arXiv:2305.03369v1 [cs.LG] 5 May 2023

MuSe 2023 多模态情感分析挑战赛：模仿情绪、跨文化幽默与个性化

Lukas Christ EIHW，奥格斯堡大学，奥格斯堡，德国

Shahin Amiriparian
EIHW，奥格斯堡大学
奥格斯堡，德国

Alice Baird
Hume AI
纽约，美国

Alexander Kathan
EIHW，奥格斯堡大学 奥格斯堡，德国

Niklas Müller
帕绍大学
帕绍，德国

Steffen Klug
帕绍大学
帕绍，德国

Chris Gagne
Hume AI 纽约，美国

Panagiotis Tzirakis
Hume AI
纽约，美国

Eva-Maria Meßner
乌尔姆大学
乌尔姆，德国

Andreas König 帕绍大学 帕绍，德国

Alan Cowen
Hume AI
纽约，美国

Erik Cambria
南洋理工大学
新加坡，新加坡

Björn W. Schuller
GLAM，帝国理工学院 伦敦，英国

摘要

Mu模态**Se**情感分析挑战赛（MuSe）2023 是一组共享任务，旨在解决当代多模态情感与情感分析领域的三个不同问题：在模仿情绪子挑战赛（MuSe-Mimic）中，参赛者需要预测三个连续情绪目标。该子挑战赛采用由用户生成视频构成的 Hume-Vidmimic 数据集。在跨文化幽默检测子挑战赛（MuSe-Humour）中，提供了 Passau 自发足球教练幽默（Passau-SFCH）数据集的扩展版本。参赛者需要在跨文化环境中预测自发幽默的存在。个性化子挑战赛（MuSe-Personalisation）基于乌尔姆-特里尔社会压力测试（Ulm-TSST）数据集，其中包含受试者处于压力情境下的录音。在该任务中，需要预测唤醒度和效价信号，同时提供部分测试标签，以促进个性化。MuSe 2023 旨在汇聚来自不同研究社区的广泛受众，包括视听情绪识别、自然语言处理、信号处理和健康信息学等领域。在这篇基线论文中，我们介绍了所提供的数据集、子挑战赛以及所提供的特征集。作为一种具有竞争力的基线系统，采用门控循环单元（GRU）-循环神经网络（RNN）。在各自子挑战赛的测试数据集上，该系统在 MuSe-Mimic 中取得了 0.4727 的皮尔逊相关系数（对三个连续强度目标取平均），在 MuSe-Humour 中取得了 0.8310 的曲线下面积（AUC），并在 MuSe-Personalisation 子挑战赛中分别取得了唤醒度 0.7482 和效价 0.7827 的一致性相关系数（CCC）。

允许出于个人或课堂使用目的，免费制作本作品的全部或部分内容的数字副本或纸质副本，前提是这些副本不得用于营利或商业利益目的，不得为此进行复制或分发，并且副本首页必须包含本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的组成部分，必须尊重其版权。注明出处后，可以进行摘要引用。如需以其他方式复制、重新发表、发布到服务器或分发至列表，则必须事先获得明确许可和/或支付相应费用。请通过 permissions@acm.org 申请许可。第四届多模态情感分析挑战赛与研讨会，*2023*年*10*月，加拿大渥太华© 2023 版权所有归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

CCS 概念

• **计算方法** → **神经网络**；**人工智能**；计算机视觉；自然语言处理。**关键词**

多模态情感分析；情感计算；幽默检测；情绪识别；多模态融合；挑战赛；基准

ACM 参考格式：
Lukas Christ、Shahin Amiriparian、Alice Baird、Alexander Kathan、Niklas Müller、Steffen Klug、Chris Gagne、Panagiotis Tzirakis、Eva-Maria Meßner、Andreas König、Alan Cowen、Erik Cambria 和 Björn W.Schuller。2023。《MuSe 2023 多模态情感分析挑战赛：模仿情绪、跨文化幽默与个性化》。载于《*MuSe 2023* 论文集（第四届多模态情感分析挑战赛与研讨会）》。ACM，美国纽约，10 页。
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 引言

在第 4 届会议中，多模态情感分析（MuSe）挑战赛提出了三项不同的任务，即类别情绪

预测、跨文化幽默检测，以及个性化维度情绪回归。在情绪模仿子挑战赛（MuSe-Mimic）中，我们通过推出首个大规模（557 名受试者， 30+ 小时）、多模态（音频、视频和文本）数据集来探索情绪模仿。这些数据是在自然环境中收集的：受试者通过网络摄像头录制自己对各种“种子”视频的面部和声音模仿。受试者从提供的 63 个类别中选择自己在每段视频中感知到的情绪，此外还可选择“此视频中看不到或听不到任何人”，并以 0–100 的强度等级为每种选定的情绪评分。在这一子挑战赛中，参赛者将采用多输出回归，根据受试者针对与决策相关的情绪类别所录制的多模态回应，预测三种自我报告情绪的强度：认可、失望和不确定性。

对于跨文化幽默检测子挑战赛（MuSe-Humour），参赛者将训练模型，从德国足球新闻发布会录音中预测幽默。与训练数据不同，未见过的测试集由英语足球新闻发布会视频组成，从而提供了跨文化评估场景。Passau 自发性足球教练幽默（Passau-SFCH）数据集 [18]作为 MuSe 2022 [17, 38, 63] 中的数据集提供，用作训练集。对于测试集，我们在 Passau-SFCH 的基础上加入了 7 名来自英格兰足球超级联赛的不同教练于 2016年9月至2020年9月期间进行的新闻发布会录音。训练集和测试集都只包含相应教练发言的片段，总计超过 17 小时的数据。尽管所有提供的视频最初都依据 Martin 等人提出的幽默风格问卷（HSQ） [42], 进行了标注，但 MuSe-Humour 的预测目标是二元的，即幽默的存在或缺失。

在新增的个性化子挑战赛（MuSe-Personalisation）中，参赛者需采用个性化方法，预测效价和唤醒度的连续估计值。与近年来常用的说话人独立挑战赛设置不同，参赛者还将获得测试分区中每位受试者的带标签数据。因此，MuSe-Personalisation 鼓励探索如何将多模态情绪识别系统适配到不同个体，并将其特定特征纳入考量。MuSe-Personalisation 使用 MuSe 2021 [54, 56] 中介绍的乌尔姆-特里尔社会压力测试数据集（Ulm-TSST）。Ulm-TSST 包含 69 名参与者接受特里尔社会压力测试（TSST）时的录制数据。该测试是一种旨在诱发压力的情境。除音视频录制及其文本转录外，Ulm-TSST 还包括受试者的皮肤电活动（EDA）、心电图（ECG）、呼吸（RESP）和心率（BPM）信号，所有信号的采样率均为 1 kHz。

凭借多样化的数据和预测目标，这三个子挑战赛面向广泛的受众，包括但不限于对情感计算、多模态表征学习、自然语言处理、机器学习和信号处理感兴趣的研究人员。MuSe 旨在成为一个交流平台，用于比较针对所提任务的不同方法，从而深入了解不同方法、模态和特征的适用性。

第2节将更详细地介绍上述子挑战、其对应的数据集以及挑战赛协议。

表1：各子挑战数据集的统计信息。给出了唯一受试者的数量（#）以及采用时:分:秒格式表示的视频时长。请注意，对于 MuSe-Personalisation，测试受试者的部分录制内容也包含在训练集和开发集中，如圆括号所示。

	MuSe-Humour		MuSe-Mimic		MuSe-Personalisation	
分区	#	时长	#	时长	#	时长
训练集	7	7 :44 :49	328	19 :24 :23	41 (+14)	3 :39 :56
开发集	3	3 :06 :48	107	6 :22 :22	14 (+14)	1 :20 :10
Test	6	6 :35 :16	122	6 :33 :10	14	0 :47 :21
∑	16	17 :26 :53	557	32 :19 :56	69	5 :47 :27

