arXiv:2302.02419v1 [cs.CL] 5 Feb 2023

对话语音情感识别的片段级特征表示深度学习

Jiachen Luo[1], *Huy Phan*[2*], *Joshua Reiss*[1]

[1]数字音乐中心，伦敦玛丽女王大学，英国[2]Amazon Alexa，马萨诸塞州剑桥，美国

摘要

由于对话中情绪及其动态变化错综复杂，准确检测对话中的情绪是一项必要但具有挑战性的任务。说话者的情绪状态可能受到许多不同因素的影响，例如对话者刺激、对话场景和话题。在这项工作中，我们提出了一种对话语音情感识别方法，用于捕捉注意力上下文依赖关系和说话者敏感交互。首先，我们使用预训练的 VGGish 模型提取单个话语中的片段级音频表示。其次，采用注意力双向门控循环单元（门控循环单元），以动态方式对上下文敏感信息进行建模，并联合探索说话者内及说话者间依赖关系。在标准对话数据集 MELD 上进行的实验表明，与最先进的方法相比，所提出的方法效果显著。

索引词— 上下文信息、情感计算、说话者敏感

1. 引言

人类情绪自动识别在对话生成、社交媒体分析和人机交互等领域有着广泛的应用 [1]。语音是人们能够清晰、直观地感知情绪变化的主要交流媒介。与常规的句子/话语情绪识别不同，对话情绪识别（ERC）理想情况下依赖于从包含两个或更多交谈者的对话中挖掘人类情绪，并且需要对单个话语进行上下文建模，还需要对单个话语进行上下文建模 [2]。如何从语音信号中捕捉这类信息是一项具有挑战性的任务。

在这项工作中，我们聚焦于交互式对话中的语音信号。语音信号天然能够承载说话者的情绪特征。传统上，对话情绪识别通常需要具备强大的能力，以建模上下文敏感属性、筛选关键信息并捕捉说话者敏感的依赖关系 [3]。在所有因素中，说话者信息对于追踪对话中的情绪特征至关重要，尤其是说话者内及说话者间依赖关系。

在互动式对话中，这些因素会带来多种多样的情绪动态。图 1 展示了多模态 EmotionLines 数据集（MELD）中的一些示例，呈现了这类模式。对话 (a) 展现了情绪惯性的存在，即

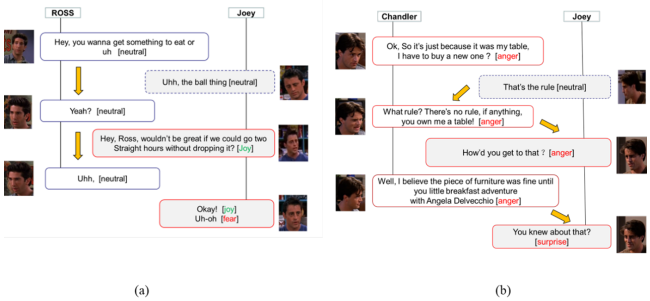

图 1。对话中说话者情绪动态的对比。

说话者对自身的影响。角色 Ross 不受另一位说话者的影响，始终保持中性情绪状态。另一方面，对话 (b) 指的是说话者之间的依赖关系，即对话对象会影响说话者。“我得买个新的？”表现出负面态度。带有特定的愤怒的语音色彩时，这句话会影响受话者／听者的感受。“你怎么会得出这个结论？”向听者强调，Joey 受到说话者 Chandler 回应所表达情绪的影响。

要对这类对话建模，架构需要应对以下挑战：如何捕捉说话者自身及说话者之间的依赖关系，以刻画情绪动态；以及如何根据对话流程中的上下文信息解读潜在情绪。此外，原始情绪还可能根据相邻话语中的上下文信息得到增强、减弱或逆转 [5]。对于话语级语音情绪识别，将语音压缩为话语级特征会导致动态时序信息和短期情绪动态的丢失，这是一个根本问题 [6]。然而，对于单个话语中基于片段的特征表示、对话中的上下文敏感信息及说话者影响，关于分析其情绪估计方面的研究进展甚微。Devamanyu等人使用文本模态特征，在对话情感识别任务中对上下文信息进行建模，以刻画说话者自身及说话者之间的情绪影响 [7]。

