Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

深度学习语音情绪识别技术

硕士研究生 杨桢弘

2024年11月10日

问题回溯

- 相关内容
  - 暂无

内容提要

- 预期收获
- 内涵解析与研究目标
- 研究背景与意义
- 研究历史与现状
- 知识基础
- 算法原理
  - AMRWC
  - TIM-Net
- 特点总结与未来展望
- 参考文献

预期收获

- 预期收获
  - 掌握语音情绪识别的研究现状与基本概念
  - 理解语音情绪识别的基本模型及其原理
  - 了解语音情绪识别未来发展方向

内涵解析与研究目标

- 内涵解析
  - 语音识别：
    - 语音→文本（仅限）
    - 侧重语言特征
  - 情绪识别：文本/语音/图像/视频→情绪类别/强度
  - 语音情绪识别
    - 语音→情绪类别/强度
    - 侧重非语言特征，更关注声音中隐含的情感信息

- 研究目标
  - 结合深度学习、信号处理等技术
  - 准确识别情绪类别和强度，实现在不同场景、语言和说话者的情况下均达到较优效果，提升语音情绪识别的泛化性

研究背景与意义

- 研究背景
  - 智能设备的普及，语音助手（如Siri、Alexa）等已成为日常生活中的一部分
  - 传统的人机交互往往缺乏情感理解，导致机器的反馈显得机械化
- 研究意义
  - 人机交互中，机器对用户的情绪变化做出恰当反应，提供更具个性化的服务
    - 心理健康
    - 移动服务
    - 车载系统
    - 呼叫中心

研究历史与现状

1996 Dellaert等人提出了基于统计模式识别的方法，首次使用统计学方法进行语音情绪识别，并用KNN和GMM对情绪进行分类

2011 Liu等人首次将DBN应用于语音情绪识别，验证了深度学习在情绪识别中的可行性和有效性

2016 Trigeorgis等提出了端到端的语音情绪识别方法，通过DCRN实现情感识别，避免手工设计特征，开创性表明深度神经网络可以直接从原始信号中学习到情感特征

2019 Zhang等人首次在语音情绪识别中应用Transformer，运用注意力机制在全局范围内捕获语音中的情绪依赖关系

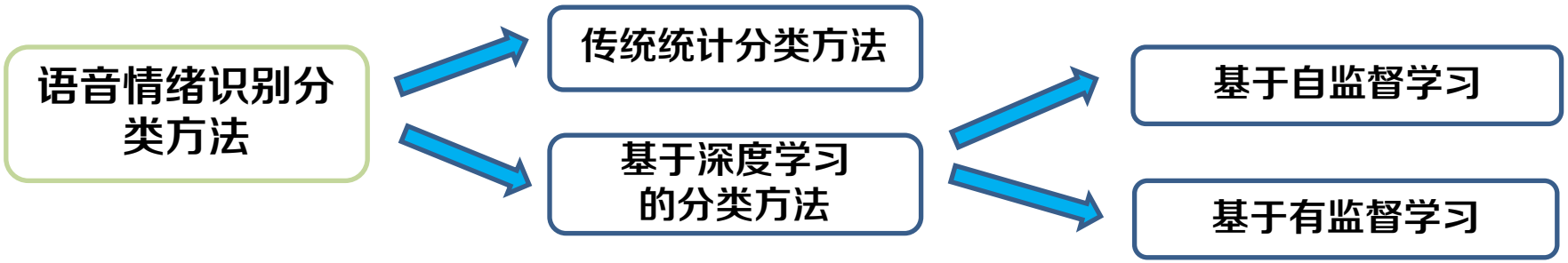
2021 Chen等人提出了WavLM，自监督语音学习，引入噪声模拟和多任务训练方法

2022 Neil等人提出了SERAB基准，包含九个语音情感识别任务，涵盖六种语言，提供DNN、手工特征、基于Transformer的基线

2023 Ye等人提出了TIM-Net，整合过去和未来的信息，融合不同时间尺度的特征，有效捕捉情感的时序动态特征

2024 Li等人提出了通过交叉注意力融合AM-Resnet和Wav2vec2.0的特征，解决数据稀疏和无声帧影响的问题

2024 Chen等人提出Vesper预训练模型，采用初始压缩、任务特定预训练、掩码过程等方法，在保证性能同时降低计算负担

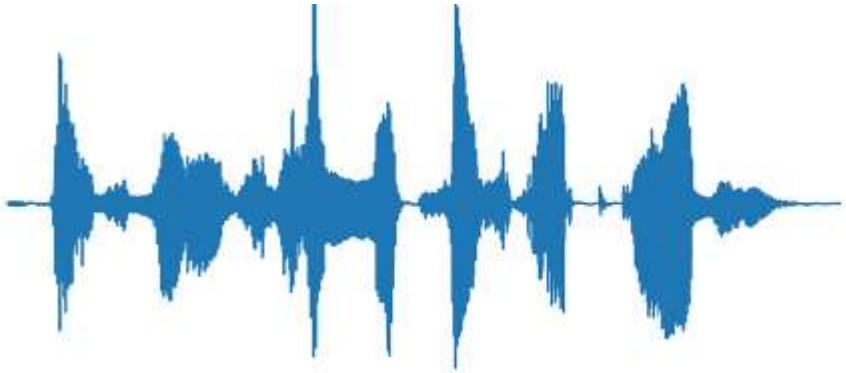


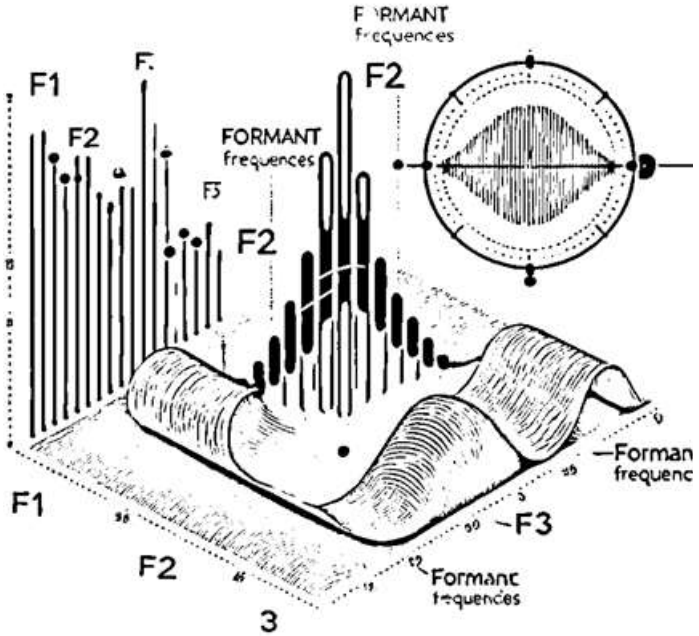
研究历史与现状

- 现有存在问题及解决方案
  - 带标注的数据量少 → 自监督学习
    - 表演
    - 自发
    - 诱发
  - 泛化能力弱 → 跨语种，增强数据集多样性
    - 跨语言、文化
  - 实时、个性化 → 情绪动态识别，时序建模
  - 抗噪能力弱
    - 回旋失真（更便宜的接收器）
    - 干扰扬声器（环境背景音）
    - 预处理阶段本身的语音分离
  - 情绪的复杂性 → 多标签标注