接下来，第 3 节介绍我们的预处理与特征提取流程，以及用于计算各子挑战赛基线结果的实验设置。随后，第 4 节展示实验结果，最后第 5 节对全文进行总结。

2 三项子挑战

接下来，我们将详细介绍各项子挑战、数据集以及挑战赛协议。

2.1 TheMuSe-Mimic 子挑战

在情感计算领域，获取有关人类表达行为的数据仍然是研究人员面临的一项艰巨任务。为了解决这一问题，研究人员采用了各种策略，包括在真实环境中模仿人类表情的技术 [12]。这种方法能够更高效地针对特定情绪进行研究。情绪模仿已在儿童发展过程中得到观察，通常会在社会互动中自发发生 [32]。此外，研究表明，人类具有高度精准的情绪模仿能力，甚至能够精确模仿特定特征 [33]。

为应对获取与人类表达行为相关数据这一挑战，现已推出一个名为 Hume-Vidmimic 数据集的多模态模仿数据集，供 MuSe-Mimic 子挑战赛使用。该数据集包含超过 30 小时的视频数据，平均时长为 6.4 秒，采集自美国的 557 名受试者（有关数据集划分详情，请参见表1）。数据集中的受试者年龄介于 19 至 59 岁之间，通过 Prolific 招募，并获得了参与时间的报酬。

在该数据集中，每位受试者都被要求模仿一段种子视频中的某人表达情绪的样子，然后从一组选定的情绪类别中选择并评定种子视频中情绪的强度。本子挑战的目标情绪表达为认可、失望和不确定性。这三个连续强度目标的标签，是通过对经过筛选的相关矩阵采用凝聚聚类方法确定的；该相关矩阵涵盖所有 63 个可用标签[1]（以及视频中未出现或未听到任何人的情况），相关阈值为 [1] 。

[1]钦佩、爱慕、审美欣赏、愉悦、愤怒、恼怒、焦虑、认可、敬畏、尴尬、无聊、平静、专注、自信、困惑、沉思、轻蔑、满足、渴望、决心、失望、不赞同、不适、厌恶、痛苦、怀疑、狂喜、尴尬、共情痛苦、热情、出神、嫉妒、兴奋、恐惧、挫折、感激、内疚、犹豫、惊恐、兴趣、喜悦、爱、怀旧、疼痛、恐慌、

对于每个样本，从 63 种情绪中剔除相关性较低的情绪，并对各情绪的强度进行归一化后，通过计算种子-模仿者配对在相应标签簇中的平均强度评分，确定目标标签。聚类结果如下：认可包含认可、钦佩、爱慕、热情、兴奋、喜悦和爱；不确定性包含不确定性、怀疑、浪漫、性欲、震惊和惊讶（负向）；最后，失望包含失望、愤怒、恼怒、轻蔑、不赞同、厌恶和挫折感。

数据准备过程包括将数据集按说话人独立划分为训练集、开发集和测试集。每个样本都提供了自动转录文本，并以每秒 2 帧的频率检测视频中人物的面部。

对于 MuSe-Mimic 子挑战赛，目标是根据从多模态（音频、视频和文本）数据中提取的特征，对 3 个情绪强度目标进行多输出回归。皮尔逊相关系数（ρ）用作评估指标。.

2.2 MuSe 跨文化幽默子挑战赛

幽默——被定义为一种旨在逗乐他人、通过建立出人意料或不协调的关系或含义来表达的形式 [27]——是人类社会互动中最复杂的现象之一，具有多种潜在的正面或负面影响 [13]。因此，幽默一直是情感计算和人机交互等领域的重点研究课题，例如自然语言接口 [10]。幽默既可以通过语言表达，也可以通过非语言方式表达，因此多模态方法尤其适用于幽默检测。目前已有多个用于多模态幽默识别的数据集 [30, 44, 62], , 但通常依赖 TED 演讲或电视剧等人为设计的场景。此外，一些研究将观众笑声等同于（成功的）幽默，这是一种常见却至关重要的缺陷。相比之下，据我们所知，Passau-SFCH 是唯一一个用于预测自发、非脚本化幽默表现，并采用细致幽默测量方式的数据库。

由于幽默融入了语言和语境因素，跨文化研究有助于理解幽默运用中的共性与差异。近年来，研究者开始深入探讨不同国家幽默的多模态复杂性，例如美国人和法国人在微笑行为表现上的差异 [50] ，以及德巴互动中用于构建幽默的手势和韵律等方面的差异 [36]。然而，据我们所知，迄今尚无自动化多模态跨文化幽默检测研究，而这类研究有助于揭示幽默的可迁移性。MuSe-Humour 旨在为这一主题提供初步见解。

参赛者将使用 Passau-SFCH 提供的 10 位德国足球甲级联赛足球教练的新闻发布会录音来训练模型。英语测试集包含新闻发布会录音内容，来自英格兰足球超级联赛的 6 位足球教练。然而，他们中只有一位以英语为母语，其余五位教练分别来自五个不同的国家（阿根廷、法国、德国、葡萄牙和西班牙）。训练集/开发集和测试集中的教练均为男性。训练对象的年龄介于 30 至 53 岁之间，测试受试者的年龄介于 47 至 57 岁之间。我们将德国教练划分为训练集和开发集，用于基线实验；其中开发集的划分与 2022年 MuSe-Humour 挑战赛 [2, 19] 的划分完全相同。数据集的详细统计信息见表1。

我们对所有视频进行分段，仅保留教练实际说话的部分。训练数据和测试数据均包含分段后的音视频录制内容，以及带时间戳的人工转录文本。首先，我们以 2 Hz 的频率对所有视频进行标注，标记幽默话语的情感及其指向（针对自身或他人），标注依据为幽默风格问卷 [42]。随后，我们按照 [19], 所述，获取表示幽默是否存在的二元标签，从而将视频划分为时长为 2 秒的片段，每个片段均被判定为幽默或非幽默。金标准中，训练数据、开发集和测试数据中被标记为幽默的部分分别占 4 38%、2 81% 和 6 17%。

与 2022年的 MuSe-Humour 子挑战赛类似，我们采用 AUC指标作为评估标准。

2.3 MuSe-Personalisation 个性化子挑战赛

在处理现实世界的数据时，常常可以观察到个体之间存在显著差异（例如，同一个人的语音音高变化幅度较大，或在人格特征方面存在差异，如性别、年龄范围或文化背景） [1, 37, 61]。然而，如今大多数方法往往忽略这些个体差异，导致模型基于广泛人群的数据进行训练，因此对训练集中未出现的受试者并不总能很好地泛化 [22]。为解决这一问题，需要采用能够纳入个体特征的个性化方法，从而构建能够提供更准确预测结果的个性化模型 [37]。MuSe-Personalisation 旨在成为多模态情感分析中个性化方法的基准。

本子挑战赛的参赛者将获得多模态 Ulm-TSST 数据集。该数据集包含受试者参与特里尔社会压力测试（TSST）的录制内容 [35], , 该测试设定了一种会诱发压力、类似求职面试的情境。参赛者需进行一场五分钟的即兴演讲，阐述自己为何适合某个假想职位。Ulm-TSST 收录了 69 名受试者（年龄范围为 18 至 39 岁）的演讲录制内容，其中 49 人为女性，总计约 6 小时的视频数据。

数据以说话人独立的方式划分为训练集、开发集和测试集。训练数据集包含41段视频，开发集和测试集则各包含14段视频。请注意，此划分与2021年和2022年情绪压力子挑战赛（MuSe-Stress）所采用的Ulm-TSST划分完全相同 [19, 54]。为了便于针对测试受试者进行个性化，我们提供其视频中带有标签的片段，具体如下：取每段测试视频的前60秒作为受试者特定训练数据；接下来的60秒则作为受试者特定开发集。每位测试受试者视频的剩余部分

自豪、领悟、宽慰、怨恨、浪漫、悲伤、讽刺、满足、平静、性欲、羞耻、震惊、惊讶（负面）、惊讶（正面）、同情、疲倦、胜利和不确定性。

将作为子挑战赛的测试数据，也就是说，这些片段的标注将予以保密。表1列出了Ulm-TSST这一划分的关键统计数据。

Ulm-TSST 已由三位标注者以 2 Hz 频率标注唤醒度和效价信号。金标准与 2022 年 MuSe-Stress 挑战赛中的金标准相同 [19]，该挑战赛吸引了相当多的关注，例如 [31, 40, 47, 64]。对于效价金标准，我们采用评定者对齐标注加权（Rater Aligned Annotation Weighting，RAAW）方法融合这三份标注 [57]。RAAW 将利用典型时间扭曲（Canonical Time Warping，CTW）[67] 进行的时间对齐，与通过评估者加权估计器（Evaluator Weighted Estimator，EWE）[28]进行的标注融合相结合。对于唤醒度，每段视频中，将一致性最高的两位标注者的信号与视频拍摄对象经降采样和平滑处理的皮肤电活动（EDA）信号合并，因为研究表明，与不可避免地带有主观性的标注相比，EDA 是唤醒度的客观指标 [15]，。金标准创建的详细信息见 [19]，；关于合并生物信号和标注信号以估计唤醒度的详尽实验，见 [9]。

