(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 118675552 B
(45)授权公告日 2025.04.01

(21)申请号 202410638236.1
(22)申请日 2024.05.22
(65)同一申请的已公布的文献号
申请公布号 CN 118675552 A
(43)申请公布日 2024.09.20
(73)专利权人 大连外国语大学
地址 116044 辽宁省大连市旅顺南路西段六号
(72)发明人 祁瑞华 郭旭
(74)专利代理机构 大连星河彩舟专利代理事务所(普通合伙) 21263
专利代理师 马新月
(51)Int.Cl.
G10L 25/63(2013.01)
G10L 25/30(2013.01)

(56)对比文件
CN 118038901 A,2024.05.14

审查员 李荣

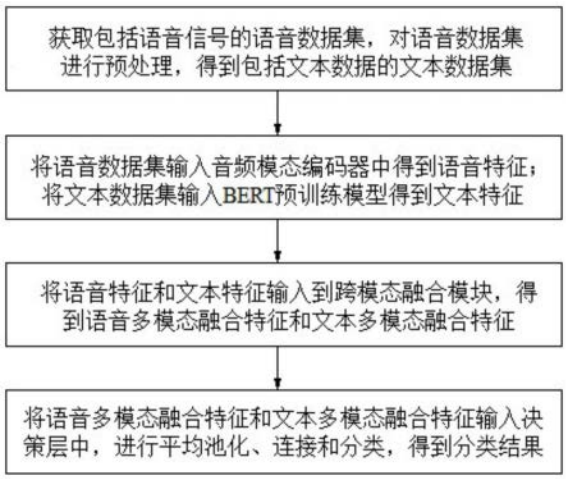
权利要求书3页 说明书11页 附图1页

(54)发明名称

一种基于语境信息增强和交叉注意力的语音情绪分类方法

(57)摘要

本发明提出一种基于语境信息增强和交叉注意力的语音情绪分类方法,包括获取包括语音信号的语音数据集,对语音数据集进行预处理,得到包括文本数据的文本数据集;将语音数据集输入音频模态编码器中进行上下文表示提取,得到语音特征;将文本数据集输入BERT预训练模型进行文本特征提取,得到文本特征;将语音特征和文本特征输入到跨模态融合模块,语音特征在音频模态特征学习时融入文本特征,得到语音多模态融合特征;文本特征在文本模态特征学习时融入语音特征,得到文本多模态融合特征;将语音多模态融合特征和文本多模态融合特征输入决策层中,进行平均池化、连接和分类,得到分类结果。本发明能够使语音情绪的分类结果更加准确。



CN 118675552 B

1.一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,包括如下步骤:

S1. 获取包括语音信号的语音数据集,对所述语音数据集进行预处理,得到包括文本数据的文本数据集;

S2. 将步骤S1中得到的所述语音数据集输入音频模态编码器中进行上下文表示提取,得到语音特征;将所述文本数据集输入BERT预训练模型进行文本特征提取,得到文本特征;

S3. 将步骤S2中得到的所述语音特征和文本特征输入到跨模态融合模块,所述语音特征在音频模态特征学习时融入所述文本特征,得到语音多模态融合特征;所述文本特征在文本模态特征学习时融入所述语音特征,得到文本多模态融合特征;

S4. 将步骤S3中得到的所述语音多模态融合特征和文本多模态融合特征输入决策层中,进行平均池化、连接和分类,得到分类结果;

所述步骤S3中,所述跨模态融合模块包括交叉注意力层和模态特征对齐融合层;

所述交叉注意力层用于音频模态特征学习,所述交叉注意力层通过模态间多头交叉注意力对所述语音特征进行深层语义抽取,如公式(1)所示:

$$mht_A = \text{Concat}(head_{A1}, head_{A2}, head_{A3}, ..., head_{A8})W_A^O \quad (1)$$

其中,mht_A 表示语音特征的深层语义抽取结果,Concat 表示模态间多头交叉注意力,$head_{Ai}$ 表示语音特征第i个头的注意力抽取,$i \in (1,2,...8)$,W_A^O 表示语音输出变换矩阵,

所述交叉注意力层还用于文本模态特征学习,所述交叉注意力层通过模态间多头交叉注意力对所述文本特征进行深层语义抽取,如公式(2)所示:

$$mht_{TE} = \text{Concat}(head_{TE1}, head_{TE2}, head_{TE3}, ..., head_{TE8})W_{TE}^O \quad (2)$$

其中,mht_{TE} 表示文本特征的深层语义抽取结果,$head_{TEi}$ 表示文本特征第i个头的注意力抽取,$i \in (1,2,...8)$,W_{TE}^O 表示文本输出变换矩阵,

所述模态特征对齐融合层用于将所述语音特征的深层语义抽取结果 mht_A 与语音特征进行拼接,得到语音多模态融合特征,如公式(3)所示:

$$CrossAtt_A = mht_A + hidden_A \quad (3)$$

其中,$CrossAtt_A$ 表示语音多模态融合特征,$hidden_A$ 表示语音特征,

所述模态特征对齐融合层还用于将所述文本特征的深层语义抽取结果 mht_{TE} 与文本特征进行拼接,得到文本多模态融合特征,如公式(4)所示:

$$CrossAtt_{TE} = mht_{TE} + hidden_{TE} \quad (4)$$

其中,$CrossAtt_{TE}$ 表示文本多模态融合特征,$hidden_{TE}$ 表示文本特征。

2.根据权利要求1所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,所述步骤S1中,对所述语音数据集进行预处理包括通过Wave2Vect2模型生成文本数据,如公式(5)所示:

$$x_t=\begin{cases}\mathrm{ASR}(X_0) & \text{当}X\text{为起始会话}X_0\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1) & \text{当}X\text{为起始会话}X_1\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1)+\mathrm{ASR}(X_2) & \text{当}X\text{为起始会话}X_2\\ \mathrm{ASR}(X_{n-3})+\mathrm{ASR}(X_{n-2})+\mathrm{ASR}(X_{n-1})+\mathrm{ASR}(X_n) & \text{当}X_n\text{中的}n\geq 3\end{cases} \tag{5}$$