本文提出了一种方法，能够在情绪感知语音对话系统中，实现话语内各片段局部和全局上下文信息的协同演化，并适应自我影响、说话者内部状态和说话者间状态。我们首先使用预训练的 VGGish 模型提取话语中的片段级音频表示。随后，采用统计策略确定语音中话语级信息的情绪动态。为动态整合上下文依赖和说话者敏感交互，我们采用双向门控循环单元 对这些关系进行建模。总体而言，我们的贡献总结如下：

*这项工作是在 H. Phan 就职于英国伦敦玛丽女王大学数字音乐中心期间、加入亚马逊之前完成的。感谢中国国家留学基金管理委员会和伦敦玛丽女王大学提供资助。

• 片段级特征提取策略能够为对话系统提供细粒度的时间情绪表示。

• 我们采用双向门控循环单元层来捕捉上下文敏感信息以及对话情绪识别中的说话者内及说话者间依赖关系，并结合注意力机制突出重要的全局上下文话语。

• 结果表明，我们提出的方法在对话情绪识别方面优于当前最先进的方法。

2. 现有文献

语音情感识别系统中的全局和局部音频特征通常分为以下四类：韵律特征、频谱特征、嗓音质量特征，以及基于 Teager 能量算子的特征 [8]。传统上，许多频谱特征通常采用现有的倒谱表示之一来描述。常用的方法包括梅尔频率倒谱系数（MFCC）或梅尔尺度频谱图；一些研究还使用了共振峰及其他信息 [8]。Suraj 等人展示了卷积神经网络结合 MFCC 进行情感分类的有效性 [5]。此外，直接使用梅尔尺度频谱图进行对话情感识别也被证明是有效的 [9]。在本研究中，我们使用预训练模型从低层梅尔尺度频谱图中提取高层声学特征，以进行情绪识别。

对话情感识别需要深入理解对话中的人际互动 [10~13]。一些重要研究指出，情绪动态是一种交互现象 [13,16],，而非仅存在于个体内部。我们在模型设计中利用了这一特性，在对话中纳入说话者间的动态变化。由于对话具有天然的时间属性，上下文也在情绪分析中发挥着至关重要的作用 [4,17]。Poria 等人采用双向 LSTM，从同一说话者的相邻话语中捕捉时间上下文信息，以推断情绪 [4]。然而，现有方法并未对上下文和说话者间的交互影响进行建模。

需要指出的是，我们的方法不同于以往的对话情感识别策略。我们提出通过层次循环网络捕捉这种对上下文敏感的信息。此外，我们提出的方法采用一种交互机制，主动对对话中的说话者内部及说话者之间的情绪动态进行建模。

3. 方法

我们提出的方法由三个主要模块组成：1）预训练的片段级音频表示；2）分组并行统计操作，称为统计单元（SU）；3）对话感知交互模块，旨在对对话中的交互进行建模，并据此预测情绪状态。

3.1. 预训练音频嵌入

在音频端，将时长为 0.96 秒的音频片段转换为对数梅尔频谱图，并使用从预训练 VGGish 模型最后一个全连接层提取的 128 维特征向量进行编码。特别是，较短的样本在转换为对数梅尔频谱图之前会进行零填充。分段时使用非重叠片段。该模型在 AudioSet 上训练，该数据集包含 1 亿个 YouTube 视频 [14]。对于每个话语，都会生成一个由 L 个嵌入向量组成的序列，其中L 表示音频信号被划分成的音频片段数量（见图 2）。

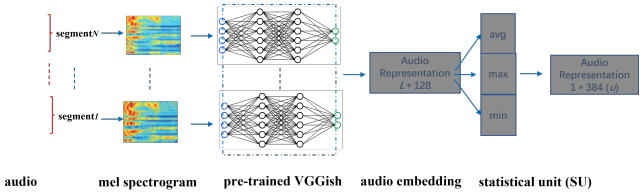

图2。将从预训练 VGGish 模型中提取的片段级嵌入进行聚合，以生成话语级嵌入。

3.2. 统计单元

我们使用一个统计单元，沿序列方向并行计算三种一维统计量，以缩减由 L 个片段级嵌入向量组成的序列，并生成话语级嵌入向量：平均值、最大值和最小值，如图 2 所示。最后，将它们拼接成一个特征向量，用于话语级表示。

