、文本特征以及编码特征;利用融合模块将音频特征、文本特征以及编码特征进行融合处理;再利用自注意力模块和预测模块对融合特征进行全局处理和情感预测。由此能够充分整合不同模态(声学、文本、外部知识)的信息,并且能够对融合后的信息进行全局处理和情感预测,从而增强模型的表达能力。

[0053] 本发明在特征提取模块中,设计音频特征提取单元、文本特征提取单元和编码单元,其中编码单元利用预训练的中文BERT模型提取文本的编码特征,以作为外部知识,使得模型能够更好地将来自外部世界的知识融入到情感识别任务中。

[0054] 本发明在融合模块中设计第一融合单元和第二融合单元,通过引入两轮融合的设计,首先融合音频和文本信息,然后再融合外部知识,使得模型能够更全面、深入地整合不同模态的信息。这种阶段化设计提高了模型对时序和语义结构的敏感性,使其更适应复杂多变的情感表达场景。

[0055] 本发明在第一融合单元中,考虑到音频数据和词嵌入特征时序相似的情况,采用了直接的元素相加方式,以增强它们的时序结构;在第二融合单元中,考虑到与BERT编码特

征时序不同的情况,采用了拼接方式,以保留更多信息。

[0056] 本发明引入共注意力单元,以实现模态间的信息交互,使得每个模态能够更好地利用来自其他模态的知识。

[0057] 本发明引入自注意力层和两层MLP对融合后的信息进行全局处理和情感预测,增强了模型的表达能力。这种设计使得模型能够更好地捕捉多模态信息的复杂关系,提高了情感识别的准确性和鲁棒性。

附图说明

[0058] 图1为本发明的方法流程示意图;

[0059] 图2为实施例中搭建的多模态模型流程示意图。

具体实施方式

[0060] 下面结合附图和具体实施例对本发明进行详细说明。

[0061] 实施例

[0062] 如图1所示,一种基于声纹和文本的多模态情感识别方法,包括以下步骤:

[0063] S1、收集多个中文日常单句对话语料,构建中文会话情感语料库,包括多个音频信号及对应的文本信息、对应的标签;

[0064] S2、构建多模态模型,并利用中文会话情感语料库进行训练,得到情感识别模型;

[0065] S3、将待识别的中文语料切分为单句内容后输入情感识别模型,输出得到相应的情感预测结果。

[0066] 本实施例应用上述技术方案,首先创建了一个名为VCEMO的大型中文会话情感语料库,专门用于单句中文情感识别任务。该数据集包含了日常生活中的单句对话,具有以下几个优势:

[0067] 1)丰富的声纹信息:数据集包含来自100多人的日常语音数据,包括广泛的中文发音口音和口语特征。

[0068] 2)丰富的文本信息:数据集的文本内容专门来自日常生活中的自发对话。因此,这些文本之间存在实质性差异,并且信息丰富。

[0069] 3)适应多模态融合:由于数据来源于日常对话,个体自然地使用各种文本表达方式来传达内在的情感,从而可有效地利用音频信号和文本信息的多模态融合进行情感识别任务。充分利用这个多样性的中文数据集,可增强模型的泛化能力,使其在日常对话等实际场景中更为有效。

[0070] 本实施例创建的中文情感识别数据集,是一个包含来自100多个个体的7477个音频信号样本的中文日常对话语料库。

[0071] 之后搭建多模态模型,如图2所示,该模型具有三种输入模态,对于声学输入,利用梅尔频谱图,通过对音频信号应用短时傅里叶变换(STFT)生成。梅尔频谱图以可视化的方式呈现了音频信号在不同频段随时间变化的能量,频率轴经过调整以更好地匹配人类听觉感知。

[0072] ASR是一种将音频数据转换为文本数据的技术,便于机器转录和理解口语。本方案使用ASR模块从音频信号中提取已识别的文本。对于文本输入,使用文本嵌入直接学习文本

特征。同时,结合预训练的BERT从外部知识中提取转录特征。

[0073] 在获取音频信号的梅尔频谱图后,应用经典的Conv-BatchNorm-ReLU结构来提取时间和频率维度的特征。然后,应用一个LSTM层来提取时间维度中更深层次的特征。此外,单词嵌入在每个时间槽中具有更好的时间结构且更为直观。因此,在使用1D卷积层整合整个时间线的信息之前,对单词嵌入应用了LSTM。从BERT中提取的特征则是一个768维的向量,由于它已经结构良好且包含丰富的信息,因此应用了一个线性层来修改其大小,以便进行后续的多模态融合和信息压缩。