其中,x_t表示文本数据,X表示语音信号,ASR表示语音文本识别算法,X_n表示语音信号中的起始会话,$n\in(0,1,2,3,4\ldots)$。

3.根据权利要求1所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,所述步骤S2中,所述音频模态编码器包括特征编码器模块、Transformer上下文表示模块和量化模块;

所述特征编码器模块用于将输入的所述语音信号处理为低级特征;

所述Transformer上下文表示模块用于将输入的所述语音信号映射到更能代表数据特征的特征空间;

所述量化模块用于将低级特征离散到一个可训练的码本中。

4.根据权利要求3所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,所述步骤S2中,所述音频模态编码器为Wav2vec2.0模型。

5.根据权利要求1所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,所述步骤S2中,所述BERT预训练模型通过bert-base-uncased对所述文本数据集中的文本数据进行文本特征提取,如公式(6)所示:

$$hidden_{TE}=BERT(x_t) \tag{6}$$

其中,$hidden_{TE}$表示文本特征,x_t表示文本数据,BERT表示bert-base-uncased。

6.根据权利要求5所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法,其特征在于,语音特征第i个头的注意力抽取$head_{Ai}$如公式(7)所示:

$$head_{Ai}=\mathrm{Attention}_A(Q_AW_{Ai}^{Q},K_AW_{Ai}^{K},V_AW_{Ai}^{V}) \tag{7}$$

其中,$\mathrm{Attention}_A$表示音频模态交叉注意力,Q_A表示语音特征的查询矩阵,W_{Ai}^{Q}表示语音特征第i个头的查询变换矩阵,K_A表示语音特征的键矩阵,W_{Ai}^{K}表示语音特征第i个头的键变换矩阵,V_A表示语音特征的值矩阵,W_{Ai}^{V}表示语音特征第i个头的值变换矩阵,

音频模态交叉注意力$\mathrm{Attention}_A$的计算公式如(8)所示:

$$\mathrm{Attention}_A(Q_A,K_{TE},V_{TE})=Softmax\left(\frac{Q_A{K_{TE}}^{T}}{\sqrt{d_k}}\right)V_{TE} \tag{8}$$

其中,K_{TE}表示文本特征的键矩阵,V_{TE}表示文本特征的值矩阵,T表示矩阵的转置操作,$\sqrt{d_K}$表示缩放因子,

文本特征第i个头的注意力抽取$head_{TEi}$如公式(9)所示:

$$head_{TEi}=\mathrm{Attention}_{TE}(Q_{TE}W_{TEi}^{Q},K_{TE}W_{TEi}^{K},V_{TE}W_{TEi}^{V}) \tag{9}$$

其中,$\mathrm{Attention}_{TE}$表示文本模态交叉注意力,Q_{TE}表示文本特征的查询矩阵,W_{TEi}^{Q}表示文

本特征第i个头的查询变换矩阵，K_{TE} 表示文本特征的键矩阵，W_{TEi}^{K} 表示文本特征第i个头的键变换矩阵，V_{TE} 表示文本特征的值矩阵，W_{TEi}^{V} 表示文本特征第i个头的值变换矩阵，

文本模态交叉注意力 Attention_{TE} 的计算公式如(10)所示：

$$\text{Attention}_{TE}(Q_{TE},K_A,V_A) = Softmax\left(\frac{Q_{TE}{K_A}^T}{\sqrt{d_k}}\right)V_A \quad (10)。$$

7.根据权利要求6所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法，其特征在于，所述步骤S3中，所述交叉注意力层通过8个头的模态间多头交叉注意力对所述语音特征进行深层语义抽取，所述交叉注意力层通过8个头的模态间多头交叉注意力对所述语音特征进行深层语义抽取。

8.根据权利要求6所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法，其特征在于，步骤S4中，所述语音多模态融合特征在所述决策层进行平均池化得到音频特征向量，如公式(11)所示：

$$x'_W = \begin{cases} AveragePool(CrossAtt_A(x_W)) & 当x_w为起始片段 \\ (AveragePool(CrossAtt_A(x_W)) + AveragePool(CrossAtt_A(x'_{W-1})))/2 & 当x_w为非起始片段 \end{cases} \quad (11)$$

其中，x'_W 表示音频特征向量，x_W 表示语音数据集中的语音片段，x'_{W-1} 为上一次语音片段生成的音频特征向量，

所述文本多模态融合特征在所述决策层进行平均池化得到文本特征向量，如公式(12)所示：

$$x_{TE}{}'_W = \begin{cases} AveragePool(CrossAtt_{TE}(x_{TEW})) & 当x_{TEW}为起始片段 \\ (AveragePool(CrossAtt_{TE}(x_{TEW})) + AveragePool(CrossAtt_{TE}(x_{TE}{}'_{W-1})))/2 & 当x_{TEW}为非起始片段 \end{cases} \quad (12)$$

其中，$x_{TE}{}'_W$ 表示文本特征向量，x_{TEW} 表示文本数据集中的文本片段，$x_{TE}{}'_{W-1}$ 为上一次语音片段生成的文本特征向量，

将所述音频特征向量 x'_W 和文本特征向量 $x_{TE}{}'_W$ 连接起来得到连接特征向量，使用线性分类器将对所述连接特征向量进行多类别分类，得到分类结果，如公式(13)所示：

$$\hat{y} = w*[x'_W; x_{TE}{}'_W] + b \quad (13)$$

其中，$\hat{y}$ 表示分类结果，w表示待学习的权重，b表示待学习的偏置量。

9.根据权利要求8所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法，其特征在于，步骤S4中，损失函数为类别交叉熵损失函数。

10.根据权利要求1所述的一种基于语境信息增强和交叉注意力的语音情绪分类方法，其特征在于，步骤S4中，所述分类结果为快乐、中立、悲伤或愤怒中的一种。

一种基于语境信息增强和交叉注意力的语音情绪分类方法

技术领域

[0001] 本发明涉及语音情感识别领域,具体涉及一种基于语境信息增强和交叉注意力的语音情绪分类方法。

[0002] 背景介绍

[0003] 语音情感识别(Speech Emotion Recognition,SER)技术主要聚焦于从语音信号中识别和理解情感状态。该技术广泛应用于人机交互、智能客服等领域,增强机器对人类情绪的感知与响应能力。语音情感识别方法主要分为两个步骤:语音特征提取和模型识别。在语音特征提取阶段,常见的做法是从语音信号中提取诸如语谱图、Mel频率倒谱系数MFCC、音高及其谐波、或抖动等声学特征,然后使用GMM、HMM、SVM等传统分类器对其进行分类,但是这些声学特征难以准确捕捉到复杂的语音情感,因此效果有限。