我们向参与者提供视听录制内容、人工转录文本以及 ECG、RESP 和 BPM 信号。

2.4 挑战赛协议

要参加挑战赛，参与者需要具备学术机构 affiliation，并完成 MuSe 2023 主页 [2] 上提供的最终用户许可协议（ EULA ）。组织者不会以参赛者身份参加任何子挑战。比赛期间，参与者需在 CodaLab 平台[3]上提交针对测试标签的预测结果。每个子挑战最多可提交 5 个预测结果。我们鼓励所有参赛团队提交论文，介绍其试验和结果。若要正式赢得某个子挑战，必须提交一篇论文并获得录用。论文将接受双盲同行评审。

3 基线方法

我们提供一套丰富的预提取特征，以帮助参赛者高效地开发模型。对于每项子挑战赛，均提供七种特征集（3 种基于音频、3 种基于视频、1 种基于文本），具体如下[4]。

3.1 预处理

如第2节所述，我们将每个子挑战赛的数据集划分为训练集、开发集和测试集。在此过程中，我们考虑了时长、说话人独立性和标签分布等因素（参见表1）。对于 MuSe-Mimic，如表1所示，采用了 60-20-20% 的划分策略。未进行额外的分段处理。每个样本均包含一个种子视频的单次模仿，并将标签按样本归一化至 [0:1] 范围。

Passau-SFCH 新闻发布会的录音被切分为片段，每个片段中都有相应的教练在讲话。我们手动删除教练使用英语以外语言的少数片段。此外，我们还舍弃音频质量受到严重影响的片段。

Ulm-TSST 数据集中的录制视频经过剪辑，仅保留特里尔社会压力测试（TSST）协议所规定的求职面试陈述环节。为保护测试对象的隐私，我们还删除了他们透露姓名的视频片段。除第2节2.3节中所述的测试受试者视频分段处理外，MuSe-Personalisation 不进行进一步的分段处理。

3.2 音频

在提取音频特征之前，我们将所有音频文件归一化至 -3 分贝，并将其转换为单声道、16 kHz、16 位格式。随后，我们使用 openSMILE [25] 工具包计算人工设计的特征。此外，我们还利用 DeepSpectrum [4] 和 Wav2Vec2 的一种变体0 [7]计算高维音频表征。

这两种系统已被证明在基于音频的语音情感识别（SER）任务中具有重要价值 [6, 11, 26]。

3.2.1 eGeMAPS。 我们使用 openSMILE 工具包 [25][5], 提取 88 个扩展日内瓦极简声学参数集（eGeMAPS）特征 [24]，这些特征已被证明适用于情感分析和语音情感识别（SER）任务 [8, 59]。对于每个子挑战，我们均采用标准配置，以 2 秒为窗口大小、500ms 为步长提取特征。

3.2.2 DeepSpectrum。 我们使用 DeepSpectrum [4][6] 从音频数据中提取基于深度卷积神经网络（CNN）的表征。具体而言，我们首先为每个音频文件生成梅尔频谱图（128 个梅尔频带，采用 *viridis* 色彩映射），窗口大小为 1 秒，步长为 500 毫秒。随后，我们将频谱图表征输入 DenseNet121，并将最后一个池化层的输出作为 1024 维特征向量。研究已证明，DeepSpectrum 在语音和音频识别任务中效果显著 [3, 5, 46]。

3.2.3 Wav2Vec2.0. 近年来，自监督预训练 Transformer 模型在计算机听觉领域引起了广泛关注 [39]。此类基础模型中一个广受欢迎的例子是 Wav2Vec2.0 [7]，，它常用于 SER [45, 48]。由于所有子挑战都与情感有关，我们使用在 MSP-Podcast [41] 数据集上针对连续情绪识别 [60][7]进行微调的 Wav2Vec2.0 大型版本。我们通过对该模型最后一层中的音频信号表示取平均值来提取特征，得到 1024 维嵌入。对于 MuSe-Humour 和 MuSe-Personalisation，我们以 500 ms 的步长在每个音频文件上滑动 3 s 窗口，从而获得 2 Hz 特征。由于 MuSe-Mimic 中的片段相对较短，因此采用 2 s 的窗口大小。

3.3 视频

我们基于受试者的面部计算视觉模态的特征。因此，首先使用多任务级联卷积网络（MTCNN）自动提取面部。随后，从这些面部中提取面部动作单元（FAUs）、FaceNet512 和视觉 Transformer（ViT）表征。

[2]https://www.muse-challenge.org [3]https://codalab.lisn.upsaclay.fr/。相应挑战赛的链接将通过电子邮件发送给参赛者。[4]注意：参赛者可以使用其他外部资源（例如特征、数据集、预训练基础模型）。使用额外资源、工具等的情况，须在相应论文中明确说明。

[5]https://github.com/audeering/opensmile [6] https://github.com/DeepSpectrum/DeepSpectrum [7] https://huggingface.co/audeering/wav2vec2-large-robust-12-ft-emotion-msp-dim

3.3.1 MTCNN。我们采用 MTCNN [65] 人脸检测模型8，提取受试者的面部图像。

在 Passau-SFCH 的视频中，通常会出现好几个人，但只有教练与我们相关。首先，我们尝试通过对人脸嵌入进行聚类，自动筛选出每段视频中对应教练的人脸。之后，我们会手动校正由此得到的人脸集合，只保留与教练对应的人脸。对于 Ulm-TSST 和 Hume-Vidmimic，无需进行此类后处理。在这两个数据集的录像中，通常每段视频只显示一个人。随后，提取的人脸图像将作为基础，通过 Py-Feat、FaceNet512 和视觉 Transformer 提取特征。

3.3.2 面部动作单元。 面部动作单元 [23]通过参照特定面部肌肉的激活情况，为面部表情的编码提供了一种可解释的方法。由于面部表情包含有关个体情感状态的重要线索，面部动作单元已受到 AC 社区的广泛关注 [66]。我们通过 Py-Feat [9] 库中包含的相应支持向量分类器模型，自动估计 20 种不同面部动作单元的激活情况。

3.3.3 FaceNet512。 为了计算高维人脸表征，我们使用 deepface 库 [53]中实现的 FaceNet512 模型 [52]。FaceNet512 经过人脸识别任务训练，可为每张人脸生成一个 512 维嵌入。

3.3.4 视觉 *Transformer*（视觉 *Transformer*）。 作为一种基于视觉的替代方案，我们采用经过 DINO 训练的视觉 Transformer。该模型使用无标签自蒸馏（DINO）方法[14]，以自监督方式在 ImageNet-1K 数据集上进行预训练。该模型已展现出处理各种基于图像的任务的有效性，包括根据面部表情进行情绪识别 [16, 58]。模型处理提取的人脸图像，并为每张图像输出一个 384 维嵌入。我们不进行任何进一步的预训练或微调。

3.4 文本：Transformer

文本特征通过类似 BERT [21] 的预训练 Transformer 模型获得。我们取模型最后一层对 CLS 标记的编码，计算句子表示；CLS 标记是指代整个输入句子的特殊标记。对于 MuSe-Mimic，我们使用一个小型预训练 ELECTRA [20] 模型对英语转录文本进行编码，得到 256 维嵌入。由于 MuSe-Humour 的训练集和开发集为德语，而测试集为英语，因此我们选择在此使用多语言版本的 BERT [21][10]。该模型在包含德语和英语在内的 104 种语言的维基百科条目上进行预训练。研究表明，该模型能够有效地从一种语言泛化到另一种语言 [49]。对于 MuSe-Personalisation 仅有德语的转录文本，我们使用德语 BERT 模型[11]。上述两种 BERT 变体均生成 768 维表示。请注意，我们没有对任何模型进行微调。