知识基础 特征

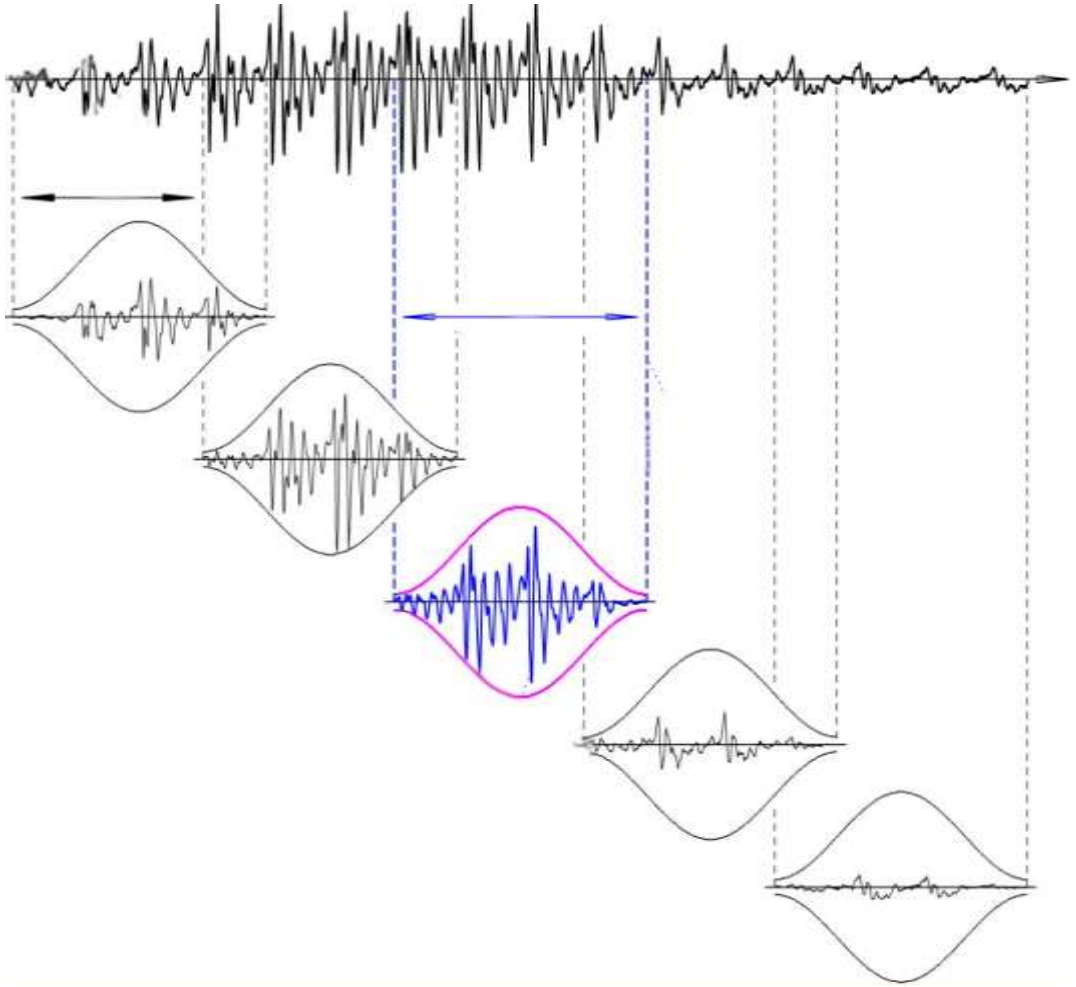
- 韵律
  - 音调、能量、持续时间
  - 基频：语音信号中频率最低值
  - 过零率：语音信号在单位时间内经过零值的次数
- 质量
  - 共振峰：声道在发声时，由于声带的振动和声道的特定形状，形成的共振频率
    - F1（第一共振峰）：与口腔开度有关
    - F2（第二共振峰）：与舌头在口腔中的前后位置有关
    - F3（第三共振峰）：影响口音和个性化语音特征，与嘴唇形状和其他细微的发音机制有关





知识基础 预处理

- 短时分析
  - 将语音信号分为短时帧进行处理，每帧通常为20-40毫秒，以捕捉语音信号在不同时间片的变化
- 滑动窗口
  - 一种用于分割语音信号以进行帧级处理的技术，滑动窗口步长（通常10毫秒）用于控制帧的重叠率
- 语音活动检测
  - 语句由三部分组成：清音、浊音、无声音
- 降噪
  - MMSE



知识基础 特征提取

- 频谱
  - MFCC：模拟人类耳朵对声音的感知方式，将音频信号转换成一组特征参数
  - 线性预测编码（LPC）

$$s[n] = \sum_{k=1}^{p} a_k s[n-k] + e[n]$$

最小二乘法求解预测系数，使预测误差最小化

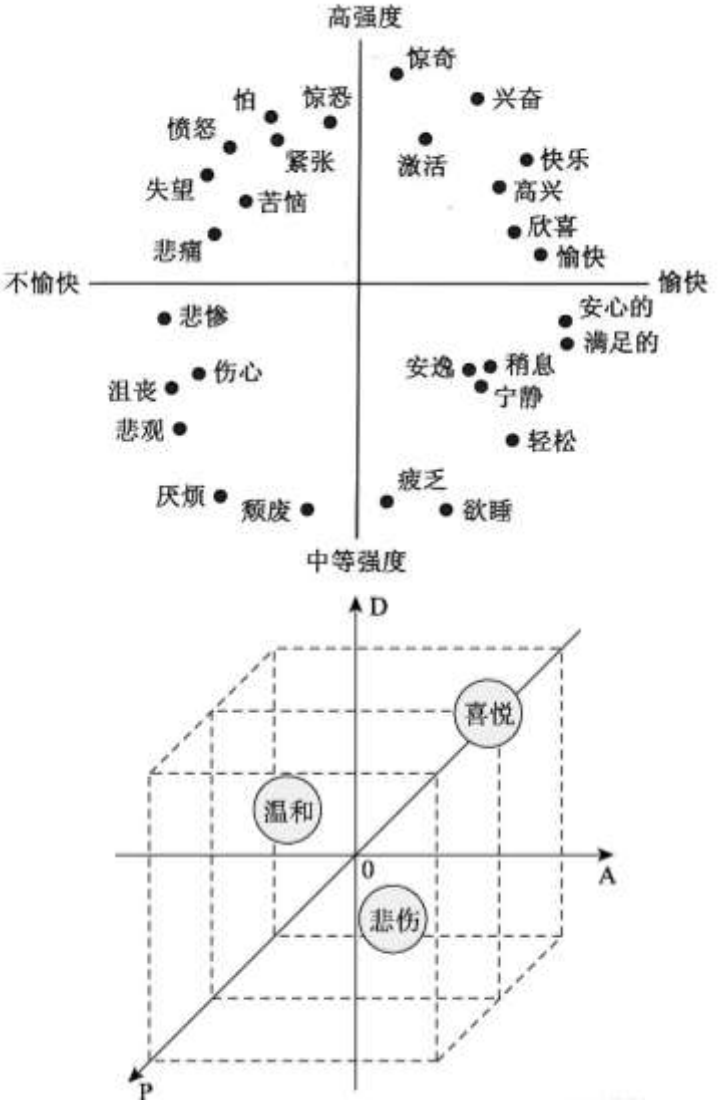
  - 感知线性产生（PLP）
    - ERB滤波：频率越高分辨率越低，滤波器的宽度随频率增加而增加
    - 响度压缩：对数压缩、平方根压缩
- TEO特征
  - 捕获信号的瞬时能量变化

$$\psi(x(n)) = x^2(n) - x(n-1)x(n+1)$$

$x(n)$是采样的语音信号

知识基础 情绪分类

- 离散
  - 基本情绪
    - 快乐、悲伤、愤怒、害怕、厌恶、惊讶、轻蔑
  - 扩展情绪
    - 信任、期待、害羞等
- 连续
  - 情绪环
    - 愉悦度：负面→正面
    - 唤醒度：平静→激动
  - PAD模型
    - 支配度：用户与外部环境相互主导强弱　弱→强
    - 愤怒、恐惧