[0004] 随着深度学习的发展,基于深度学习的语音情感识别方法日益成为主流。例如,Yang等提出了利用波形和频谱图提取互补信息的方法,使用堆叠的BLSTM层进行唤醒和效价分类,改善了单一特征的语音情感识别效果。但这些描述特征难以全面捕捉情感信息,导致分类效果有限。近年来,通过构建综合多模态数据的深度学习模型,能够充分利用来自不同来源的有效信息,采用如卷积神经网络CNN或循环神经网络RNN等技术,用于同时编码和提取语音和文本数据的特征。现有研究证明,与传统的单模态方法相比,双模态情感识别方法在准确性和鲁棒性方面取得了显著的改进。例如,Yoon等提出采用双重递归神经网络RNNs编码音频和文本序列中的信息,然后结合这些来源的信息来预测情感类别,该模型同时利用文本数据和音频信号来更好地理解语音数据。此外,采用自注意力机制和跨模态注意力可以将语音和文本信息的注意力权重进行融合,以获得更精确的情感识别结果。例如,Xu等使用注意力网络来学习语音和文本之间的对齐,旨在产生更准确的多模态特征表示,然后使用Bi-LSTM网络来学习情绪。Zou等利用CNN、BiLSTM和wav2vec2.0提取多级声学信息,包括MFCC、声谱图和嵌入的高级声学信息。然后将这些提取的特征作为多模态输入,并通过所提出的共同注意机制进行融合。

[0005] 尽管近年来在语音情绪识别领域取得了显著进步,但现有技术仍在充分利用上下文信息方面存在局限,现有方法在处理较长对话中的情感累积和变化时,往往无法有效地识别和利用关键时刻的情感信息。例如,Wu等提到的基于预训练模型的情感识别方法虽然在提取语音特征和文本的上下文表示方面取得了一定进展,但这类模型依赖当前语音之后的语音和脚本,而在在现实应用场景中,当前语音之后的信息是未知信息,因此并不适用于现实应用场景。此外,Chen等提出的基于连接注意机制的多尺度SER并行网络,融合了细粒度帧级手动特性和粗粒度话语级深度特性,虽然在捕获语音信号特征方面表现良好,但对于复杂的对话情境和语境演变的敏感度仍有待提高。现有技术在交叉注意力机制具有局限性,交叉注意力机制在整合多模态信息如音频和文本时显示出了潜力,但现有模型如Sun等提出的MCSAN和Zou等提出的共同注意力网络仍面临一些挑战。这些模型虽然能够处理音频和文本间的交互,但在处理对话中不断变化的情感动态和多方面交互时,其效果并不理想。此外,Xu等的研究虽然在对齐音频和文本信号方面取得了进展,但在捕捉对话中细粒度情

感变化和复杂交互模式方面仍有限。此外,当前的技术还面临着在复杂对话环境下准确识别情绪的挑战,对话中的情感表达常常受到多种因素的影响,如说话者的个性、对话的环境背景以及说话者之间的互动关系等。现有技术在捕捉这些复杂因素对情感表达的影响方面还有很大的提升空间。针对上述存在的问题,研究设计一种新型的基于语境信息增强和交叉注意力的语音情绪分类方法,克服现有语音情绪分类方法中存在的问题是十分必要的。

发明内容

[0006] 本发明为解决现有语音情绪分类方法难以准确捕捉到复杂的语音情感、上下文信息未充分利用以及交叉注意力机制存在局限性而导致分类准确性低的问题,提出了一种基于语境信息增强和交叉注意力的语音情绪分类方法。

[0007] 本发明提供了一种基于语境信息增强和交叉注意力的语音情绪分类方法,包括如下步骤:

[0008] S1.获取包括语音信号的语音数据集,对所述语音数据集进行预处理,得到包括文本数据的文本数据集;

[0009] S2.将步骤S1中得到的所述语音数据集输入音频模态编码器中进行上下文表示提取,得到语音特征;将所述文本数据集输入BERT预训练模型进行文本特征提取,得到文本特征;

[0010] S3.将步骤S2中得到的所述语音特征和文本特征输入到跨模态融合模块,所述语音特征在音频模态特征学习时融入所述文本特征,得到语音多模态融合特征;所述文本特征在文本模态特征学习时融入所述语音特征,得到文本多模态融合特征;

[0011] S4.将步骤S3中得到的所述语音多模态融合特征和文本多模态融合特征输入决策层中,进行平均池化、连接和分类,得到分类结果。

[0012] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S1中,对所述语音数据集进行预处理包括通过Wave2Vect2模型生成文本数据,如公式(1)所示:

[0013] $$x_t=\begin{cases}\mathrm{ASR}(X_0) & \text{当}X\text{为起始会话}X_0\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1) & \text{当}X\text{为起始会话}X_1\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1)+\mathrm{ASR}(X_2) & \text{当}X\text{为起始会话}X_2\\ \mathrm{ASR}(X_{n-3})+\mathrm{ASR}(X_{n-2})+\mathrm{ASR}(X_{n-1})+\mathrm{ASR}(X_n) & \text{当}X_n\text{中的}n\geq 3\end{cases} \quad (1)$$

[0014] 其中,x_t表示文本数据,X表示语音信号,ASR表示语音文本识别算法,X_n表示语音信号中的起始会话,$n\in(0,1,2,3,4\cdots)$。

[0015] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S2中,所述音频模态编码器包括特征编码器模块、Transformer上下文表示模块和量化模块;

[0016] 所述特征编码器模块用于将输入的所述语音信号处理为低级特征;

[0017] 所述Transformer上下文表示模块用于将输入的所述语音信号映射到更能代表数据特征的特征空间;

[0018] 所述量化模块用于将低级特征离散到一个可训练的码本中。

[0019] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类

方法,所述步骤S2中,所述音频模态编码器为Wav2vec2.0模型。

[0020] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S2中,所述BERT预训练模型通过bert-base-uncased对所述文本数据集中的文本数据进行文本特征提取,如公式(2)所示:

[0021] $hidden_{TE}=BERT(x_t)$ (2)

[0022] 其中,$hidden_{TE}$表示文本特征,x_t表示文本数据,BERT表示bert-base-uncased。

[0023] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S3中,所述跨模态融合模块包括交叉注意力层和模态特征对齐融合层;

[0024] 所述交叉注意力层用于音频模态特征学习,所述交叉注意力层通过模态间多头交叉注意力对所述语音特征进行深层语义抽取,如公式(3)所示:

[0025] $mht_A=Concat(head_{A1},head_{A2},head_{A3},...,head_{A8})W_A^0$ (3)

[0026] 其中,mht_A表示语音特征的深层语义抽取结果,Concat表示模态间多头交叉注意力,$head_{Ai}$表示语音特征第i个头的注意力抽取,$i\in(1,2,\cdots8)$,W_A^0表示语音输出变换矩阵,

[0027] 语音特征第i个头的注意力抽取$head_{Ai}$如公式(4)所示:

[0028] $$head_{Ai}=\text{Attention}_A(Q_AW_{Ai}^Q,K_AW_{Ai}^K,V_AW_{Ai}^V) \quad (4)$$

[0029] 其中,$Attention_A$表示音频模态交叉注意力,Q_A表示语音特征的查询矩阵,W_{Ai}^Q表示语音特征第i个头的查询变换矩阵,K_A表示语音特征的键矩阵,W_{Ai}^K表示语音特征第i个头的键变换矩阵,V_A表示语音特征的值矩阵,W_{Ai}^V表示语音特征第i个头的值变换矩阵,

[0030] 音频模态交叉注意力$Attention_A$的计算公式如(5)所示:

[0031] $$\text{Attention}_A(Q_A,K_{TE},V_{TE})=Softmax\left(\frac{Q_AK_{TE}{}^T}{\sqrt{d_k}}\right)V_{TE} \quad (5)$$

[0032] 其中,K_{TE}表示文本特征的键矩阵,V_{TE}表示文本特征的值矩阵,T表示矩阵的转置操作,$\sqrt{d_K}$表示缩放因子,

[0033] 所述交叉注意力层还用于文本模态特征学习,所述交叉注意力层通过模态间多头交叉注意力对所述文本特征进行深层语义抽取,如公式(6)所示:

[0034] $mht_{TE}=Concat(head_{TE1},head_{TE2},head_{TE3},...,head_{TE8})W_{TE}^0$ (6)

[0035] 其中,mht_{TE}表示文本特征的深层语义抽取结果,$head_{TEi}$表示文本特征第i个头的注意力抽取,$i\in(1,2,\cdots8)$,W_{TE}^0表示文本输出变换矩阵,

[0036] 文本特征第i个头的注意力抽取$head_{TEi}$如公式(7)所示:

[0037] $$head_{TEi}=\text{Attention}_{TE}(Q_{TE}W_{TEi}^Q,K_{TE}W_{TEi}^K,V_{TE}W_{TEi}^V) \quad (7)$$

[0038] 其中,$Attention_{TE}$表示文本模态交叉注意力,Q_{TE}表示文本特征的查询矩阵,W_{TEi}^Q表示文本特征第i个头的查询变换矩阵,K_{TE}表示文本特征的键矩阵,W_{TEi}^K表示文本特征第i个头的键变换矩阵,V_{TE}表示文本特征的值矩阵,W_{TEi}^V表示文本特征第i个头的值变换矩阵,

[0039] 文本模态交叉注意力$Attention_{TE}$的计算公式如(8)所示:

[0040] $$\text{Attention}_{TE}(Q_{TE},K_A,V_A)=Softmax\left(\frac{Q_{TE}K_A{}^T}{\sqrt{d_k}}\right)V_A \quad (8)$$

[0041] 所述模态特征对齐融合层用于将所述语音特征的深层语义抽取结果mht_A与语音

特征进行拼接,得到语音多模态融合特征,如公式(9)所示:

[0042] $CrossAtt_A = mht_A + hidden_A$ (9)

[0043] 其中,$CrossAtt_A$表示语音多模态融合特征,$hidden_A$表示语音特征,

[0044] 所述模态特征对齐融合层还用于将所述文本特征的深层语义抽取结果mht_{TE}与文本特征进行拼接,得到文本多模态融合特征,如公式(10)所示:

[0045] $CrossAtt_{TE} = mht_{TE} + hidden_{TE}$ (10)

[0046] 其中,$CrossAtt_{TE}$表示文本多模态融合特征,$hidden_{TE}$表示文本特征。

[0047] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S3中,所述交叉注意力层通过8个头的模态间多头交叉注意力对所述语音特征进行深层语义抽取,所述交叉注意力层通过8个头的模态间多头交叉注意力对所述语音特征进行深层语义抽取。

[0048] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S4中,所述语音多模态融合特征在所述决策层进行平均池化得到音频特征向量,如公式(11)所示:

[0049] $$x'_W = \begin{cases} AveragePool(CrossAtt_A(x_W)) & 当x_w为起始片段 \\ (AveragePool(CrossAtt_A(x_W)) + AveragePool(CrossAtt_A(x'_{W-1})))/2 & 当x_w为非起始片段 \end{cases} \quad (11)$$

[0050] 其中,x'_W表示音频特征向量,x_W表示语音数据集中的语音片段,x'_{W-1}为上一次语音片段生成的音频特征向量,

[0051] 所述文本多模态融合特征在所述决策层进行平均池化得到文本特征向量,如公式(12)所示:

[0052] $$x_{TE}{}'_W = \begin{cases} AveragePool(CrossAtt_{TE}(x_{TEW})) & 当x_{TEW}为起始片段 \\ (AveragePool(CrossAtt_{TE}(x_{TEW})) + AveragePool(CrossAtt_{TE}(x_{TE}{}'_{W-1})))/2 & 当x_{TEW}为非起始片段 \end{cases} \quad (12)$$

[0053] 其中,$x_{TE}{}'_W$表示文本特征向量,x_{TEW}表示文本数据集中的文本片段,$x_{TE}{}'_{W-1}$为上一次语音片段生成的文本特征向量,