[0074] 考虑到存在三种模态,本方案设计两轮融合来全面结合从这些不同模态提取的所有信息,并且确定融合的顺序也是需要考虑的一个重要因素。在本方案设计的模型中,首先融合音频特征和单词嵌入特征,二者相似的时间结构使它们适合进行初始融合,因为这个过程通过利用它们的共同特点来增强时间维度,以放大共有信息并补偿某个模态特有的缺失数据。随后,将初始融合后有时间结构的特征与BERT编码的特征融合,以将外部知识从外界整合到数据集内的知识中。

[0075] 且在每次融合中,均有两个阶段:用另一种模态的知识从一种模态中提取附加特征,然后将这些额外提取的特征合并成单一的表示。具体的,在第一阶段,采用了共注意力层来向每个模态传递另一种模态的存在。共注意力层的结构采用编码器-解码器结构来堆叠多层注意力模块。在共注意力层中,第一模态仅使用自注意力来从自身提取更深层次的信息。随后,第二模态经历一个自注意力操作,在此期间进行了引导注意力步骤,以在考虑两种模态的同时提取更多信息。与简单地在引导注意力的输入中使用另一模态相同深度的自注意力输出相比,利用自注意力层的最终输出可以提供更丰富的信息和更准确的引导。自注意力和引导注意力都基于注意机制。注意模块有助于构建在语音过程中整个时间跨度的整体视角。在自注意力中,所有的q、k和v来自同一模态。然而,在引导注意力中,v和k来自同一模态,而q来自另一个不同的模态。

[0076] 两轮融合的第一阶段是相同的,但在第二阶段则存在分歧:考虑到时间结构的相似性,对于从音频数据和单词嵌入中融合的特征,采用直观的逐元素加法,这种方法增强了它们的时间结构,并相对于串联减少了特征大小。在第二次融合中,特征是不相似的,缺乏共享的时间结构,这导致使用逐元素加法时信息丢失和无序。因此采用串联来保留更多信息,这对于有效利用数据集内知识和外部世界知识至关重要。在最终的融合之后,则应用额外的自注意力来全面处理所有模态的集体信息,并使用两层MLP进行预测。

[0077] 综上可知,上述多模态模型的工作过程包括:

[0078] 音频特征提取:对音频信号进行处理,提取梅尔频谱图。这是通过应用短时傅里叶变换(STFT)得到的,梅尔频谱图提供了音频信号在不同频段随时间变化的能量的可视化表示。

[0079] 文本特征提取:利用自动语音识别(Automatic Speech Recognition,ASR)将音频信号转换为文本信息。对文本信息进行处理,使用文本嵌入技术学习文本特征,并利用预训练的中文BERT模型提取文本的编码特征。

[0080] 第一轮融合:将音频特征和文本特征进行融合。由于它们具有相似的时间结构,采用元素相加的方式进行融合,以增强它们的时序结构。

[0081] Co-attention(共注意力)层:引入Co-attention层,用于实现模态间的信息交互。

这一阶段通过co-attention结构,使每个模态能够更好地利用来自其他模态的知识。

[0082] 第二轮融合:将Co-attention层输出的时序特征与BERT编码的特征进行融合。由于它们的时序不同,采用拼接的方式进行融合,以保留更多信息。

[0083] Self-attention(自注意力)处理:在最终的融合特征上应用self-attention层,全面处理来自所有模态的信息,以更好地捕捉多模态信息的复杂关系。

[0084] 情感预测:最后通过两层多层感知机(MLP)进行情感预测。MLP对全局特征进行处理,输出最终的情感预测结果。

[0085] 本方案一方面创建了一个新的中文情感识别数据集VCEMO,包含来自100多个个体的7477个音频信号样本的中文日常对话语料库,补充了中文语料在情绪识别领域中的不足;

[0086] 另一方面提出了一个多模态模型,用于处理声学数据和文本数据(词嵌入和预训练的BERT嵌入),使用共注意结构进行多模态特征融合,该多模态模型利用BERT模型提取的特征作为外部知识,使得模型能够更好地将来自外部世界的知识融入到情感识别任务。这带来了更好的泛化能力,使得模型在未见过的情境中仍能表现出色,提高模型鲁棒性。