知识基础 评价指标

- 情绪分类评价指标
  - 从分类和回归的角度采用不同的衡量指标
    - 从分类的角度来看
      - 预测准确率ACC
      - ROC曲线下面积AUC
      - 未加权、加权平均召回率UAR、WAR
      - 综合评估精确率与召回率F1值
      - 混淆矩阵
    - 从回归的角度来看，量化预测情绪与实际情绪之间的误差
      - 均方根误差MSE
      - Pearson相关系数PCC

- 情绪分类评价指标
  - 谐波噪声比（HNR）
    - 谐波成分的强度与噪声成分的强度之比
      - HNR值越高，语音越清晰
      - HNR值越低，噪声越多，语音更沙哑或含有杂音

$$HNR = 10 \cdot log_{10} \frac{P_{harmonics}}{P_{noise}}$$

$P_{harmonics}$是谐波成分的功率

P_{noise}是噪声成分的功率

Cross-feature fusion speech emotion recognition based on attention mask residual network and Wav2vec 2.0

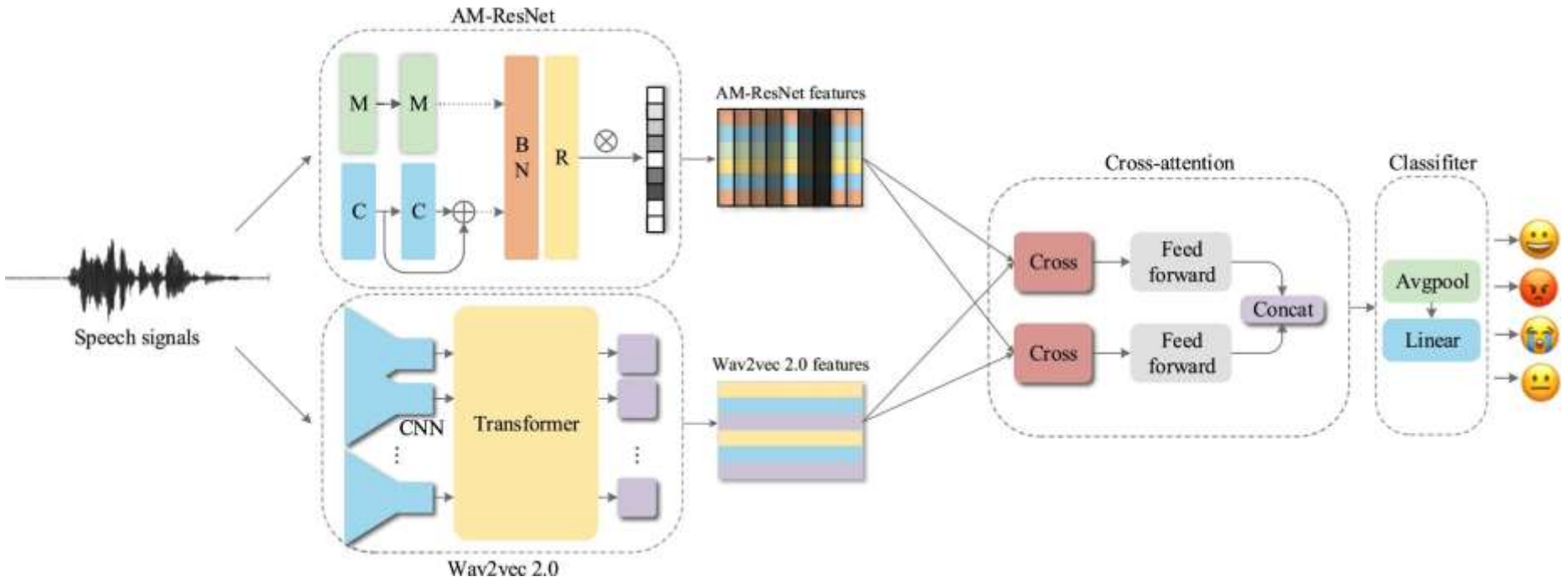
AMRWC TIPO

T	目标	减少不相关信息对语音信号和数据稀疏性的影响
I	输入	1个数据集（10位演员，5男5女，5个会话，每个会话至少三个注释者标记的分类标签）
P	处理	1. 分别通过AM-ResNet和Wav2vec2.0提取对应的特征 2. 交叉注意力模块动态交互融合AM-ResNet和Wav2vec2.0的特征 3. 预测最终情绪
O	输出	4种情绪分类

P	问题	1.现有方法存在静音帧和无声帧的干扰，语音情绪识别的精度较低 2.现有方法数据集规模有限，会限制深度学习的性能
C	条件	需要预训练的Wav2vec2.0模型
D	难点	1.如何处理静音帧和无声帧，增强语音情绪识别的精度 2.如何有效缓解数据集稀缺
L	水平	2024 SCI 1区

AMRWC 算法原理

- 算法原理图
  - 分别通过AM-ResNet和Wav2vec2.0提取对应的特征
  - 交叉注意力模块动态交互融合AM-ResNet和Wav2vec2.0的特征
  - 预测最终情绪



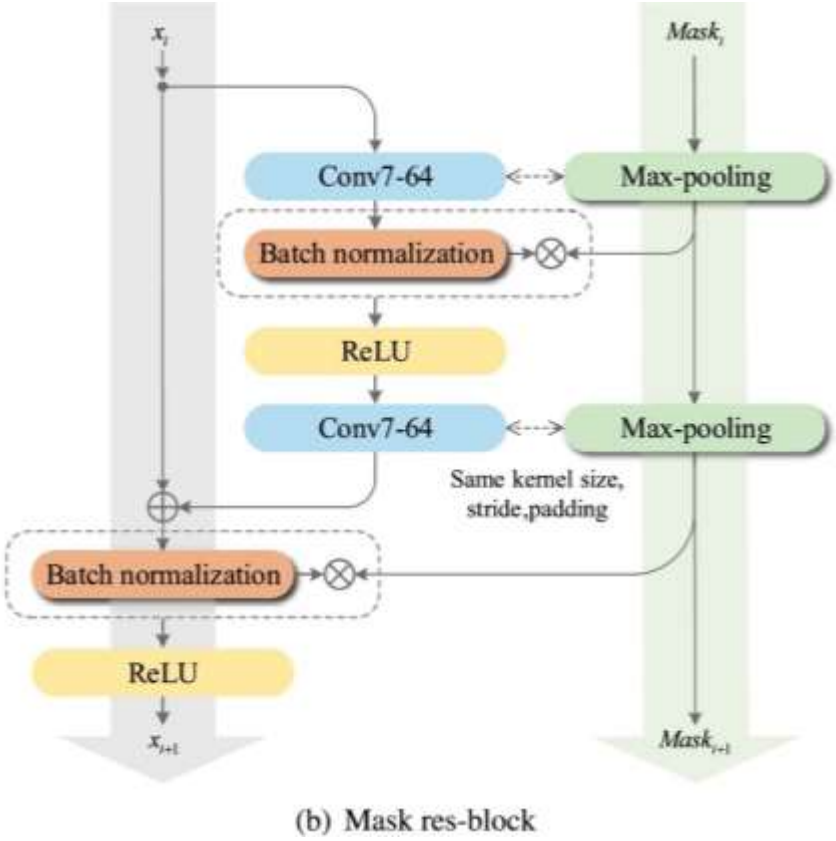
- 现有方法存在问题
  - 语音信号一般由浊音、清音和无声帧组成，清音和无声帧可能会增加计算复杂性并降低情绪识别的准确性

- VAD的解决方法
  - 去除无声帧或清音
  - 存在问题
    - 去除无声帧会破坏时序信号的连续性，不易于识别情绪
    - 清音中包含一些语音信息，直接去除会降低分类精度