[0054] 将所述音频特征向量x'_W和文本特征向量$x_{TE}{}'_W$连接起来得到连接特征向量,使用线性分类器将对所述连接特征向量进行多类别分类,得到分类结果,如公式(13)所示:

[0055] $$\hat{y} = w*[x'_W; x_{TE}{}'_W] + b \quad (13)$$

[0056] 其中,$\hat{y}$表示分类结果,w表示待学习的权重,b表示待学习的偏置量。

[0057] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S4中,损失函数为类别交叉熵损失函数。

[0058] 根据本发明一些实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,所述步骤S4中,所述分类结果为快乐、中立、悲伤或愤怒中的一种。

[0059] 本发明提出的一种基于语境信息增强和交叉注意力的语音情绪分类方法,能够充分的利用上下文信息,能够基于语境信息增强上下文语义信息,同时利用文本数据和音频数据的语境信息增强来更好地理解语音信号的累积情绪,此外,本方法能够通过语境信息增强的交叉注意力机制获得多模态信号之间的对齐和不同特征映射的权重,能够更加关注关键信息,从而使语音情绪的分类结果更加准确。

附图说明

[0060] 图1本发明的一种基于语境信息增强和交叉注意力的语音情绪分类方法的流程示意图。

具体实施方式

[0061] 下面结合附图和实施例对本发明的实施方式作进一步详细描述。以下实施例用于说明本发明,但不能用来限制本发明的范围。

[0062] 实施例1

[0063] 一种基于语境信息增强和交叉注意力的语音情绪分类方法,如图1所示,包括如下步骤:

[0064] S1.获取包括语音信号的语音数据集,对语音数据集进行预处理,得到包括文本数据的文本数据集;

[0065] 具体的,对语音数据集进行预处理包括通过Wave2Vect2模型生成文本数据,如公式(1)所示:

[0066] $$x_t=\begin{cases}\mathrm{ASR}(X_0) & \text{当}X\text{为起始会话}X_0\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1) & \text{当}X\text{为起始会话}X_1\\ \mathrm{ASR}(X_0)+\mathrm{ASR}(X_1)+\mathrm{ASR}(X_2) & \text{当}X\text{为起始会话}X_2\\ \mathrm{ASR}(X_{n-3})+\mathrm{ASR}(X_{n-2})+\mathrm{ASR}(X_{n-1})+\mathrm{ASR}(X_n) & \text{当}X_n\text{中的}n\geq 3\end{cases} \quad (1)$$

[0067] 其中,x_t表示文本数据,X表示语音信号,ASR表示语音文本识别算法,X_n表示语音信号中的起始会话,n∈(0,1,2,3,4…)。本实施例中,预处理过程中采用的Wave2Vect2模型版本可以为wav2vec2-base-960h;

[0068] S2.将步骤S1中得到的语音数据集输入音频模态编码器中进行上下文表示提取,得到语音特征;将文本数据集输入BERT预训练模型进行文本特征提取,得到文本特征;

[0069] 具体的,音频模态编码器包括特征编码器模块、Transformer上下文表示模块和量化模块;

[0070] 其中,特征编码器模块用于将输入的语音信号处理为低级特征;Transformer上下文表示模块用于将输入的语音信号映射到更能代表数据特征的特征空间;量化模块用于将低级特征离散到一个可训练的码本中,在特征编码器模块训练过程中,部分低级特征被Transformer上下文表示模块掩码,目的是基于其上下文来识别掩码特征的数值向量,此外,本实施例的特征编码器模块可以包含12个Transformer上下文表示模块。

[0071] 音频模态编码器可以为Wav2vec2.0模型,可以应用于各种任务目标。本实施例在训练Wav2vec2.0模型的过程中进行预训练微调,改进Wav2vec2.0以适应本实施例的语音情绪识别任务,语音数据集中的原始音频信号直接送到wav2vec2.0处理器生成wav2vec2.0模型预训练表示,使用的原始音频信号在16kHz处进行采样,将每个音频会话分割成若干语音片段,当一个语音片段的长度小于7.7秒时,通过padding以保持相同的长度。

[0072] 步骤S2中,BERT预训练模型通过bert-base-uncased对文本数据集中的文本数据进行文本特征提取,如公式(2)所示:

[0073] $$\mathrm{hidden}_{TE}=\mathrm{BERT}(x_t) \quad (2)$$

[0074] 其中,hidden_{TE}表示文本特征,x_t表示文本数据,BERT表示bert-base-uncased。

[0075] S3.将步骤S2中得到的语音特征和文本特征输入到跨模态融合模块,语音特征在音频模态特征学习时融入文本特征,得到语音多模态融合特征;文本特征在文本模态特征学习时融入语音特征,得到文本多模态融合特征;

[0076] 步骤S3中,跨模态融合模块包括交叉注意力层和模态特征对齐融合层;

[0077] 交叉注意力层用于音频模态特征学习,交叉注意力层通过模态间多头交叉注意力对语音特征进行深层语义抽取,如公式(3)所示:

[0078] $mht_A=Concat(head_{A1},head_{A2},head_{A3},...,head_{A8})W_A^0$ (3)

[0079] 其中,mht_A表示语音特征的深层语义抽取结果,Concat表示模态间多头交叉注意力,$head_{Ai}$表示语音特征第i个头的注意力抽取,i∈(1,2,…8),W_A^0表示语音输出变换矩阵,

[0080] 语音特征第i个头的注意力抽取$head_{Ai}$如公式(4)所示:

[0081] $$head_{Ai}=\text{Attention}_A(Q_AW_{Ai}^Q,K_AW_{Ai}^K,V_AW_{Ai}^V) \quad (4)$$

[0082] 其中,$Attention_A$表示音频模态交叉注意力,Q_A表示语音特征的查询矩阵,W_{Ai}^Q表示语音特征第i个头的查询变换矩阵,K_A表示语音特征的键矩阵,W_{Ai}^K表示语音特征第i个头的键变换矩阵,V_A表示语音特征的值矩阵,W_{Ai}^V表示语音特征第i个头的值变换矩阵,

[0083] 音频模态交叉注意力$Attention_A$的计算公式如(5)所示:

[0084] $$\text{Attention}_A(Q_A,K_{TE},V_{TE})=Softmax\left(\frac{Q_AK_{TE}{}^T}{\sqrt{d_k}}\right)V_{TE} \quad (5)$$

[0085] 其中,K_{TE}表示文本特征的键矩阵,V_{TE}表示文本特征的值矩阵,T表示矩阵的转置操作,$\sqrt{d_K}$表示缩放因子,

[0086] 交叉注意力层还用于文本模态特征学习,交叉注意力层通过模态间多头交叉注意力对文本特征进行深层语义抽取,如公式(6)所示:

[0087] $mht_{TE}=Concat(head_{TE1},head_{TE2},head_{TE3},...,head_{TE8})W_{TE}^0$ (6)

[0088] 其中,mht_{TE}表示文本特征的深层语义抽取结果,$head_{TEi}$表示文本特征第i个头的注意力抽取,i∈(1,2,…8),W_{TE}^0表示文本输出变换矩阵,

[0089] 文本特征第i个头的注意力抽取$head_{TEi}$如公式(7)所示:

[0090] $$head_{TEi}=\text{Attention}_{TE}(Q_{TE}W_{TEi}^Q,K_{TE}W_{TEi}^K,V_{TE}W_{TEi}^V) \quad (7)$$

[0091] 其中,$Attention_{TE}$表示文本模态交叉注意力,Q_{TE}表示文本特征的查询矩阵,W_{TEi}^Q表示文本特征第i个头的查询变换矩阵,K_{TE}表示文本特征的键矩阵,W_{TEi}^K表示文本特征第i个头的键变换矩阵,V_{TE}表示文本特征的值矩阵,W_{TEi}^V表示文本特征第i个头的值变换矩阵,

[0092] 文本模态交叉注意力$Attention_{TE}$的计算公式如(8)所示:

[0093] $$\text{Attention}_{TE}(Q_{TE},K_A,V_A)=Softmax\left(\frac{Q_{TE}K_A{}^T}{\sqrt{d_k}}\right)V_A \quad (8)$$

[0094] 模态特征对齐融合层用于将语音特征的深层语义抽取结果mht_A与语音特征进行拼接,得到语音多模态融合特征,如公式(9)所示:

[0095] $CrossAtt_A=mht_A+hidden_A$ (9)

[0096] 其中,$CrossAtt_A$表示语音多模态融合特征,$hidden_A$表示语音特征,

[0097] 模态特征对齐融合层还用于将文本特征的深层语义抽取结果mht_{TE}与文本特征进

行拼接,得到文本多模态融合特征,如公式(10)所示:

[0098] $CrossAtt_{TE}=mht_{TE}+hidden_{TE}$ (10)

[0099] 其中,$CrossAtt_{TE}$表示文本多模态融合特征,$hidden_{TE}$表示文本特征。

[0100] 步骤S3中,交叉注意力层通过8个头的模态间多头交叉注意力对语音特征进行深层语义抽取,交叉注意力层通过8个头的模态间多头交叉注意力对语音特征进行深层语义抽取。

[0101] 通过跨模态融合模块处理转换后的文本数据和语音信号时,能够同时关注语音信号和文本数据的关键特征,通过将音频特征和文本数据映射到另一种专门用于融合的高维表示,识别出音频特征和文本特征的原始数据及其之间的关键情感信息,并加强对这些信息的理解,减少信息损失。

[0102] S4.将步骤S3中得到的语音多模态融合特征和文本多模态融合特征输入决策层中,进行平均池化、连接和分类,得到分类结果。

[0103] 步骤S4中,语音多模态融合特征在决策层进行平均池化得到音频特征向量,如公式(11)所示:

[0104] $$x'_W=\begin{cases}AveragePool(CrossAtt_A(x_W)) & 当x_w为起始片段\\ (AveragePool(CrossAtt_A(x_W))+AveragePool(CrossAtt_A(x'_{W-1})))/2 & 当x_w为非起始片段\end{cases} \quad (11)$$

[0105] 其中,x'_W表示音频特征向量,x_W表示语音数据集中的语音片段,x'_{W-1}为上一次语音片段生成的音频特征向量,以最后一个语音片段的表示输出作为音频特征向量。通过语音片段增强的决策方式可以有效地获得捕捉对情绪识别有用的片段级信息的能力,有助于语音情绪识别和跨模态的注意力对齐。

[0106] 文本多模态融合特征在决策层进行平均池化得到文本特征向量,如公式(12)所示:

[0107] $$x_{TE}{}'_W=\begin{cases}AveragePool(CrossAtt_{TE}(x_{TEW})) & 当x_{TEW}为起始片段\\ (AveragePool(CrossAtt_{TE}(x_{TEW}))+AveragePool(CrossAtt_{TE}(x_{TE}{}'_{W-1})))/2 & 当x_{TEW}为非起始片段\end{cases} \quad (12)$$

[0108] 其中,$x_{TE}{}'_W$表示文本特征向量,x_{TEW}表示文本数据集中的文本片段,$x_{TE}{}'_{W-1}$为上一次语音片段生成的文本特征向量,以最后一个语音片段的表示输出作为文本特征向量,通过文本片段增强的决策方式可以有效地获得捕捉对情绪识别有用的片段级信息的能力,有助于语音情绪识别和跨模态的注意力对齐。

[0109] 将音频特征向量x'_w和文本特征向量$x_{TE}{}'_W$连接起来得到连接特征向量,使用线性分类器将对连接特征向量进行多类别分类,以类别交叉熵损失函数作为损失函数为,得到分类结果,如公式(13)所示:

[0110] $$\hat{y}=w*[x'_W;x_{TE}{}'_W]+b \quad (13)$$

[0111] 其中,$\hat{y}$表示分类结果,w表示待学习的权重,b表示待学习的偏置量。

[0112] 步骤S4中,分类结果为快乐、中立、悲伤或愤怒中的一种。

[0113] 实施例2

[0114] 本实施例的一种基于语境信息增强和交叉注意力的语音情绪分类方法,在语音情感分析公开数据集IEMOCAP(Interactive Emotional Dyadic Motion Capture)数据集上对本发明方法的有效性进行了验证,语音情感分析公开数据集IEMOCAP(Interactive Emotional Dyadic Motion Capture)数据集是由南加州大学的Sail实验室收集的,内含语