[8]https://github.com/ipazc/mtcnn [9]https://py-feat.org [10] https://huggingface.co/bert-base-multilingual-cased [11] https://huggingface.co/dbmdz/bert-base-german-cased

3.5 对齐

每个数据集都包含音频、视频和转录文本。为了便于开发利用所提供特征的多模态模型，我们将不同特征彼此对齐，并与相应任务的标注方案对齐。

对于所有任务，音频和基于面部的特征均以 2 Hz 的频率提取：音频采用步长为 500 ms 的滑动窗口，视频则以 2 Hz 的频率采样面部。为了根据人工转录文本获取 MuSe-Humour 和 MuSe-Personalisation 的逐句时间戳，我们采用一个包含三个步骤的流程：首先，使用蒙特利尔强制对齐器（MFA） [43] 工具包生成词级时间戳。然后，通过 deepmultilingualpunctuation 工具 [29]自动为转录文本添加标点。最后，使用 PySBD [51] 将转录文本切分为句子，从而根据词级时间戳推断逐句时间戳。随后，我们通过对时间戳与相应 500 ms 窗口重叠的所有句子的嵌入取平均值，计算 2 Hz 的文本特征。由于 Hume-Vidmimic 中的视频通常只有几秒长，因此我们不对其转录文本进行时间对齐。Ulm-TSST 中的生物信号最初以 1 KHz 的频率采样。此外，我们将其降采样至 2 Hz，并随后通过萨维茨基–戈莱滤波器进行平滑处理。我们同时提供 1 kHz 和降采样后的生物信号。由于 Ulm-TSST 也以 2 Hz 的频率标注，因此为 MuSe-Personalisation 提供的特征已与标注信号对齐。由于 Passau-SFCH 中的标签对应于大小为 2 s 的窗口，因此 MuSe-Humour 的特征并未与标注直接对齐，但可以轻松匹配。在 MuSe-Mimic 中，标签对应于完整视频，因此无需与标签进行对齐。

3.6 基线训练

鉴于所有任务都基于视频录制，因此具有序列特性，我们选用 GRU-RNN，后接两个前馈层，作为一种有竞争力且简单的基线系统。我们针对每个子挑战和特征，对该模型进行超参数搜索。具体而言，我们优化 GRU 隐藏表示的维度、堆叠的 GRU 层数以及学习率。此外，我们还考察单向和双向 GRU。训练完所有单模态模型后，我们还尝试采用简单的后期融合方法来构建多模态基线。具体来说，我们根据模型在开发集上的表现，为表现最佳的单模态模型的预测结果加权并求平均。我们在 GitHub[12] 上公开提供代码、超参数和最佳模型检查点。接下来，我们将介绍各任务特定的训练细节。

3.6.1 MuSe-Mimic. 为了预测 Hume-Vidmimic 中三个片段级情绪标注的分数，我们将视频的特征输入 GRU模型，并取其最终隐藏表示来对视频进行编码。模型使用均方误差（MSE）损失进行训练。由于 Hume-Vidmimic 的数据规模较大，我们在包含 5 000 个数据点的子集上进行超参数搜索。

[12]https://github.com/EIHW/MuSe-2023

3.6.2 MuSe-Humour. 由于 Passau-SFCH 的标签对应于 2 秒的帧，而特征每 500 毫秒采样一次，因此每个训练数据点对应于最多 4 个步骤。后期融合基于每个 2 秒帧的预测结果。对于每种超参数配置，训练流程使用 5 个不同且固定的随机种子运行。由于预测目标为二元目标，我们采用二元交叉熵损失函数。

3.6.3 MuSe-Personalisation. 为了充分利用完整的 Ulm-TSST 数据集和特定于测试对象的数据，我们在 MuSe-Personalisation 基线中采用两阶段方法。按照 [34]，，我们首先在整个数据集上训练模型。具体而言，我们使用 3 个固定随机种子进行超参数搜索。随后，根据模型在开发集上的表现，选出最佳的单个模型检查点。第二步中，我们为每个测试对象复制该模型，并仅使用相应对象的数据进一步训练，采用 5 个固定随机种子。为此，我们另外进行超参数搜索，以优化训练轮数和学习率。最终预测结果通过以下方式获得：针对每个测试对象，从 5 个特定于该对象的模型中选出表现最佳者，并用其预测该对象对应的测试标签。

然而，如果针对特定受试者的最佳模型在该受试者的开发集数据上的表现不如初始模型，我们就改用初始模型对该受试者的预测结果。在两个阶段中，训练数据都会进行分段。分段的窗口大小和步长通过超参数搜索进行优化。与以往的 MuSe挑战赛类似 [19, 54, 55]，，我们采用CCC损失作为损失函数。我们的实验仅限于基于音频和视频的特征，并将生理信号和文本特征的实验留给参赛者。

4 基线结果

我们按照上一节所述训练门控循环单元。接下来，我们将呈现并讨论基线结果。

4.1 MuSe-Mimic

MuSe-Mimic的结果见表3。

实验结果表明，在所有单模态方法中，Wav2Vec2.0特征表现最佳，在开发集和测试集上的皮尔逊均值分别为.4317和.4296。ELECTRA特征的表现与之相近，在开发集和测试集上的皮尔逊均值分别为.4079和.3855。相比之下，其他五种基于音频和视频的特征表现明显较差，其中FaceNet512在测试集上的结果（0.0292）表明相关性极弱。此外，所有面部相关特征的皮尔逊值都相对较低，其中面部动作单元表现最佳，在测试集上的数值为.1337。就音频模态而言，Wav2Vec2.0特征明显优于eGeMAPS和DeepSpectrum，远远超过二者的相关性。这可能是因为Wav2Vec2.0模型在自动语音识别（ASR）任务上进行预训练，因而融入了语言知识 [60]。此外，纯文本特征（ELECTRA）取得的皮尔逊值相对较高，这表明文本模态编码了该任务所需的重要信息。因此，模型很可能主要利用了Wav2Vec2.0嵌入中包含的语言信息，

表2：MuSe-Mimic 的结果。我们报告 3 个情绪目标 认可、失望、不确定性 **均值的皮尔逊相关系数（ρ）。报告的分数为 5 个固定随机种子中的最佳结果，以及这些种子的均值 ρ 和相应的标准差。**

	[均值 ρ]	
特征	开发集	Test
音频		
eGeMAPS	.0842 (.0739 ± .0067)	.0546 (.0462 ±.0070)
DeepSpectrum	.0800 (.0748 ± .0043)	.0708 (.0734 ±.0072)
Wav2Vec2.0	.4317 (.4290 ± .0020)	.4296 (.4330) ±.0029)
视频		
FAU	.1280 (.1241 ± .0032)	.1337 (.1319 ±.0019)
ViT	.1202 (.1151 ± .0041)	.1068 (.1046 ±.0098)
FaceNet512	.0669 (.0540 ± .0072)	.0292 (.0275 ±.0160)
Text		
ELECTRA	.4079 (.4027 ± .0028)	.3855 (.3902 ±.0037)
后期融合		
A+T	.4718 (.4695 ± .0022)	.4679 (.4657 ±.0025)
A+V	.4234 (.4131 ± .0094)	.4281 (.4209 ±.0079)
T+V	.4027 (.3983 ± .0049)	.3965 (.3869 ±.0075)
A+T+V	.4789 (.4761 ± .0024)	**.4727** (.4711 ± .0023)

表3：各类别按模态划分的最佳特征及其后期融合结果的皮尔逊相关值（ρ）。报告的是 5 个固定随机种子的平均结果及相应的标准差。

	[按类别 ρ]		
特征	认可	失望	不确定性
Wav2Vec2.0（A）	.5139 ± .0032	.4813 ± .0023	.3045 ±.0036
ELECTRA (T)	.4590 ± .0055	.4227 ±.0054	.2889 ±.0034
FAU（V）	.1786 ± .0030	.1407 ±.0041	.0763 ±.0064
A+T+V	**.5536** ± .0029	。**5139** ± 。0032	**.3395** ± .0015