- 解决方法 AM-ResNet
  - 最大幅度差异检测（MADD）
    - 利用滑动窗口中的最大幅度减去最小幅度得到的最大幅度差（MAD）作为检测特征，检测语音存在区域
  - 掩模残差网络（M-ResNet）

$$\hat{Z}_{i,j} = a_i \times \frac{Z_{i,j} - E(Z_i)}{\sqrt{Var(Z_i) + \epsilon}} + b_i$$

    - 在静默区和填充区，掩码将相应的特征值保持为0
    - 掩码会随输入特征的长度而变化



(b) Mask res-block

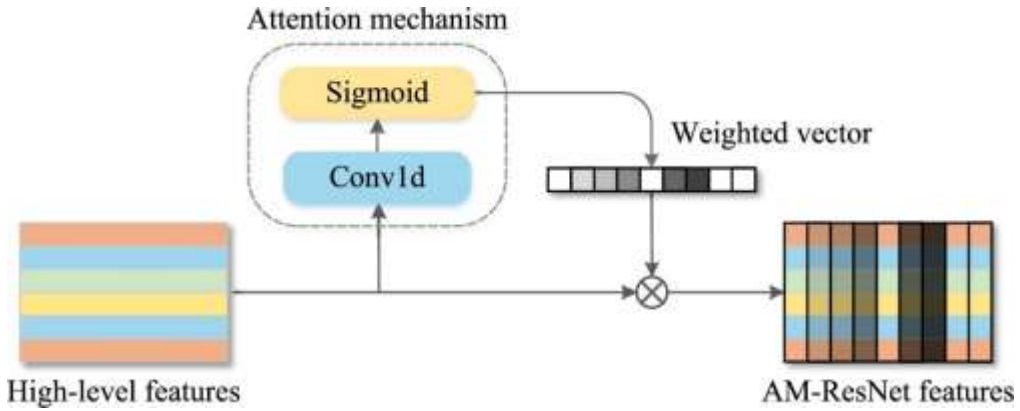
- 解决方法 AM-ResNet
  - 注意力机制获取语音信号的AM-ResNet特征
    - 从M-ResNet中提取高级特征生成加权向量

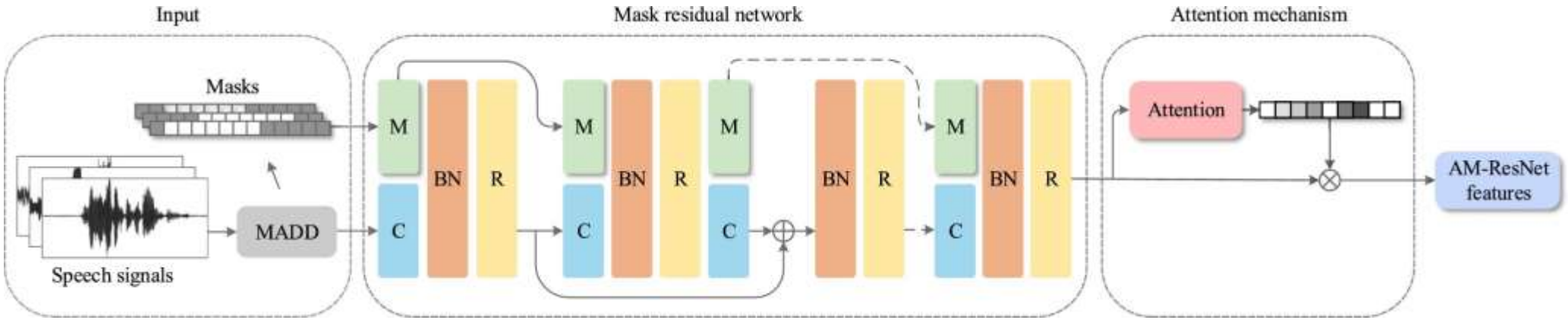
$$W = Sigmoid\big(O(E, W_{att})\big)$$

$O(\cdot)$表示卷积算法，W_{att}表示卷积权重矩阵

    - 加权向量乘高级特征向量

$$\bar{E} = E \otimes w$$

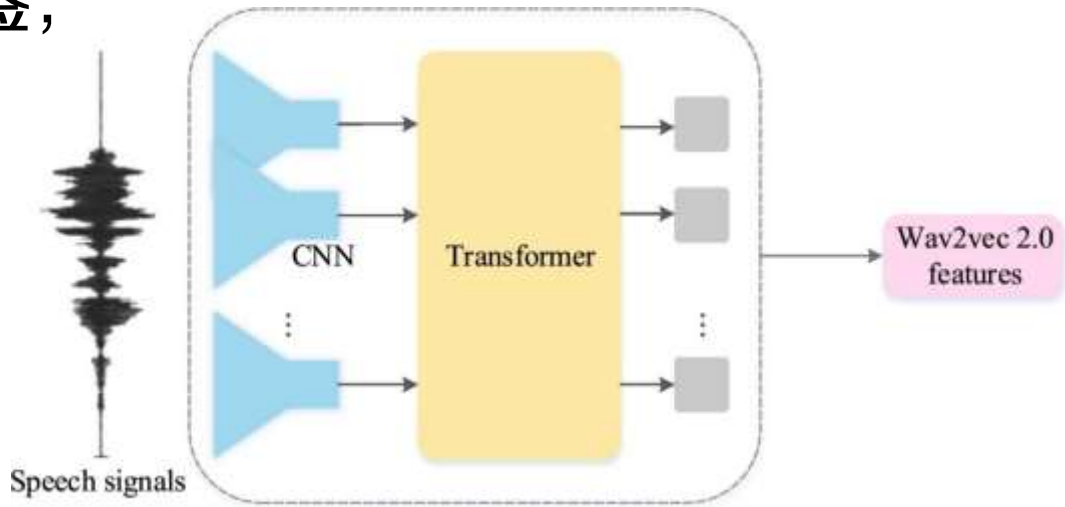




- 现有方法存在问题
  - 大规模带注释的语音情感数据相对稀缺，模型泛化能力会受到限制

- 单标签学习的解决方法
  - 侧重于注释者的多数同意
  - 存在问题
    - 缺乏多数共识，模糊的情感话语会被丢弃，降低情感识别准确率

- 解决方法
  - Wav2vec2.0: 可以提高SER性能，其提取的特征优于传统特征
    - 大规模未标记语音数据集上预训练
    - 标记数据微调，进行语音识别
    - 使用在960小时的LiberSpeech数据集上预训练的Wav2vec2.0模型
    - 将Wav2vec2.0的最后一层隐藏状态作为特征
  - 多标签学习：为每个话语生成情感比例标签，每个话语都可以分为明确或模糊
    - 明确的话语是多标签中唯一获得多数票的情感，反之是模糊的话语