3.3. 模型

所提出的模型包含四个双向门控循环单元 分支，用于捕捉参与者的累积上下文、说话人内状态、说话人间影响和情绪状态（见图3）。

3.3.1. 注意力上下文状态

在对话情绪识别中，为了确定时间戳 t处某个话语的情绪状态，可以将其之前的话语视为其累积上下文。上下文状态沿着对话流程的序列存储并传递整体话语级信息。前一个话语的上下文状态 C_{t-1}、说话人内状态 S_{t-1} 和说话人间状态 I_{t-1}，以及时间戳 t 处的音频表示u_t，用于将上下文信息从 C_{t-1} 更新至 C_t （见图3）。注意上下文状态更新 C_t 的步骤通过以下公式进行描述，并如图3所示。

$$C_t = GRU_C(C_{t-1}, (S_{t-1} \oplus I_{t-1} \oplus u_t)) \tag{1}$$

其中，$\oplus$ 表示拼接。在时间步 $t = 0$，上下文状态被随机初始化。

为了增强上下文丰富信息的贡献，我们采用来自历史交互上下文的软注意力，以结合长上下文中的说话者交互影响和对话依赖关系 [15]。我们使用软注意力从周围的上下文历史 $[C_1, C_2, \ldots, C_{t-1}]$ 中汇聚注意力向量 a_t。该上下文注意力向量 a_t 可按如下方式计算：

$$\begin{aligned} u_i &= \tanh(WC_i + b), \ 1 \le i \le t-1, \\ \alpha_i &= \frac{\exp(u_i^{\mathsf{T}})}{\sum_{i=1}^{t-1} \exp(u_i^{\mathsf{T}})}, \\ a_t &= \sum_{i=1}^{t-1} \alpha_i C_i. \end{aligned} \tag{2}$$

3.3.2. 自说话人状态

自我影响取决于说话者在对话中维持情绪的倾向。这种状态也称为情绪惯性，因为说话者不一定总会通过反应明确表达自己的感受或观点。具体而言，自我影响仅涉及说话者本人。

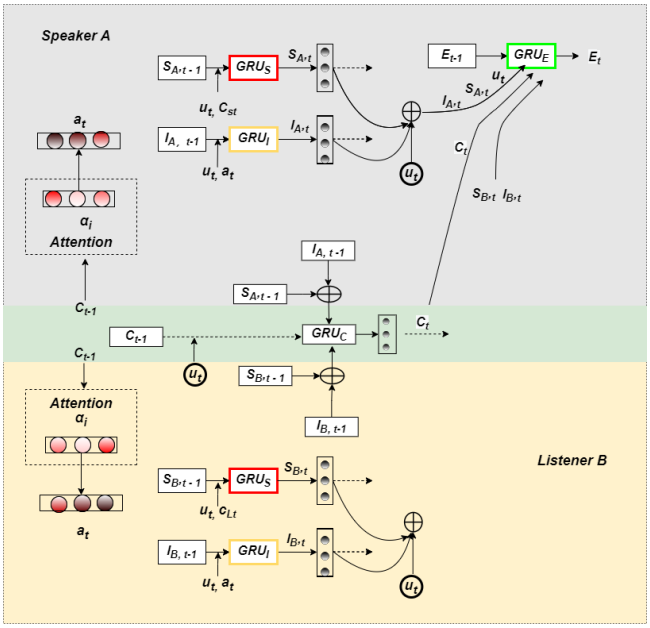

图3。我们提出的用于对话情感预测的模型整体架构。

自我影响模块由两个 $GRU_{S\lambda}$组成：分别是 GRU_{SS} 和 GRU_{SL}。对于 $\lambda \in \{S, L\}$， S 表示说话者， L 表示听者。 $GRU_{S\lambda}$ 旨在记忆 P_λ 的情绪惯性，这体现了个体对自身先前状态的情感依赖。一次对话包含两方：说话者，以及归属于听者的其他人。在时间步 t，个体 $S_{\lambda t}$ 的自说话人状态由该个体 $S_{\lambda(t-1)}$先前的自说话人状态、 $C_{\lambda t}$累积上下文向量以及 $u_{\lambda t}$话语更新。在时间步 t，自说话人状态 $S_{\lambda t}$ 可计算如下：

$$S_{\lambda t} = GRU_{S\lambda}(S_{\lambda(t-1)}, (C_{\lambda t} \oplus u_{\lambda t})) \tag{3}$$