因此，其表现优于仅使用音频的特征 DeepSpectrum 和 eGeMAPS。

表3进一步揭示了最佳模型在各个情绪类别上的表现。分析表明，对于全部三个标签，Approval 的类别相关性均值最高，而 Uncertainty 始终是最具挑战性的情绪目标。例如，在 Wav2Vec2.0 中，Approval 的 ρ值为 .5139，而 Uncertainty 的对应值仅为 .3045。后期融合结果（*A+T+V*）也体现了这一趋势，其表现优于所有单模态类别结果；Approval、Disappointment 和 Uncertainty 的相关性均值分别为 .5536、.5139 和 .3395。

此外，如表2和表3所示，多模态后期融合方法能够提升该性能。值得注意的是，

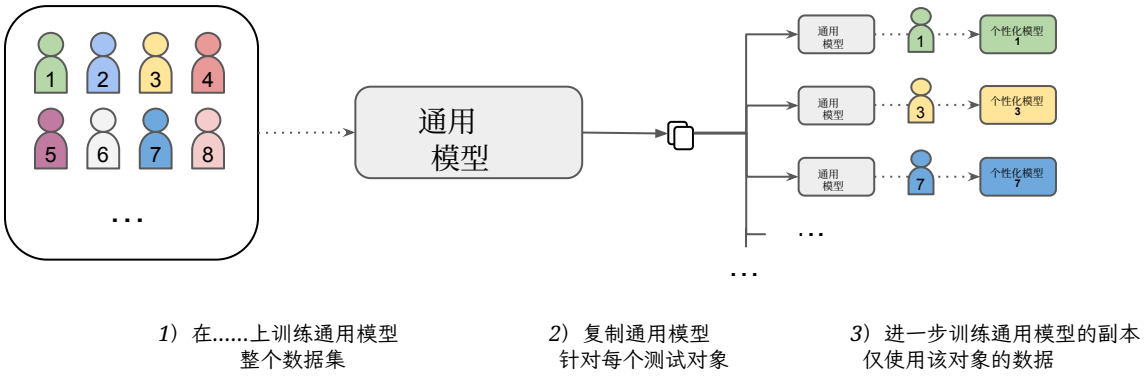

图 1：MuSe-Personalisation 中采用的训练方法。请注意，图中的 ID 与 Ulm-TSST数据集中的实际 ID 不对应。

音频与文本的结合取得了令人鼓舞的结果，相较于 Wav2Vec2.0 的结果，在开发集和测试集上的表现均提升了约.04。最后，将每种模态中采用最佳特征获得的预测结果进行融合，即 Wav2Vec2.0、面部动作单元和 ELECTRA，得到的 Pearson 相关系数在开发集和测试集上分别为.4789 和.4727。因此，尽管文本模态似乎在当前任务中发挥着关键作用，但预计多模态方法将优于仅使用文本的方法，这表明三种模态都包含能够为该任务作出贡献的有价值信息。

4.2 MuSe-Humour

表 4 报告了 MuSe-Humour 的结果。

所有特征的结果都高于随机水平，即 0.5 AUC。Wav2Vec2.0 在开发集和测试数据上的 AUC 分数分别为 .8435 和 .7940，整体表现最佳。与 MuSe-Mimic 的结果类似，一个可能的原因是，Wav2Vec2.0 在 ASR 任务上进行预训练，也编码了语言信息 [60]，因此不能被视为严格意义上的单模态特征提取器。它明显优于 eGeMAPS 和 DeepSpectrum，后两者按构造仅基于音频信号。eGeMAPS 和 DeepSpectrum 对测试数据的泛化能力良好：eGeMAPS 的平均 AUC 从 .6836 降至 .6554，而 DeepSpectrum 的平均 AUC 则从开发数据上的 .6936 略微提升至测试集上的 .7019。相比之下，视频模态存在明显的泛化问题，尤其是 FAU 和 FaceNet512 特征。使用 FAU 特征得到的平均 AUC 分数从开发数据上的 .7702 降至跨文化测试数据评估时的 .5948。仅考虑表现最佳的随机种子时，文本特征（.7572 AUC）在测试集上的表现优于除 Wav2Vec2.0 以外的所有其他特征。不过，其表现相当不稳定。在开发数据和测试数据上的标准差分别为 .0717 和 .0830，高于所有其他特征。

对于简单的后期融合方法，所有可能的模态组合均优于各自对应的单模态结果，这表明幽默表达具有多模态特性。

表 4：MuSe-Humour 基线结果。每行对应使用 5 个固定随机种子进行的实验，并报告其中最佳的 AUC 分数，以及这 5 个随机种子的平均 AUC 分数和标准差。

特征	开发集	<escape>[AUC]</escape> Test
音频		
eGeMAPS	.7235 (.6836 ± .0254)	.6672 (.6554 ±.0169)
DeepSpectrum	.6969 (.6936 ± .0022)	.7012（.7019 ±.0025）
Wav2Vec2.0	.8435 (.8332 ± .0082)	.7940 (.7929 ±.0113)
视频		
FAU	.7879 (.7702 ± .0087)	.6398 (.5948 ±.0546)
ViT	.8277 (.7890 ± .0257)	.7457 (.7478 ±.0093)
FaceNet512	.7342 (.6608 ± .0671)	.5350 (.5446 ±.0188)
Text		
BERT	.8105 (.7635 ± .0717)	.7572 (.7108 ±.0830)
后期融合		
A+T	.8791 (.8600 ± .0218)	.8218 (.8067 ±.0149)
A+V	.8656 (.8362 ± .0205)	.8222 (.8125 ±.0112)
T+V	.8428 (.8125 ± .0246)	.7907 (.7780 ±.0259)
A+T+V	.8759 (.8504 ± .0209)	**.8310**(.8244 ± .0168)

音频、文本和视频最佳模型的组合在整体测试集上取得了最高的 AUC 分数，即。8310。

4.3 MuSe-Personalisation

表 5 报告了 MuSe-Personalisation 的结果。在除使用 Wav2Vec2.0 进行效价预测之外的每项实验中，测试分区上的结果都显著低于开发集上的 CCC 值。对于两种模态和预测目标，至少存在一个特征集，能够在测试分区上取得大于 0.5 的 CCC 值。就视频模态而言，FaceNet512 明显优于另外两个特征集，二者均用于

唤醒度和效价方面，FaceNet512 在测试数据上的唤醒度 CCC 值达到 .5959，比排名第二的视频特征（面部动作单元，FAUs ）高出 0.200 以上。此外，FaceNet512 是效价表现最佳的特征，在开发集和测试数据上的表现均优于所有其他单模态结果。对于音频而言，并不存在如此明显的优势特征集。虽然 eGeMAPS 在开发集的唤醒度和效价两项任务中均取得了最佳结果，但在测试集上，其表现分别不及 DeepSpectrum（唤醒度）和 Wav2Vec2.0（效价）。总体而言，与基于面部的特征集相比，音频特征在唤醒度预测方面表现更好。具体来说，eGeMAPS 和 DeepSpectrum 在测试集唤醒度任务中的 CCC 值分别为 .7395 和 .7482，均优于视觉特征取得的最佳结果，即使用 FaceNet512 获得的 CCC 值 .5959。对于效价预测，则不存在这种趋势。在测试数据上，每种模态仅有一个特征集的 CCC 值略高于 0.5，分别是 Wav2Vec2.0，其 CCC 值为 .5232，以及 FaceNet512，其 CCC 值达到 .5654。

后期融合结果表明，音频和视频模态具有互补性。尤其是，对于效价，融合 eGeMAPS 和 FaceNet512 的表现显著优于它们各自的表现，在测试集上的 CCC 值为 .7827。对于唤醒度，这种效果没有那么明显：后期融合在测试集上的结果为 .7450，仅略高于仅使用 eGeMAPS 所取得的相应结果，即 .7395。总体而言，采用后期融合方法，在测试集上取得了最佳综合得分 .7639。某一特征在开发集与测试数据上的表现可能相差很大。通过更精细地利用所提供的数据，或许可以解决这一问题。具体而言，参与者并非必须将前 60 秒用作训练数据、将第二分钟用作开发数据。

表 5：MuSe-Personalisation 的结果。报告的是针对每位受试者进行专门训练后，效价和生理唤醒度的 CCC 值，即基于每位受试者五个随机种子中最佳的个性化模型。综合指相应特征的效价和唤醒度 CCC 值的平均值。