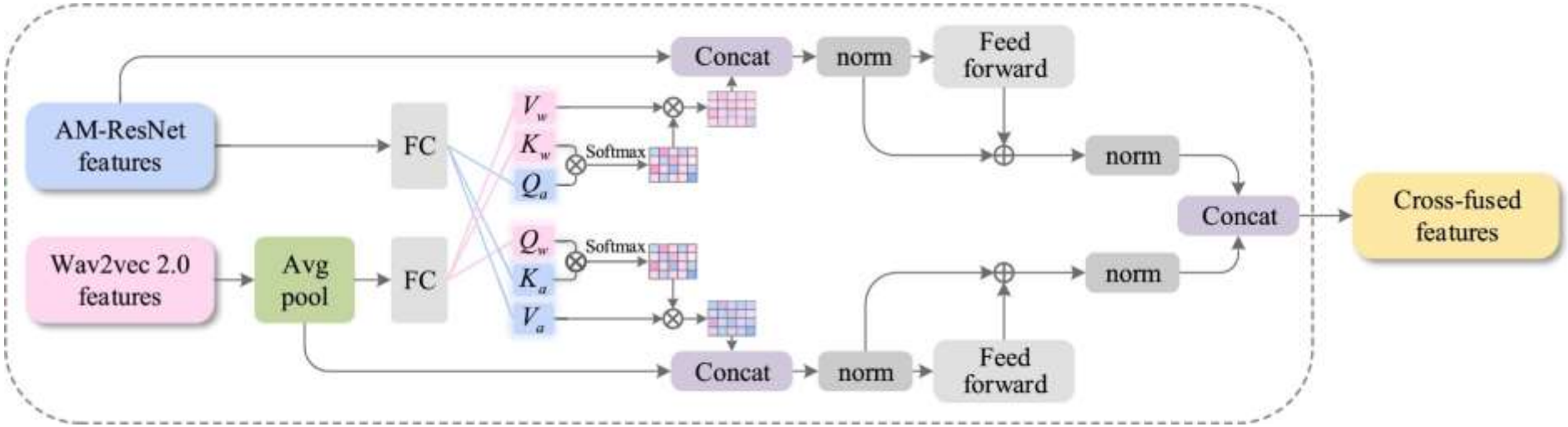


AMRWC 创新点分析 交叉注意力机制

- 交叉注意力机制融合两个特征
  - 对Wav2vec2.0特征进行自适应平均化
  - 对两个特征向量分别映射为查询(Q)、键(K)、值向量(V)
  - 执行查询和键之间的点积交叉计算，以估计两个特征之间的相关性

$$F_{a\to w} = softmax(\boldsymbol{Q}_w \boldsymbol{K}_a^T/\sqrt{d})V_a$$

$$F_{w\to a} = softmax(\boldsymbol{Q}_a \boldsymbol{K}_w^T/\sqrt{d})V_w$$



- 交叉注意力机制融合两个特征
  - 层归一化和前馈神经网络，更新特征

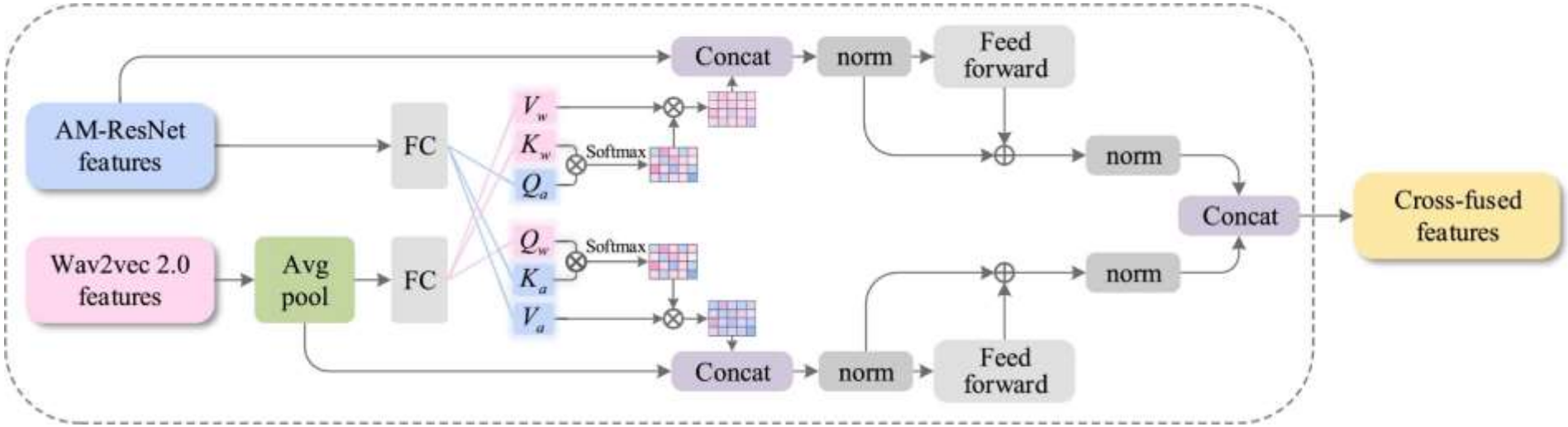
$$\acute{F}_{\mathrm{a}} = LN(F_{\mathrm{a}\to w} + FeedForward(F_{\mathrm{a}\to w}))$$

$$\acute{F}_{w} = LN(F_{w\to a} + FeedForward(F_{w\to a}))$$

LN表示层归一化操作，$FeedForward$表示全连接的前馈层

  - 特征拼接融合

$$F_{\mathrm{c}} = Concat(\acute{F}_{\mathrm{a}}\acute{F}_{w})$$



实验设计 数据资源

- 数据集
  - IEMOCAP（10位演员，5男5女，5个会话，每个会话至少三个注释者标记的分类标签）
- 对比方法

  Chen(2018)、Chou(2019)、Pepino(2021)

  Li(2022)、Yue(2022)、Pastor(2023)

  Etienne(2018)、Ando(2018)、Upadhyay(2024)
- 评价指标
  - UAR：未加权平均召回率
  - WA：加权准确率
  - 混淆矩阵

实验设计 对比实验

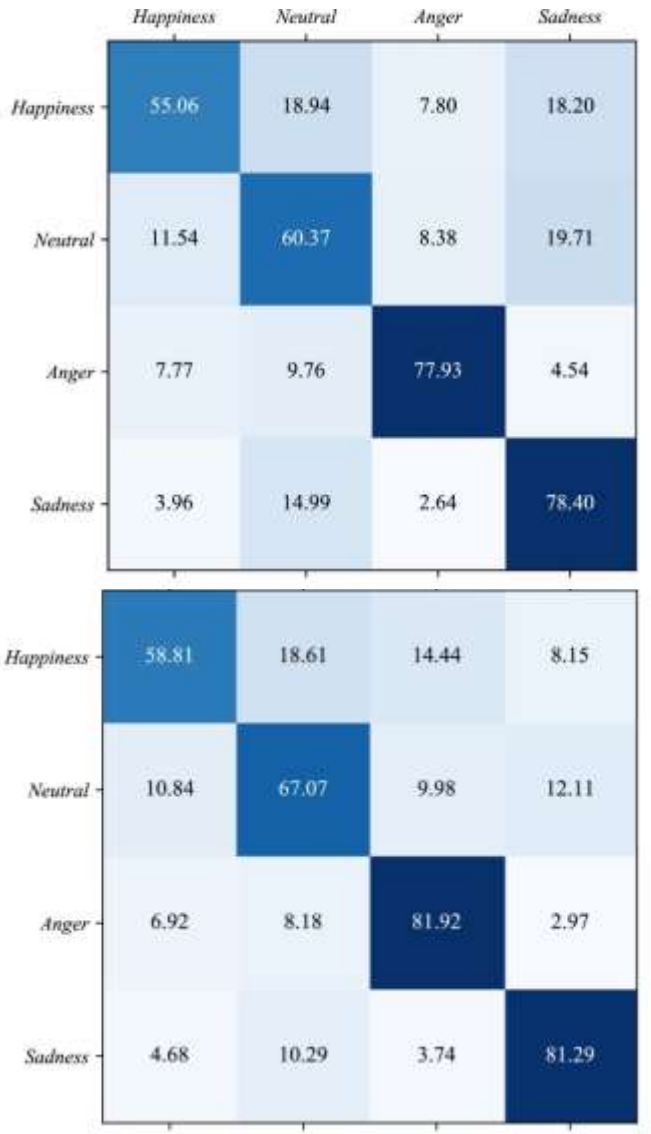
- 评估AMRWC在标注清晰和模糊数据集上的表现
  - AMRWC在清晰和模糊数据集上准确率都优于对比方法，效果提升较大
  - 仅使用清晰的话语，训练集优于UAR或WA中的所有方法，表明AMRWC可以利用Wav2vec2.0的自监督学习能力来解决数据稀疏的问题