音、视觉、文本、动作姿势四种模态的信息。该数据集涵盖由五位男演员和五位女演员进行录制的大约为12小时的音视频文件,文件分为Session1至Session5,共5组。参与者使用英语进行对话表演,表演分为即兴表演和固定的脚本场景表演。数据集IEMOCAP中对话平均持续时间为4.5秒,平均单词数为11.4。选用数据集IEMOCAP中的语音信息和文本数据作为实验数据集,对数据集IEMOCAP中常见的四类情感标签进行识别,包括愤怒、悲伤、快乐和中立,共计5531条。本实施例实验中的IEMOCAP数据集样本分布情况如表1所示。

[0115] 表1 IEMOCAP实验数据集样本分布

[0116]

情感类别	会话数量	整体占比
愤怒 Anger	1103	19.94%
悲伤 Sadness	1084	19.60%
快乐 Happy	1636	29.58%
中立 Neutral	1708	30.88%
合计	5531	100%

[0117] 本实施例所使用的语音信号在16kHz处进行采样,将每个音频会话分成若干个长度为7.7秒的语音片段,当一个语音片段小于7.7秒时,将对该语音片段应用具有0的填充操作,以保持相同的长度。每个音频会话话语的最终预测结果将由该话语的所有分割片段来决定。本实施例的方法在PyTorch中实现,模型的优化器采用AdamW,学习率为1e-5。训练的Batch Size为36,设置训练早停周期为20个epoch。整个方法都在Nvidia V100 32G GPU上进行预训练和训练。为了与以前的方法进行公平的比较,本实施例所提出的方法基于Session独立的原则分割数据集进行五折交叉验证,即当一个Session作为测试集时,其他的四个Session作为训练集。

[0118] 根据文献中的多类别语音情绪分类的通用评价指标,本实施例中语音情绪分类方法的评估指标采用加权精确率(Weight Accuracy,WA)和未加权精确率(Unweight Accuracy,UA)。加权精确率WA假设每个样本权重相等,对各个类别的预测结果进行加权平均后得到加权精确率,计算如公式(14)所示:

[0119] $$WA = \frac{\sum_{i=1}^{N} TP_i}{\sum_{i=1}^{N} TP_i + FP_i} \tag{14}$$

[0120] 其中,TP表示正确分类为第i类情感的样本数,FP表示错误分类为第i类的样本数,N表示样本数,

[0121] 未加权精确率UA是对所有类别的预测结果进行平均后的精确率,着重考察各类情感的综合平均识别性能,如公式(15)所示:

[0122] $$UA = \frac{\sum_{i=1}^{N} Acc_i}{N} \tag{15}$$

[0123] 其中,Acci表示每个情感类别的分类精确率,如公式(16)所示:

[0124] $$Acc_i = \frac{TP_i}{TP_i + FP_i} \tag{16}$$

[0125] 本实施例选取六个先进的语音情绪分类方法,及其在IEMOCAP公开数据集上基于Session独立的数据集分割的五折交叉验证实验结果来与本发明的方法进行比较,六个先进的语音情绪分类方法包括:

[0126] CMAspec:于2021年提出的一种基于跨模态注意力和语音语谱图的卷积神经网络。

[0127] CMARawwaveform:于2021年提出的一种基于跨模态注意力和原始波形的卷积神经

网络。

[0128] TSIN:于2021年提出的基于时间和语义一致性的多模态情感识别模型。

[0129] TSB&TAB ASR文本:于2021年提出的融合时间同步和时间异步表示的多模态情感识别模型,文本模态采用语音识别生成的文本作为输入,文本上下文窗口为[-3,+3]。

[0130] TSB&TAB原文本:于2021年提出的融合时间同步和时间异步表示的多模态情感识别模型,文本模态采用数据集中的原文本作为输入,文本上下文窗口为[-3,+3]。

[0131] Coattention&Fusion:于2022年提出的多模态共注意力多级融合的情感识别模型。

[0132] AuxiliaryTasks:于2022年提出的基于随机重组文本和音频输入的多模态情感识别模型。

[0133] 实验结果如表3所示,其中“--”表示文献中只提供了UA结果,未提供WA结果。

[0134] 表2与现有模型的对比实验结果

[0135]

模型	WA	UA
CMAspec, 2020	--	72.24
CMARawwaveform, 2020	--	72.84

[0136]

TSIN，2021	74.92	76.64
TSB&TAB ASR 文本，2021	70.96	71.90
TSB&TAB 原文本，2021	81.60	81.22
Coattention&Fusion，2022	--	76.31
AuxiliaryTasks，2023	78.42	79.71
本发明语音情绪分类方法+ASR ASR 文本多模态融合特征[-3,0]	76.38	77.31
本发明语音情绪分类方法+原文本的文本多模态融合特征[-3,0]	**81.64**	**82.50**

[0137] 为公平起见,首先比较使用语音和原文本作为输入的对照组。从表2可以看出,与采用原文本的基于跨模态注意力和语音语谱图的卷积神经网络CMAspec机器改进方法,以及基于跨模态注意力和原始波形的卷积神经网络CMARawwaveform相比,本实施例方法的语音情感识别精确率有显著提升,未加权精确率UA分别提升10.26%和9.66%。与基于时间和语义一致性的多模态情感识别模型TSIN相比,本实施例方法的语音情感识别加权精确率WA和未加权精确率UA分别提高6.72%和5.86%。与使用语音和原文本作为输入的Coattention&Fusion和AuxiliaryTasks模型相比,本实施例方法的未加权精确率UA均有明显提高,分别提高了6.19%和2.79%。与使用语音和原文本作为输入的TSB&TAB模型相比,在TSB&TAB使用了更多的文本信息[-3,+3]作为输入的情况下,本实施例方法的未加权精确率UA高出1.28%。

[0138] 然后比较使用语音和ASR文本作为输入的模型,TSB&TAB ASR文本模型采用了[-3,+3]的文本上下文窗口,使用的文本输入信息比本实施例方法多,但本实施例方法提出ASR模型的语音情感识别加权精确率WA和未加权精确率UA均高于TSB&TAB ASR文本模型,并且优势更为显著,加权精确率WA和未加权精确率UA分别高于TSB&TAB ASR文本模型5.42%和5.41%。