	唤醒度 [CCC]		效价 [CCC]		综合 [CCC]
特征	Dev.	Test	Dev.	Test	Test
音频					
eGeMAPS	.9073	.7395	.5892	.3944	.5670
DeepSpectrum	.8064	**.7482**	.3536	.2836	.5159
Wav2Vec2.0	.7421	.5325	.5142	.5232	.5279
视频					
FAU	.6382	.3766	.1468	.1076	.2421
ViT	.2691	.0001	.6050	.4490	.2246
FaceNet512	.8260	.5959	.6491	.5654	.5807
后期融合					
A + V	.9145	.7450	.8559	**.7827**	**.7639**

5 结论

我们介绍了 MuSe 2023——第 4 届多模态情感分析挑战赛。MuSe 2023 包含三个子挑战：

对于 MuSe-Mimic 子挑战赛，现已开放全新的 Hume-Vidmimic 数据集，其中包含人们模仿情绪视频的录制内容。参赛者需要预测每段视频中认可、失望和不确定性的程度。

MuSe-Humour 子挑战赛基于 Passau-SFCH 数据集 [18]，的扩展版本，旨在识别新闻发布会中的自发性幽默。该任务提出了一种跨文化环境：参赛者使用德语录音训练其模型，但必须预测英语数据中的幽默话语。

在 MuSe-Personalisation 子挑战赛中，我们鼓励参赛者开发方法，以针对特定个体调整模型。MuSe-Personalisation 使用了此前几届 MuSe [19, 54] 中已采用的 Ulm-TSST 数据集。虽然所有子挑战赛涉及的场景都相对简单，仅包含一个人，但我们相信，多样化的预测目标和模态也将推动共情对话系统和对话情感分析的发展。

我们使用公开可用的软件，预先计算了音频、视频和文本模态的大量特征。此外，基于这些特征，我们训练了一个简单的 GRU模型，在相应测试集上取得了以下官方基线结果：对于 MuSe-Mimic，通过对音频、视频和文本特征进行后期融合，获得了 ρ 值均值.4727。类似地，对于 MuSe-Humour，对三种模态进行后期融合，基线 AUC 值为.8310。对于 MuSe-Personalisation，DeepSpectrum 特征取得了最佳唤醒度 CCC 值.7482；而音频和视频特征的后期融合则在效价（CCC 值为.7827）以及唤醒度与效价的均值（CCC 值为.7639）上取得了最佳结果。

我们公开了代码、数据集和特征。相对简单的基线系统所取得的结果，初步揭示了不同模态和特征对于所提出的各项子挑战的适用性。我们预计，通过更先进的模型和方法，这些基线结果可以得到显著提升。MuSe 2023 旨在成为开发和评估这类新型多模态方法的通用平台。

未来的研究工作除了本挑战赛所针对的单纯性能优化之外，还需要应对更多挑战，其中包括一些亟待解决的重要问题，例如可信赖性、可解释性、公平性以及“绿色”高效处理等。

6 致谢

本项目获得了德国研究联合会（DFG）资助，资助协议编号为 461420398；同时还获得了 DFG Reinhart Koselleck 项目（编号：442218748，AUDI0NOMOUS）的资助。

参考文献

[1] Miray Akyunus、Tülin Gençöz 和 B Türküler Aka。2021。青年期不同阶段基本人格特质与人际问题的年龄和性别差异。《当代心理学》40（2021），2518–2527。[2] Shahin Amiriparian、Lukas Christ、Andreas König、Eva-Maria Meßner、Alan Cowen、Erik Cambria 和 Björn W Schuller。2022。MuSe 2022 挑战赛：多模态幽默、情绪反应与压力。载于《第*30*届 *ACM* 多媒体国际会议论文集》。计算机协会，纽约，美国，7389–7391。[3] Shahin Amiriparian、Nicholas Cummins、Sandra Ottl、Maurice Gerczuk 和 Björn Schuller。2017。使用基于图像的深度频谱特征进行情感分析。载于《与第*7*届两年一度的情感计算与智能交互会议（*ACII 2017*）联合举办的第*2*届野外自动情感分析国际研讨会论文集（*WASA 2017*）》。AAAC、IEEE，美国圣安东尼奥，26–29。[4] Shahin Amiriparian、Maurice Gerczuk、Sandra Ottl、Nicholas Cummins、Michael Freitag、Sergey Pugachevskiy 和 Björn Schuller。2017。使用基于图像的深度频谱特征进行鼾声分类。载于《*INTERSPEECH 2017*论文集：国际语音通信协会第*18*届年会》。ISCA、ISCA，瑞典斯德哥尔摩，3512–3516。[5] Shahin Amiriparian、Maurice Gerczuk、Lukas Stappen、Alice Baird、Lukas Koebe、Sandra Ottl 和 Björn Schuller。2020。面向卷积神经网络与循环神经网络音频识别的跨模态预训练及时空特征学习。《*EURASIP* 音频、语音与音乐处理期刊》2020，19（2020），1–11。[6] Shahin Amiriparian、Tobias Hübner、Vincent Karas、Maurice Gerczuk、Sandra Ottl 和 Björn W. Schuller。2022。DeepSpectrumLite：面向去中心化数据的嵌入式语音与音频处理高能效迁移学习框架。《人工智能前沿》5（2022），10页。https://doi.org/10 3389 2022 856232 。/frai. 。[7] Alexei Baevski、Yuhao Zhou、Abdelrahman Mohamed 和 Michael Auli。2020。wav2vec 2.0：语音表征自监督学习框架。《神经信息处理系统进展》33（2020），12449–12460。[8] Alice Baird、Shahin Amiriparian 和 Björn Schuller。2019。深度生成音频能否表达情绪？迈向个性化情绪音频生成方法。载于《*2019 IEEE* 第*21*届多媒体信号处理国际研讨会（*MMSP*）》。IEEE、IEEE，马来西亚吉隆坡，1–5。[9] Alice Baird、Lukas Stappen、Lukas Christ、Lea Schumann、Eva-Maria Meßner 和 Björn W Schuller。2021。压力诱发场景中唤醒度的生理适配金标准。载于《与第*29*届 *ACM* 多媒体国际会议（*ACMMM*）同期举办的第*2*届多模态情感分析挑战赛论文集》。ACM、计算机协会，中国 Changu，69–73。[10] Kim Binsted 等。1995。运用幽默让自然语言接口更加友好。载于《人工智能、人工生命与娱乐研讨会论文集，国际人工智能联合会议》。[11] Björn W. Schuller、Anton Batliner、Christian Bergler、Cecilia Mascolo、Jing Han、Iulia Lefter、Heysem Kaya、Shahin Amiriparian、Alice Baird、Lukas Stappen、Sandra Ottl、Maurice Gerczuk、Panaguiotis Tzirakis、Chloë Brown、Jagmohan Chauhan、Andreas Grammenos、Apinan Hasthanasombat、Dimitris Spathis、Tong Xia、Pietro Cicuta、Leon J. M. Rothkrantz、Joeri Zwerts、Jelle Treep 和 Casper Kaandorp。2021。INTERSPEECH 2021计算副语言学挑战赛：新冠肺炎咳嗽、新冠肺炎语音、恶化与灵长类动物。载于《*INTERSPEECH 2021*论文集：国际语音通信协会第*22*届年会》。ISCA、ISCA，捷克布尔诺，431–435。[12] Jeffrey A Brooks、Panagiotis Tzirakis、Alice Baird、Lauren Kim、Michael Opara、Xia Fang、Dacher Keltner、Maria Monroy、Rebecca Corona、Jacob Metrick 等。2023。深度学习揭示不同文化中发声爆发所表达的内容。《自然·人类行为》7，2（2023），240–250。[13] Arnie Cann、Amanda J Watson 和 Elisabeth A Bridgewater。2014。评估职场幽默：幽默氛围问卷。《幽默》27，2（2014），307–323。[14] Mathilde Caron、Hugo Touvron、Ishan Misra、Hervé Jégou、Julien Mairal、Piotr Bojanowski 和 Armand Joulin。2021。自监督视觉变换器中的新兴特性。载于《*IEEE/CVF* 计算机视觉国际会议论文集》。9650–9660。[15] Delphine Caruelle、Anders Gustafsson、Poja Shams 和 Line Lervik-Olsen。2019。使用皮肤电活动（EDA）测量理解消费者情绪：文献综述与行动倡议。《商业研究期刊》104（2019），146–160。[16] Aayushi Chaudhari、Chintan Bhatt、Achyut Krishna 和 Pier Luigi Mazzeo。2022。ViTFER：使用视觉 Transformer 进行面部情绪识别。《应用系统创新》5，4（2022），80。[17] Chengxin Chen 和 Pengyuan Zhang。2022。通过潜在表征偏移整合跨模态交互，实现多模态幽默检测。载于《第*3*届多模态情感分析研讨会与挑战赛论文集》（葡萄牙里斯本）（*MuSe' 22*）。计算机协会，纽约，美国，23–28。https://doi.org/10.1145/3551876.3554805