Method	Training set	WA (%)	UAR (%)
Chen et al. [36] (2018)	clear	-	64.74
Chou et al. [37] (2019)	clear	-	61.48
Pepino et al. [21] (2021)	clear	-	67.20
Li et al. [24] (2022)	clear	63.40	-
Yue at al. [23] (2022)	clear	68.29	-
Pastor et al. [38] (2023)	clear	-	65.90
AM-ResNet	clear	67.35	65.56
AM-ResNet+W2V2+CA	clear	68.36	67.75
Etienne et al. [26] (2018)	clear+ambiguous	64.50	-
Ando et al. [27] (2018)	clear+ambiguous	62.60	63.70
Chou et al. [37] (2019)	clear+ambiguous	-	61.68
Upadhyay et al. [39] (2024)	clear+ambiguous	-	63.92
AM-ResNet	clear+ambiguous	67.50	66.56
AM-ResNet+W2V2+CA	clear+ambiguous	**70.79**	**72.27**

实验设计 消融实验

- 评估AMRWC的不同模块在IEMOCAP数据集上的表现
  - MR：仅使用掩码残差网络
  - AMR：仅使用注意力掩码残差网络
  - AMRWF：使用AM-ResNet和wav2vec2.0，全连接层融合
  - AMRWC：使用AM-ResNet和wav2vec2.0，交叉注意力融合

Fold	Testing set	ResNet		M-ResNet		AM-ResNet	
		UAR (%)	WA (%)	UAR (%)	WA (%)	UAR (%)	WA (%)
1	Session1F	72.57	73.12	**73.95**	**75.27**	72.19	73.12
2	Session1M	68.53	66.25	**69.93**	**68.97**	**71.22**	69.81
3	Session2F	70.59	69.09	**74.41**	**71.43**	72.32	71.43
4	Session2M	**73.01**	**66.12**	72.31	63.79	**74.05**	**68.69**
5	Session3F	**64.34**	**68.78**	64.87	66.25	**69.33**	**71.31**
6	Session3M	**64.38**	63.31	63.94	**67.11**	**64.20**	63.88
7	Session4F	**63.48**	**73.80**	62.17	68.45	**64.85**	**74.06**
8	Session4M	66.10	**66.59**	**68.20**	66.35	67.07	**66.83**
9	Session5F	62.87	64.37	**64.32**	**65.75**	63.82	**66.13**
10	Session5M	59.71	63.59	**63.29**	**65.44**	60.35	**66.13**
Average		66.56	67.50	**67.74**	**67.88**	**67.90**	**68.87**



算法总结 AMRWC

- 算法贡献
  - AM-ResNet：解决了无声帧和清音会增加计算复杂度并降低情绪识别准确性的问题
    - 利用掩模残差块保证无声区域值不变
    - 利用注意力机制为清音和浊音分配不同的权重，减少清音造成的冗余情感信息
  - Wav2vec2.0 + 多标签学习：解决SER中的数据稀疏性
  - 交叉注意力：有效利用两个特征进行融合，得到更优效果

- 算法不足
  - 情感状态可能随对话的上下文变化而变化，只是基于单一语音片段进行情感预测，可能会影响情绪识别的准确性
  - 只在IEMOCAP数据集上进行实验，泛化能力有限

TIM-Net

Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition

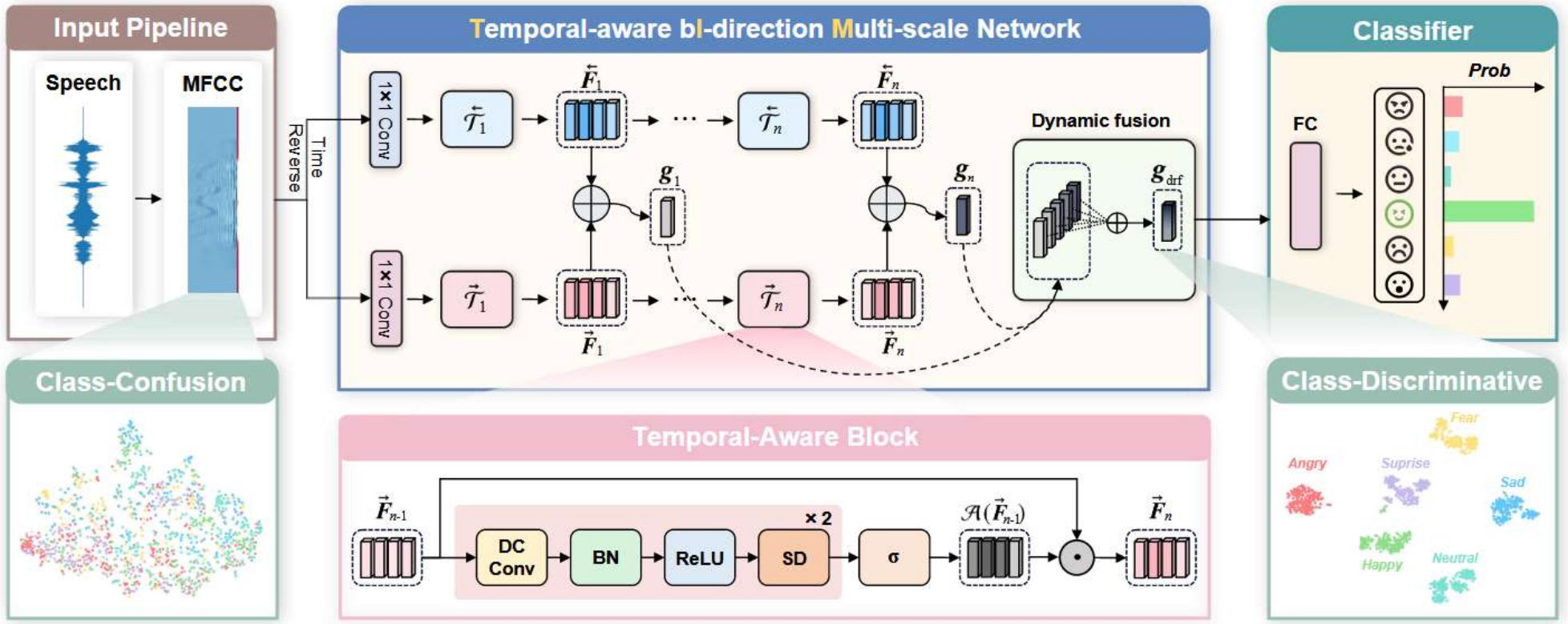
TIM-Net TIPO

T	目标	捕捉情绪的动态变化
I	输入	6个数据集（4种语言、58位说话者）
P	处理	1. 构建时间感知模块 2. 双向时间建模，从前向和后向学习长程情感依赖性，捕获帧级多尺度特征 3. 融合多尺度特征
O	输出	8种情绪分类

P	问题	1.现有方法未能有效解决动态情绪的准确识别 2.现有方法泛化能力弱，在不同数据集上效果差异大
C	条件	音频处理库预处理、6个不同的数据集
D	难点	1.如何学习长程情感依赖性 2.如何提高模型对未知数据或语料库的泛化能力
L	水平	2023 CCF B类