[0139] 为了探究语音多模态融合特征和文本多模态融合特征对语音情绪分类的作用及影响,本实施例进行消融实验来进行比较,如表3所示。

[0140] 表3模态消融实验结果

[0141]

特征	WA	UA
语音多模态融合特征+ASR 文本多模态融合特征[-3,0]	**76.38**	**77.31**
语音多模态融合特征	66.42	67.01
ASR 文本多模态融合特征[-3,0]	70.09	71.3
语音多模态融合特征+原文本的文本多模态融合特征[-3,0]	**81.64**	**82.50**
原文本的文本多模态融合特征[-3,0]	77.80	79.47

[0142] 从表3中可以看出,当同时使用语音多模态融合特征和ASR文本多模态融合特征,即同时采用语音模态增强信息和自动识别ASR文本增强信息时,比单独采用语音多模态融合特征即语音增强信息的加权精确率WA和未加权精确率UA分别高出9.96%和10.3%,比单独采用ASR文本多模态融合特征即ASR文本增强信息的加权精确率WA和未加权精确率UA分别高出6.29%和6.01%,说明语音多模态融合特征和ASR文本多模态融合特征都对情感识别任务准确率的提高做出了显著贡献。当同时采用语音多模态融合特征和原文本的文本多模态融合特征,即同时采用语音模态增强信息和原文本增强信息时,比单独采用语音多模态融合特征即语音增强信息的加权精确率WA和未加权精确率UA分别高出15.22%和15.49%,比单独采用原文本的文本多模态融合特征即原文本增强信息的加权精确率WA和未加权精确率UA分别高出3.84%和3.03%,再次说明语音多模态融合特征和原文本的文本多模态融合特征都对情感识别任务准确率的提高做出了显著贡献的同时,也证明了原文本比ASR文本具有更好的情感识别能力,原因在于自动识别ASR文本的准确率达不到原文本的精准程度。

[0143] 为了探究语境信息增强对情感识别任务的作用,在本实施例中比较了语境信息增强消融实验结果,如表4所示。

[0144] 表4语境信息增强消融实验结果

[0145]

语境信息增强	WA	UA
语音多模态融合特征+原文本的文本多模态融合特征[-3,0]	81.64	82.50
语音多模态融合特征+原文本	76.81	78.40
语音+原文本	74.58	75.89

[0146] 从表4中可以看出,当同时采用语音多模态融合特征和原文本的文本多模态融合特征,即同时采用语音模态增强信息和原文本增强信息时,比使用原文本的加权精确率WA和未加权精确率UA分别高出4.83%和4.1%。再次消融语音增强信息,仅采用原语音和原文本信息,加权精确率WA和未加权精确率UA又分别下降2.23%和2.51%,说明语音多模态融合特征和原文本多模态融合特征,即语音模态的增强信息和文本增强信息,都对情感识别任务准确率的提高做出了显著贡献。

[0147] 为了进一步探究语境信息增强对情感识别任务的作用,在本实施例中比较了文本增强信息的窗口长度对情感识别的影响实验,实验中采用IEMOCAP数据集的Session 1-4作为训练集,Session5作为测试集,实验结果如表5所示。

[0148] 表5文本增强信息窗口长度对情感识别的影响实验结果

[0149]

语境信息增强	WA	UA
语音多模态融合特征+原文本的文本多模态融合特征[-3,0]	82.68	82.89
语音多模态融合特征+原文本的文本多模态融合特征[-2,0]	79.61	81.30
语音多模态融合特征+原文本的文本多模态融合特征[-1,0]	77.68	79.20
语音多模态融合特征+原文本	75.1	76.58

[0150] 从表5中可以看出,当同时原文本增强信息的窗口长度从0增长到-3时,语音情感识别实验结果的加权精确率WA和未加权精确率UA逐步增高,说明文本增强信息窗口长度越长,对语音情绪分类的正向影响越显著。

[0151] 本发明的实施例是为了示例和描述起见而给出的,而并不是无遗漏的或者将本发明限于所公开的形式。很多修改和变化对于本领域的普通技术人员而言是显而易见的。选择和描述实施例是为了更好说明本发明的原理和实际应用,并且使本领域的普通技术人员能够理解本发明从而设计适于特定用途的带有各种修改的各种实施例。

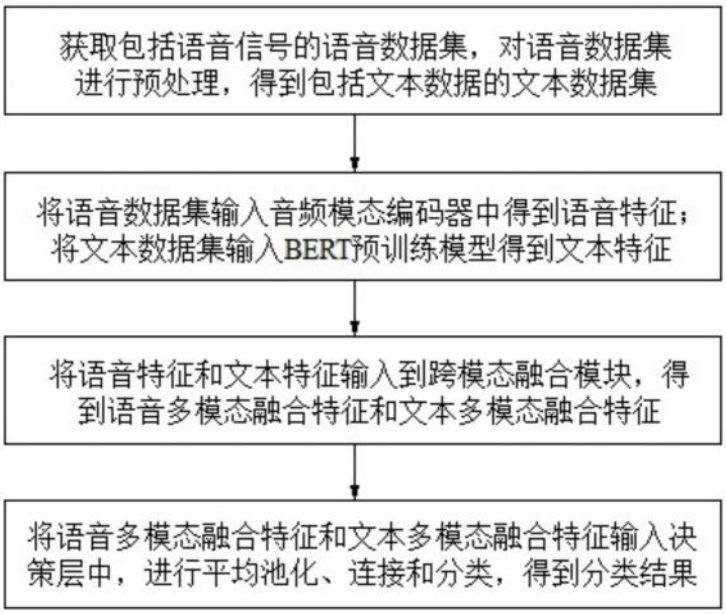

获取包括语音信号的语音数据集，对语音数据集进行预处理，得到包括文本数据的文本数据集
将语音数据集输入音频模态编码器中得到语音特征；将文本数据集输入BERT预训练模型得到文本特征
将语音特征和文本特征输入到跨模态融合模块，得到语音多模态融合特征和文本多模态融合特征
将语音多模态融合特征和文本多模态融合特征输入决策层中，进行平均池化、连接和分类，得到分类结果

图1