[18] Lukas Christ、Shahin Amiriparian、Alexander Kathan、Niklas Müller、Andreas König 和 Björn W Schuller。2022。《多模态自发幽默预测：一个新数据集及初步结果》。arXiv预印本 arXiv:2209.14272（2022）。[19] Christ, Lukas 和 Amiriparian, Shahin 和 Baird, Alice 和 Tzirakis, Panagio- tis 和 Kathan, Alexander 和 Müller, Niklas 和 Stappen, Lukas 和 Meßner, Eva-Maria 和 König, Andreas 和 Cowen, Alan 和 Cambria, Erik 和 Schuller, Björn W。2022。《MuSe 2022 多模态情感分析挑战赛：幽默、情绪反应与压力》。载于第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）。计算机协会，美国纽约，5–14。[20] Kevin Clark、Minh-Thang Luong、Quoc V. Le 和 Christopher D. Manning。2020。《ELECTRA：将文本编码器预训练为判别器而非生成器》。载于 ICLR。https://openreview.net/pdf?id=r1xMH1BtvB [21] Jacob Devlin、Ming-Wei Chang、Kenton Lee 和 Kristina Toutanova。2019。《BERT：用于语言理解的深度双向变换器预训练》。载于2019年计算语言学协会北美分会年会：人类语言技术论文集。4171–4186。[22] George Doddington、Walter Liggett、Alvin Martin、Mark Przybocki 和 Douglas Reynolds。1998。《羊、山羊、羔羊与狼：NIST 1998年说话人识别评测中的说话人表现统计分析》。技术报告。美国国家标准与技术研究院，马里兰州盖瑟斯堡。[23] Paul Ekman 和 Wallace V Friesen。1978。《面部动作编码系统》。环境心理学与非语言行为（1978）。[24] Florian Eyben、Klaus R Scherer、Björn W Schuller、Johan Sundberg、Elisabeth André、Carlos Busso、Laurence Y Devillers、Julien Epps、Petri Laukka、Shrikanth S Narayanan 等。2015。《用于语音研究与情感计算的日内瓦极简声学参数集（GeMAPS）》。《IEEE情感计算汇刊》7，2（2015），190–202。[25] Florian Eyben、Martin Wöllmer 和 Björn Schuller。2010。《openSMILE：慕尼黑通用快速开源音频特征提取器》。载于第18届 ACM 国际多媒体会议论文集。计算机协会，意大利佛罗伦萨，1459–1462。[26] Maurice Gerczuk、Shahin Amiriparian、Sandra Ottl 和 Björn Schuller。2022。《EmoNet：多语料库语音情绪识别迁移学习框架》。《IEEE情感计算汇刊》13（2022）。[27] Panagiotis Gkorezis、Eugenia Petridou 和 Panteleimon Xanthiakos。2014。《领导者积极幽默与组织犬儒主义：领导—成员交换的中介作用》。《领导力与组织发展期刊》35（2014），305–315。[28] Michael Grimm 和 Kristian Kroschel。2005。《使用自我评估人形图评估自然情绪》。载于2005年 IEEE自动语音识别与理解研讨会。IEEE，IEEE，墨西哥坎昆，381–385。[29] Oliver Guhr、Anne-Kathrin Schumann、Frank Bahrmann 和 Hans Joachim Böhme。2021。《FullStop：用于标点预测的多语言深度模型》。载于2021年瑞士文本分析会议论文集。CEUR研讨会论文集，瑞士温特图尔。http://ceur-ws.org/Vol-2957/sepp_paper4. pdf [30] Md Kamrul Hasan、Wasifur Rahman、AmirAli Bagher Zadeh、Jianyuan Zhong、Md Iftekhar Tanveer、Louis-Philippe Morency 和 Mohammed (Ehsan) Hoque。2019。《UR-FUNNY：用于幽默理解的多模态语言数据集》。载于2019年自然语言处理实证方法会议与第9届国际自然语言处理联合会议论文集（EMNLP-IJCNLP）。计算语言学协会，中国香港，2046 2056 10 18653 1 19 1211 – 。https://doi.org/ . /v /D - Yu He、Licai Sun、Zheng Lian、Bin Liu、Jianhua Tao、Meng Wang 和 Yuan Cheng。2022。《情感分析中的多模态时间注意力》。载于第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）（MuSe' 22）。计算机协会，美国纽约，61 66 10 1145 3551876 3554811 – 。[31] 2013 https://doi.org/ . / . Ursula Hess 和 Agneta Fischer。.《情绪模仿作为社会调节》。《人格与社会心理学评论》17，2（2013），142–157。[32] Ursula Hess 和 Agneta Fischer。2014。《情绪模仿：我们为何以及何时模仿情绪》。《社会与人格心理学指南》8，2（2014），45–57。[33] Alexander Kathan、Shahin Amiriparian、Lukas Christ、Andreas Triantafyllopou- los、Niklas Müller、Andreas König 和 Björn W. Schuller。2022。《视听幽默识别的个性化方法及其个体层面公平性》。载于第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）（MuSe' 22）。计算机协会，美国纽约，29 36 10 1145 3551876 3554800 – 。[34] 1993 https://doi.org/ . / . Clemens Kirschbaum、Karl-Martin Pirke 和 Dirk H Hellhammer。《“特里尔社会压力测试”——一种用于研究实验室环境中心理生物学压力反应的工具》。《神经心理生物学》28，1-2（1993），76–81。[35] Anna Ladilova 和 Ulrike Schröder。2022。《跨文化互动中的幽默：误解之源还是构建共同基础的方式？多模态分析》。《跨文化语用学》19，1（2022），71–101。[36] Jialin Li、Alia Waleed 和 Hanan Salam。2023。《人机交互中的个性化情感计算综述》。arXiv预印本 arXiv:2304.00377（2023）。[37]

[38] Jia Li、Ziyang Zhang、Junjie Lang、Yueqi Jiang、Liuwei An、Peng Zou、Yangyang Xu、Sheng Gao、Jie Lin、Chunxiao Fan、Xiao Sun 和 Meng Wang。2022。《用于情感分析的混合多模态特征提取、挖掘与融合》。载于*第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，美国纽约，81 88 10 1145 3551876 3554809 – 。https://doi.org/ . / .

[39] Shuo Liu 、 Adria Mallol-Ragolta 、 Emilia Parada-Cabaleiro 、 Kun Qian 、 Xin Jing 、 Alexander Kathan、Bin Hu 和 Björn W. Schuller。2022。《音频自监督学习：综述》。*Patterns* 3，12（2022），100616。https://doi.org/10.1016/j.patter. 2022.100616

[40] Yiping Liu、Wei Sun、Xing Zhang 和 Yebao Qin。2022。《通过逐特征融合改进维度情绪识别》。载于*第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，美国纽约，55 60 10 1145 3551876 3554804 – 。

[41] https://doi.org/ . / . R. Lotfian 和 C. Busso。2019。《通过从现有播客录音中提取情绪语音，构建自然且情绪平衡的语音语料库》。《IEEE情感计算汇刊》10，4（2019年10月至12月），471–483。https://doi.org/ 10 1109 2017 2736999 . /TAFFC. . Rod A

[42] Martin、Patricia Puhlik-Doris、Gwen Larsen、Jeanette Gray 和 Kelly Weir。2003。《幽默使用方式的个体差异及其与心理健康的关系：幽默风格问卷的开发》。《人格研究杂志》37，1（2003），48–75。[43] Michael McAuliffe、Michaela Socolof、Sarah Mihuc、Michael Wagner 和 Morgan Sonderegger。2017。《蒙特利尔强制对齐器：使用 Kaldi 进行可训练的文本—语音对齐》。载于*INTERSPEECH*论文集，第2017卷。国际语音通信协会（ISCA），瑞典斯德哥尔摩，498–502。[44] Anirudh Mittal、Pranav Jeevan、Prerak Gandhi、Diptesh Kanojia 和 Pushpak Bhattacharyya。2021。《"你觉得自己很幽默？"：为单口喜剧的幽默指数评分》。arXiv：arXiv预印本 arXiv:2110.12765