3.3.3. 说话人内状态

自然地，其他参与者很容易观察、感受到并理解说话人内状态。更具体地说，这种状态通常与表达、反应和回应有关 [3]。由于话语之间会不断相互影响，我们构建了一个注意力交互模块，称为注意力交互依赖。对于时间 t 的话语，说话人内 I_t 会根据先前的说话人内状态 I_{t-1}、注意力上下文向量 a_t和话语u_t进行更新。在时间步 t，说话人内状态 I_t 可计算如下：

$$I_t = GRU_I(I_{t-1}, (a_t \oplus u_t)). \tag{4}$$

3.3.4. 情感状态

情感状态表示话语的情绪及其情绪类别。对于时间为 t 的话语，情感状态 E_t 取决于前一个情感状态 E_{t-1}，以及注意力上下文状态 C_t、自说话人状态 S_t、说话人内部状态 I_t 和该话语 u_t 的组合。然后，情感状态 E_t 可计算如下：

$$E_t = GRU_E(E_{t-1}, (C_t \oplus S_t \oplus u_t)). \tag{5}$$

3.3.5. 分类

最终输出的情感状态 E_t 输入两个带有残差连接的全连接层。为训练模型，使用最后一层采用 Softmax激活的分类交叉熵损失作为损失函数，并通过在代价函数中添加惩罚项来应用 L2正则化。

4. 实验

4.1. 数据库与指标

我们使用了多模态多说话者对话数据集，即多模态 EmotionLines 数据集（MELD）[4]。MELD 包含电视剧《Friends》中 1433 段对话、共 13798 个话语的声学、文本和视觉信息。该数据集包含七种情感类别：愤怒、厌恶、悲伤、喜悦、中性、惊讶和恐惧。数据集分为训练集、验证集和测试集，分别包含 9989、1109 和 2610 个话语 [4]。在本研究的相关实验中，我们仅使用声学模态 [4]。鉴于各种情绪的分布存在自然的不平衡，我们选择加权平均F1值作为评估指标。

4.2. 基线方法与最先进方法

实验中总共比较了 6 种最先进方法，以验证我们所提出方法的有效性（表1）。bc-LSTM 是传统的上下文相关情感分析方法 [4]。CMN [13], 、ICON[16] 和 DialogueRNN [17] 主要用于建模说话者动态，而 M2FNet [18] 和 MMTr [19] 则是基于注意力的方法。下面简要介绍这 6 种对比方法：

- 双向上下文长短期记忆网络（bc-LSTM）利用话语级长短期记忆网络学习上下文依赖。然而，上下文长短期记忆网络模型无法处理说话人间依赖关系 [4]。
- 对话记忆网络（CMN）利用说话人相关门控循环单元，从对话历史信息中提取话语上下文。随后，利用基于注意力的跳跃合并这些记忆，以捕捉说话人间依赖关系 [13]。
- 交互式对话记忆网络（ICON）在 CMN 的基础上扩展，通过交互式记忆网络对自我和说话人间的情感影响进行建模，并存储上下文摘要[16]。
- DialogueRNN 利用 门控循环单元 捕捉对话过程中参与者的情绪状态，以及说话人之间的句子上下文表示 [17]。
- M2FNet 采用基于多头注意力的融合机制，学习音频、文本和视觉模态中富含情绪的潜在信息 [18]。
- MMTr 获取说话者自我上下文和上下文语境两个层面的情绪线索，并学习多个模态之间的信息交互 [19]。

4.3. 声学特征

音频文件被重采样至 16千赫。为了提取梅尔频谱图，首先使用短时傅里叶变换计算线性频谱图，采用 25 毫秒的汉明窗和 10 毫秒的重叠。随后，在 125–7500 Hz 的频率范围内应用一组 64 个梅尔滤波器。然后将这些特征分割成互不重叠、时长为 0.96 秒的样本。特别是，