[0087] 并且引入两轮融合的设计,首先融合音频和文本信息,然后再融合外部知识,使得模型能够更全面、深入地整合不同模态的信息。这种阶段化设计提高了模型对时序和语义结构的敏感性,使其更适应复杂多变的情感表达场景。

[0088] 此外,针对音频数据和词嵌入特征时序相似的情况,采用了直接的元素相加方式,以增强它们的时序结构。而对于与BERT编码特征时序不同的情况,则采用了拼接方式,以保留更多信息。引入自注意力层和两层MLP对融合后的信息进行全局处理和情感预测,增强了模型的表达能力。这种设计使得模型更好地捕捉多模态信息的复杂关系,提高了情感识别的准确性和鲁棒性。

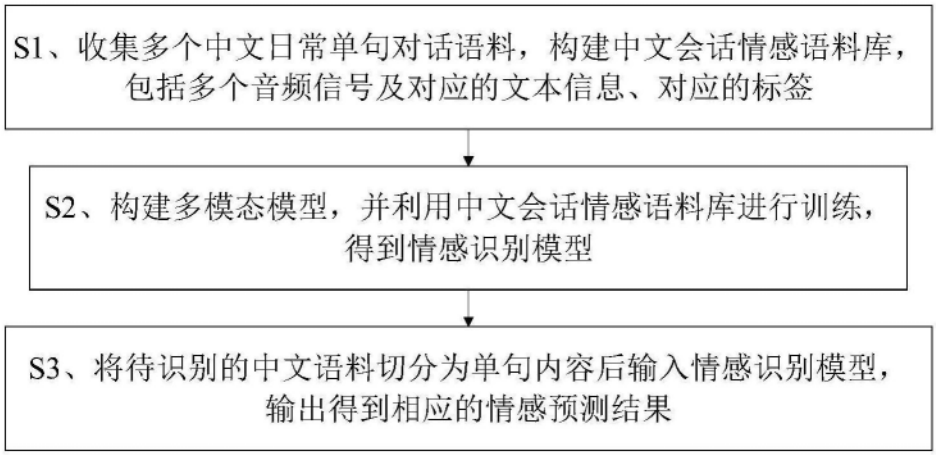

S1、收集多个中文日常单句对话语料，构建中文会话情感语料库，包括多个音频信号及对应的文本信息、对应的标签
S2、构建多模态模型，并利用中文会话情感语料库进行训练，得到情感识别模型
S3、将待识别的中文语料切分为单句内容后输入情感识别模型，输出得到相应的情感预测结果

图1

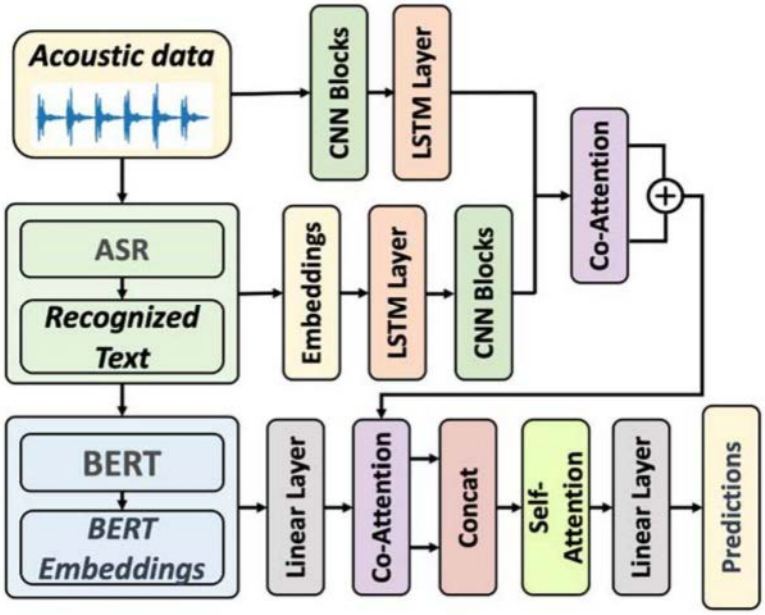

Acoustic data
CNN Blocks
LSTM Layer
ASR
Recognized Text
Embeddings
LSTM Layer
CNN Blocks
Co-Attention
BERT
BERT Embeddings
Linear Layer
Co-Attention
Concat
Self-Attention
Linear Layer
Predictions

图2