[45] Edmilson Morais、Ron Hoory、Weizhong Zhu、Itai Gat、Matheus Damasceno 和 Hagai Aronowitz。2022。《使用自监督特征进行语音情绪识别》。载于*ICASSP 2022——2022*年 *IEEE* 声学、语音与信号处理国际会议（*ICASSP*）。IEEE，6922–6926。[46] Sandra Ottl、Shahin Amiriparian、Maurice Gerczuk、Vincent Karas 和 Björn Schuller。2020。《利用深度频谱特征进行群体层面的语音情绪识别》。载于*第8届 ICMI 2020 EmotiW*——情境中的情绪识别挑战赛（*EmotiW 2020*），第*22*届 *ACM* 国际多模态交互会议（*ICMI 2020*）论文集。ACM，ACM，荷兰乌得勒支，821–826。[47] Ho-min Park、Ilho Yun、Ajit Kumar、Ankit Kumar Singh、Bong Jun Choi、Dhanan- jay Singh 和 Wesley De Neve。2022。《利用姿态特征工程和 Transformer 编码器实现时间连续情绪的多模态预测》。载于*第3届国际多模态情感分析研讨会与挑战赛论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，美国纽约，47 54 10 1145 3551876 3554807 – 。

[48] https://doi.org/ . / . Leonardo Pepino、Pablo Riera 和 Luciana Ferrer。2021。《使用 wav2vec 2.0 嵌入进行语音情绪识别》。载于*INTERSPEECH 2021*论文集。ISCA，ISCA，捷克布尔诺，3400 3404 10 21437 2021 703 – 。

[49] https://doi.org/ . /Interspeech. - Telmo Pires、Eva Schlinger 和 Dan Garrette。2019。《多语言 BERT 有多"多语言"？》。载于计算语言学协会第*57*届年会论文集。计算语言学协会，意大利佛罗伦萨，4996 5001 10 18653 1 19 1493 – 。https://doi.org/ . /v /

[50] P - Béatrice Priego-Valverde、Brigitte Bigi、Salvatore Attardo、Lucy Pickering 和 Elisa Gironzetti。2018。《幽默时微笑真的如此显而易见吗？美国英语和法国互动中幽默片段微笑行为的跨文化比较》。《跨文化语用学》15，4（2018），563–591。[51] Nipun Sadvilkar 和 Mark Neumann。2020。《PySBD：实用的句子边界消歧》。载于第二届自然语言处理开源软件研讨会（*NLP-OSS*）论文集。计算语言学协会，线上，110–114。https: //www.aclweb.org/anthology/ 2020 1 15 .

[52] nlposs- 。Florian Schroff、Dmitry Kalenichenko 和 James Philbin。2015。《FaceNet：用于人脸识别与聚类的统一嵌入》。载于*2015*年 *IEEE* 计算机视觉与模式识别会议（*CVPR*）。IEEE。https://doi.org/10. 1109 2015 7298682 /cvpr. . Sefik Ilkin

[53] Serengil 和 Alper Ozpinar。2020。《LightFace：混合式深度人脸识别框架》。载于*2020*年智能系统与应用创新会议（*ASYU*）。IEEE，23–27。https://doi.org/10.1109/ASYU50717.2020。9259802

[54] Lukas Stappen、Alice Baird、Lukas Christ、Lea Schumann、Benjamin Sertolli、Eva-Maria Messner、Erik Cambria、Guoying Zhao 和 Björn W Schuller。2021年。MuSe 2021多模态情感分析挑战赛：情感、情绪、生理情绪和压力。载于*第2届多模态情感分析挑战赛论文集*。计算机协会，纽约，美国，5–14。[55] Lukas Stappen、Alice Baird、Georgios Rizos、Panagiotis Tzirakis、Xinchen Du、Felix Hafner、Lea Schumann、Adria Mallol-Ragolta、Bjoern W. Schuller、Iulia Lefter、Erik Cambria 和 Ioannis Kompatsiaris。2020年。MuSe 2020挑战赛与研讨会：真实生活媒体中的多模态情感分析、情绪目标参与度和可信度检测。载于*第1届真实生活媒体多模态情感分析国际挑战赛与研讨会论文集*。ACM，计算机协会，纽约，美国，35–44。[56] Lukas Stappen、Eva-Maria Meßner、Erik Cambria、Guoying Zhao 和 Björn W. Schuller。2021年。MuSe 2021挑战赛：多模态情绪、情感、生理情绪和压力检测。载于*第29届 ACM 多媒体国际会议（ACMMM）*（虚拟活动，中国）。ACM，计算机协会，纽约，美国。[57] Lukas Stappen、Lea Schumann、Benjamin Sertolli、Alice Baird、Benjamin Weigel、Erik Cambria 和 Björn W Schuller。2021年。MuSe工具箱：多模态情感分析连续标注融合与离散类别转换工具箱。载于与第*29届 ACM 多媒体国际会议（ACMMM）*同期举办的第*2届多模态情感分析挑战赛论文集*。ACM，计算机协会，中国 Changu，75–82。[58] Lorenzo Vaiani、Moreno La Quatra、Luca Cagliero 和 Paolo Garza。2022。ViPER：用于情绪识别的基于视频的感知器。载于*第3届多模态情感分析研讨会与挑战赛国际论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，纽约，美国，67 73 10 1145 3551876 3554806 –。https://doi.org/。/。

[59] Bogdan Vlasenko、RaviShankar Prasad 和 Mathew Magimai.-Doss。2021年。使用监督自编码器融合声学与语言信息以改进情绪识别。载于*第2届多模态情感分析挑战赛论文集*。计算机协会，纽约，美国，51–59。[60] J. Wagner、A. Triantafyllopoulos、H. Wierstorf、M. Schmitt、F. Burkhardt、F. Eyben 和 B. W. Schuller。2023。语音情绪识别中 Transformer 时代的黎明：弥合效价差距。*IEEE*模式分析与机器智能汇刊 01 2023 1 13 10 1109 2023 3263585 （），–。

[61] https://doi.org/。/TPAMI.。Yanna J Weisberg、Colin G DeYoung 和 Jacob B Hirsh。。大五人格十个方面的人格性别差异。《心理学前沿》2（2011年），178。[62] Jiaming Wu、Hongfei Lin、Liang Yang 和 Bo Xu。2021年。MUMOR：用于对话幽默检测的多模态数据集。载于*CCF*自然语言处理与中文计算国际会议。施普林格，施普林格，中国青岛，619–627。[63] Haojie Xu、Weifeng Liu、Jiangwei Liu、Mingzheng Li、Yu Feng、Yasi Peng、Yunwei Shi、Xiao Sun 和 Meng Wang。2022。用于幽默检测的混合多模态融合。载于*第3届多模态情感分析研讨会与挑战赛国际论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，纽约，美国，15–21。https://doi.org/10.1145/3551876.3554802

[64] Sarthak Yadav、Tilak Purohit、Zohreh Mostaani、Bogdan Vlasenko 和 Mathew Magimai.-Doss。2022。连续情绪识别中生物信号与声学特征表示的比较。载于*第3届多模态情感分析研讨会与挑战赛国际论文集（葡萄牙里斯本）（MuSe' 22*）。计算机协会，纽约，美国，37–45。https://doi.org/ 10 1145 3551876 3554812 。/。Kaipeng Zhang、Zhanpeng

[65] Zhang、Zhifeng Li 和 Yu Qiao。2016年。使用多任务级联卷积网络联合进行人脸检测与对齐。*IEEE*信号处理快报23（2016年04月）。[66] Ruicong Zhi、Mengyi Liu 和 Dezheng Zhang。2020年。自动面部动作单元分析综合综述。《视觉计算机》36（2020年），1067–1093。[67] Feng Zhou 和 Fernando De la Torre。2015年。广义典范时间规整。*IEEE*模式分析与机器智能汇刊38，2（2015年），279–294。