表1：在 MELD 上与最先进方法和基线方法的性能比较

方法	愤怒	厌恶	Fear	Joy	神经网络	悲伤	惊讶	加权平均F1（%）
bc-LSTM	21.9	0	0	0	66.1	0	16	36.4
CMN	29.6	0	0	11.8	67	0	2.8	38.3
ICON	31.5	0	0	8.6	66.9	0	0	37.7
DialogueRNN	32.1	5.1	0	11.2	53	8.3	15.6	34
M2FNet	-	-	-	-	-	-	-	39.6
MMTr	-	-	-	-	-	-	-	38.8
提出的方法	**32.6**	**0**	**0**	**25.21**	**63.85**	**6.28**	**14.58**	**40.9**

较短的样本会进行零填充，然后转换为对数梅尔频谱图。最后，将每个片段输入预训练的 VGGish 模型，得到一个 128 维嵌入向量，用于识别情绪。

4.4. 模型配置

我们使用 PyTorch 1.11.0 框架实现了所提出的模型。模型采用 Adam 优化器进行训练，初始学习率为 1e-4，批量大小为 32。损失函数采用交叉熵损失。为防止过拟合，我们使用模型参数的 L2范数对网络进行正则化，权重设为 3e-4。

5. 结果与讨论

表1展示了所提出模型的总体结果以及此前的最优结果。我们的方法取得了 40.9% 的 F1分数。与当前最优方法和基线方法相比，所提出的模型表现最佳，并在愤怒和喜悦类别上取得了显著提升（表1）。消融研究（见表2）验证了说话人间和说话人内状态的重要性：在对话情绪识别中，同时考虑注意力上下文敏感依赖、说话人内影响和说话人间影响，比单独考虑单一说话者的上下文信息更有效。此外，移除注意力上下文状态模块后，模型性能会下降，这表明，与建模说话者内敏感交互相比，建模注意力长期上下文依赖更为关键。

表 2：MELD数据集上的消融研究

方法	加权平均 F1
无预训练 VGGish	36.1
无注意力上下文状态	37.5
无自说话人状态	38.9
无说话者内状态	38.3

在定量方面，我们的方法采用基于片段的特征表示进行话语级分类。情绪持续时间较短，大多数仅持续几秒。因此，逐片段的方法有助于捕捉时间信息和短期情绪交互。此外，预训练的 VGGish 模型保留了大量先验知识，有助于更好地提取高层情感音频特征。具体而言，在数据集规模有限且类别分布不均衡的情况下，使用预训练模型尤为有效。

此外，我们的模型将来自周围话语历史的注意力上下文表征融入自说话人状态

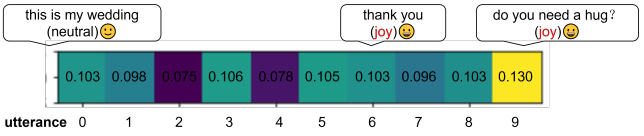

图 4。MELD 中案例的模型注意力权重可视化。

以及说话人内影响，以捕捉多轮对话中的情绪动态。在这方面，它比 bc-LSTM 等先前方法更具优势，因为后者往往无法判断这类情况。例如，当 Jill 用带有喜悦和兴奋的上扬语调告诉 Bob“我很快就要结婚了”时，Bob 的情绪会从中性突然转为开心。注意力机制用于强化重要的全局对话上下文依赖（见图 4）。此外，说话人之间的互动可能是同步的（例如，听到好消息后欢呼），也可能是异步的（例如，听到有趣的事情后发笑）。进一步分析我们的方法为何能做出更准确的预测，或许有助于阐明注意力上下文依赖和说话人敏感影响。

6. 结论

本文提出了在对话中聚合分段级语音频谱图，以进行话语级情绪分类。该方法利用推断上下文信息，纳入动态的自我影响、说话者内影响和说话者间影响。我们采用基于注意力的机制，从周边话语历史中确定重要的上下文相关信息。双向门控循环单元 用于捕捉上下文依赖、自我影响和说话者间影响。实验结果证明了所提模型的有效性。