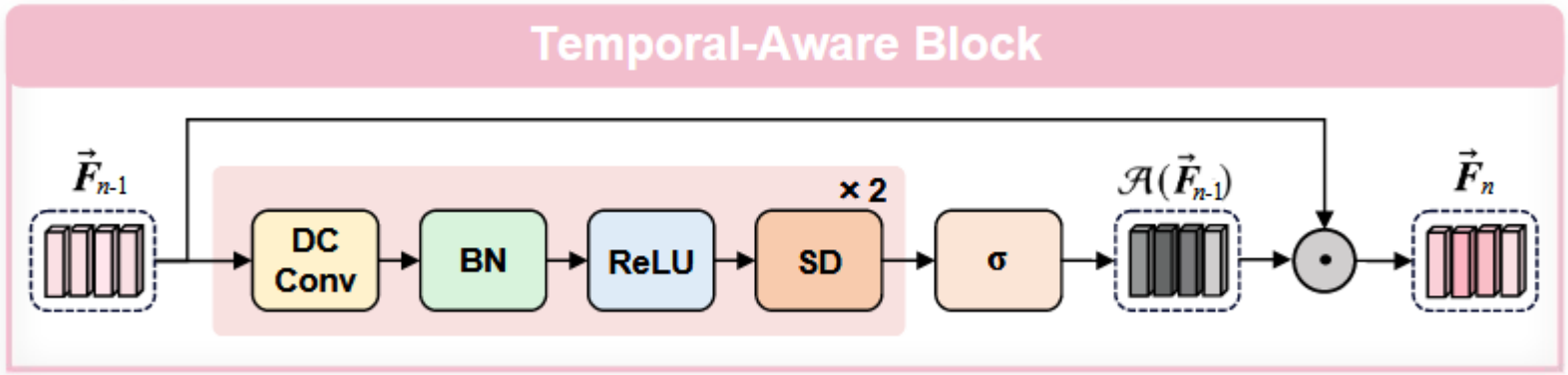
TIM-Net 算法原理

- 算法原理图
  - 构建时间感知模块
  - 双向时间建模，从前向和后向学习长程情感依赖性，捕获帧级的多尺度特征
  - 融合多尺度特征



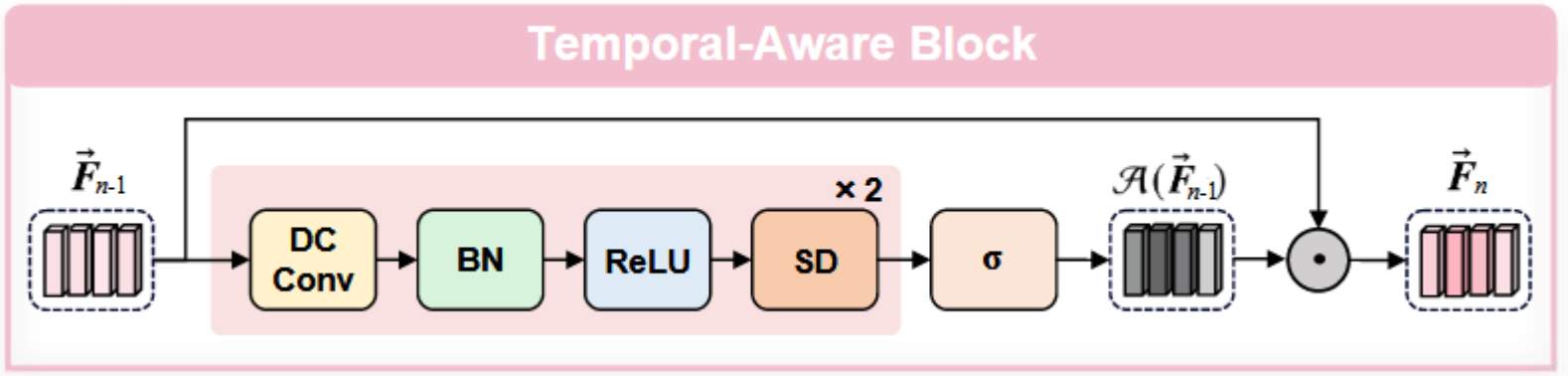
TIM-Net 创新点分析 时间感知模块(TAB)

- 现有方法存在问题
  - 缺乏足够的能力捕获情感的长时间依赖性
- 解决方法
  - 时间感知模块（TAB）
    - 膨胀因果卷积（DC Conv）
      - 膨胀卷积：通过指数递增（2^{j-1}）扩大感受野，能够捕捉长时间范围的时间依赖
      - 因果卷积：确保处理当前时间步的信息时不泄露未来的内容



TIM-Net 创新点分析 时间感知模块（TAB）

- 解决方法
  - 时间感知模块（TAB）
    - 特征归一化
    - 非线性激活
    - 空间dropout：随机选择整个图，将其全部置为零
    - 时间注意力机制：通过Sigmoid函数生成注意力权重，与输入的特征逐元素相乘，产生时间注意力特征