我们的方法存在一些局限性。为了拓展数据集的范围并推动对话情感识别研究，还需要进一步探索其泛化能力，并在其他若干基准数据集上开展详细的消融研究和分析。此外，众所周知，语音中的文本、面部表情、身体动作及其他模态都能传达情绪，因此可以提供额外信息。一项很有前景的后续研究方向是应用近期发展的方法，将多模态系统中的知识迁移到单模态系统，从而提升单模态系统的性能。

7. 参考文献

[1] Wang W, Song W, Tao A. “情感计算系统综述：情绪模型、数据库及最新进展，”*InformationFusion*，2022，83-84:19-52。[2] Aneesh M, Martin R. “使用四元数卷积神经网络进行语音情感识别，”*IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，2021，6309-6313。[3] Sun L, Liu B, Tao JH, Zheng L.“用于语音情感识别的多模态交叉与自注意力网络，”*IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，2022，4275-4279。[4] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. “MELD：用于对话情绪识别的多模态多方数据集，”第*57*届计算语言学协会年会论文集，2019，527-536。[5] Jun-Heng T-L, Ching, Yaoi, Yu-Te. “从连续普通话语音中进行基于片段的情绪识别，” 人类行为中的计算机，2011，27: 1545-1552。[6] Mao S, Ching PC. “结合多实例学习的深度学习片段级特征表示，用于话语级语音情感识别，” *Interspeech*，2019，1686-1690。[7] Ghosal D, Majumder N, Gelbukh A, Mihalcea R, Poria S. “COSMIC：用于对话情绪识别的常识知识，”计算语言学协会年会论文集（*EMNLP*），2020。[8] Akc.ay MB, O˘guz K. “语音情感识别：情绪模型、数据库、特征、预处理方法、辅助模态及分类器，” *SpeechCommunication*，2020，116 56 76: - 。[9] Neil S, Mikolaj K, Pierre B, MilosC. “SERAB：语音情感识别多语言基准，” *IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，2022，7697-7701。[10] Sidorov M, Minker W. “利用声学与视觉特征进行情绪识别和抑郁症诊断，” 第*4*届音频/视觉情绪挑战国际研讨会论文集，2014，81-86。[11] Poria S, Cambria E, Hazarika D, Majumder N, Zadeh A, Morency L-P. “用户生成视频中的上下文相关情感分析，” 第*55*届计算语言学协会年会论文集，2017，873-883。[12] Hazarika D, Poria S, Zadeh A, Cambria E, Morency LP, Zimmermann R. “用于双人对话视频情绪识别的对话记忆网络，” *2018*年北美计算语言学协会分会会议：人类语言技术论文集，2018，2122-2132。[13] Ghosal D, Majumder N, Poria S, Chhaya N, Gelbukh A. “DialogueGCN：用于对话情绪识别的图卷积神经网络，” 计算处理与第*9*届自然语言处理国际联合会议，2019，154-164。

[14] Gemmeke JF, Ellis D, Freedman D, Jansen A,Ritter M. “Audio Set：音频事件本体与人工标注数据集”，*IEEE* 国际声学会议，2017，776-780。[15] Xia X, Jiang D, Sahli H. “利用注意力时间池化学习语音情感识别的显著片段”，*IEEE Access*，2020，8:151740-151752。[16] Hazarika D, Poria S, Mihalcea R, Cambria E, Zimmermann R. “ICON：用于多模态情感检测的交互式对话记忆网络”，*2018*年自然语言处理实证方法会议论文集，2018，2594-2604。[17] Majumder N, Poria S, Hazarika D, Mihalcea R, Cambria E. “DialogueRNN：用于对话情感检测的注意力循环神经网络”，第三十三届 *AAAI* 人工智能会议论文集，2019，6818-6825。[18] Vishal C, Purbayan K, Ashish G, Nirmesh S, Pankaj W, Naoyuki O.”M2FNet：用于对话情绪识别的多模态融合网络”，计算机视觉与模式识别会议研讨会 *(CVPRW)*，2022，4652 4661-。[19] Zou SH, Huang XY, Shen XD, Liu HK. ”采用主模态变换器改进对话情绪识别中的多模态融合”，*Knowledge-Based Systems*，2022，258，1-9。