- 解决方法
  - 双向时间建模
    - 时间感知块的输出

$$F_{j+1} = A(F_j) \odot F_j$$

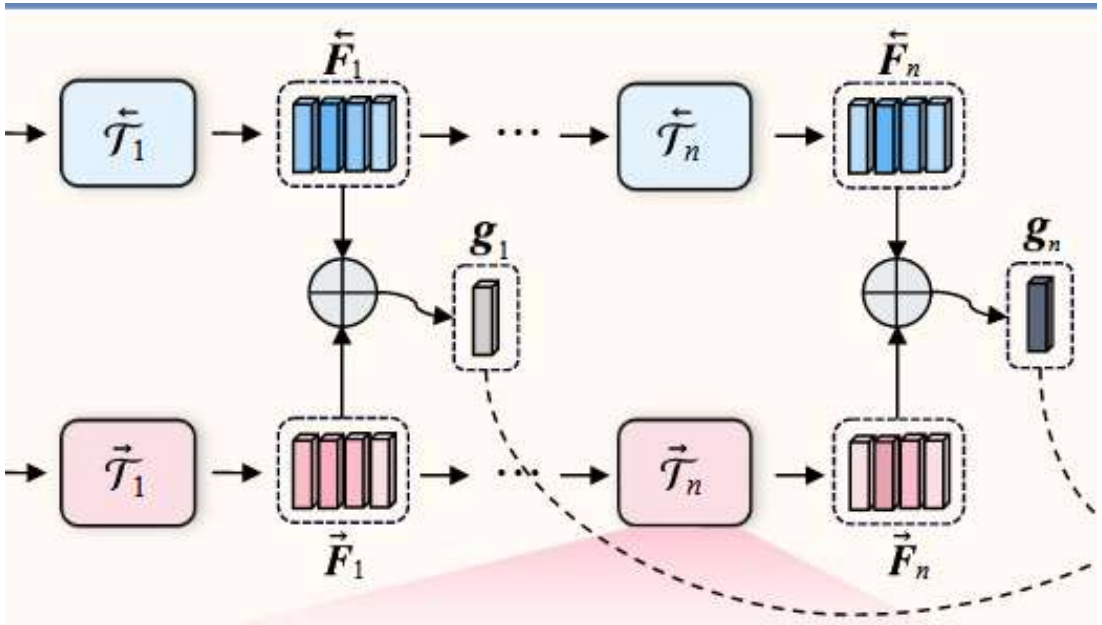
$$B_{j+1} = A(B_j) \odot B_j$$

$A(F_j)$注意力权重

    - 前向和后向信息融合

$$g_j = G(F_j + B_j)$$

$G(\cdot)$是全局时间池化，时间维度平均池化

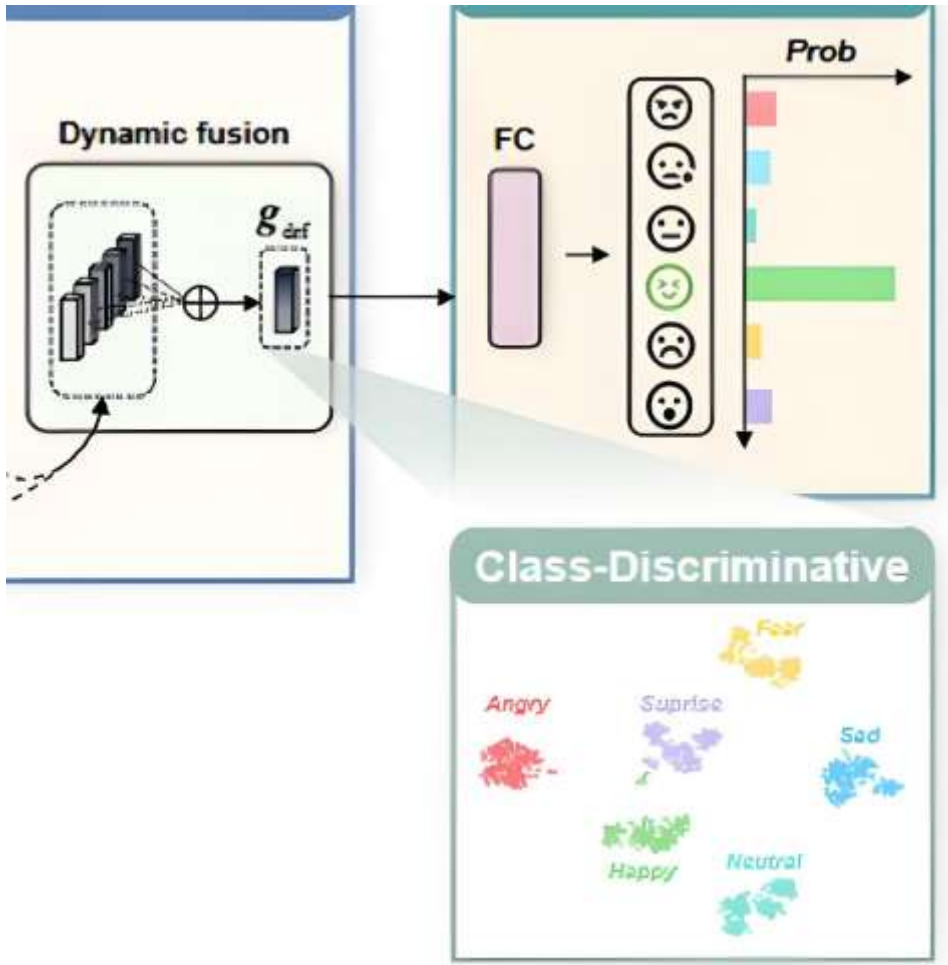


TIM-Net 创新点分析 多尺度动态融合

- 现有方法存在问题
  - 缺乏对未知数据或语料库的泛化能力
- 解决方法
  - 多尺度动态融合
    - 融合来自不同时间尺度的特征

$$g_{drf} = \sum_{j=1}^{n} w_j g_j$$

权重w_j是可训练参数



实验设计 数据资源

- 数据集

数据集名称	CASIA	EMODB	EMOVO	IEMOOCAP	RAVDESS	SAVEE
语言	中文	德语	意大利语	英语	英语	英语
数据来源	4位中国speaker	10位德国speaker	6位意大利speaker	10位美国speaker	24位英国speaker	4位英国speaker
情感类别	6	7	7	4	8	7

- 对比方法

  DT-SVM(2019)、TLFMRF(2020)、GM-TCN(2022)、CPAC(2022)、TSP+INCA(2021)、LigthtSER(2022)、CNN+INCA(2021)、GM-TCN(2022)

- 评价指标
  - UAR：非加权平均召回率
  - WAR：加权平均召回率

实验设计 对比实验

- 评估TIM-Net在不同数据集上的表现
  - 在不同数据集上准确率都有较大提升，没有过拟合问题

Model	Year	CASIA	Model	Year	EMODB	Model	Year	EMOVO
DT-SVM [12]	2019	85.08 / 85.08	TSP+INCA [2]	2021	89.47 / 90.09	RM+CNN [4]	2021	68.93 / 68.93
TLFMRF [13]	2020	85.83 / 85.83	GM-TCN** [14]	2022	90.48 / 91.39	SVM [15]	2021	73.30 / 73.30
GM-TCN** [14]	2022	90.17 / 90.17	LightSER** [16]	2022	94.15 / 94.21	TSP+INCA [2]	2021	79.08 / 79.08
CPAC** [17]	2022	92.75 / 92.75	CPAC** [17]	2022	94.22 / 94.95	CPAC** [17]	2022	85.40 / 85.40
TIM-Net*	2023	**91.08 / 91.08**	**TIM-Net***	2023	**89.19 / 90.28**	**TIM-Net***	2023	**86.56 / 86.56**
TIM-Net**	2023	94.67 / 94.67	**TIM-Net****	2023	95.17 / 95.70	**TIM-Net****	2023	92.00 / 92.00
Model	**Year**	**IEMOCAP**	**Model**	**Year**	**RAVDESS**	**Model**	**Year**	**SAVEE**
MHA+DRN [18]	2019	67.40 / -	CNN+INCA [3]	2021	- / 85.00	DCNN [19]	2020	- / 82.10
CNN+Bi-GRU [9]	2020	71.72 / 70.39	TSP+INCA [2]	2021	87.43 / 87.43	TSP+INCA [2]	2021	83.38 / 84.79
SPU+MSCNN [11]	2021	68.40 / 66.60	GM-TCN** [14]	2022	87.64 / 87.35	CPAC** [17]	2022	83.69 / 85.63
LightSER** [16]	2022	70.76 / 70.23	CPAC** [17]	2022	88.41 / 89.03	GM-TCN** [14]	2022	83.88 / 86.02
TIM-Net*	2023	**69.00 / 68.29**	**TIM-Net***	2023	**90.04 / 90.07**	**TIM-Net***	2023	**77.26 / 79.36**
TIM-Net**	2023	72.50 / 71.65	**TIM-Net****	2023	91.93 / 92.08	**TIM-Net****	2023	86.07 / 87.71

实验设计 消融实验

- 评估TIM-Net的不同模块在数据集上的表现
  - TCN：将TIM-Net替换为TCN
  - w/o-BD：删除后向TAB
  - w/o-MS：去除多尺度融合
  - w/o-DF：使用平均融合

  结论：所有组件对整体性能做出积极贡献，准确率提升

Method	TCN	w/o BD	w/o MS	w/o DF	TIM-Net
UAR_{avg}	80.45	84.92	85.45	84.85	**88.76**
WAR_{avg}	80.56	85.32	85.82	85.24	**88.97**

算法总结 TIM-Net

- 算法贡献
  - 构建时间感知模块（TAB）：更好的捕获时序中的变化，情感的长程依赖
  - 引入双向时间建模：将过去和未来的上下文信息整合，提升情感识别精度
  - 多尺度动态融合：在不同语速、停顿等变化下能够自动调整模型的时间感受野，提高模型在不同说话者、情感模式中的泛化能力

- 算法不足
  - 缺乏对非情感特征的鲁棒性（如说话者个性、背景噪声等）
  - 实时性不够强，双向建模处理前后向的信息流，会导致较高的延迟

特点总结与未来展望

特点总结与未来展望

- 特点总结
  - AMRWC
    - 建立注意力掩码残差网络（AM-ResNet），通过交叉注意力机制融合AM-ResNet和Wav2vec2.0的特征
      - 更好的解决了无声帧和清音会增加计算复杂度并降低情绪识别准确性和数据稀疏的问题
  - TIM-Net
    - 构建时间感知模块，引入双向时间建模，多尺度动态融合
      - 有效提升了动态情绪识别的准确率，增强了模型的泛化能力
- 未来发展
  - 如何在有噪声干扰的情况下进行时序建模
  - 如何做到实时语音情绪识别
  - 如何解决不同地区方言、口音等问题

参考文献

[1] Li X, Zhang Z. Cross-feature fusion speech emotion recognition based on attention mask residual network and Wav2vec 2.0[J]. Digital Communications and Networks, 2024.

[2] Ye J, Wen X C, Wei Y, et al. Temporal modeling matters: A novel temporal emotional modeling approach for speech emotion recognition[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.

[3] Akçay M B, Oğuz K. Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers[J]. Speech Communication, 2020, 116: 56-76.

道德经

知人者智，自知者明。胜人者有力，自胜者强。知足者富。强行者有志。不失其所者久。死而不亡者，寿。

谢谢！