arXiv:2304.00860v1 [cs.SD] 3 Apr 2023

语音情感识别系统的设计与评估：基于 IEMOCAP 的现实检验案例研究

Nikolaos Antoniou [†, ∗], Ath anasios Katsamanis [†, ∗], Theodoros Giannakopoulos [†], Shrikanth Narayanan [†, +]

[†] Behavioral Signal Technologies，洛杉矶，加利福尼亚州，美国

∗ 语言与语音处理研究所，雅典娜研究中心，雅典，希腊+ 南加州大学语音分析与解读实验室（SAIL），洛杉矶，加利福尼亚州，美国

{nikos.antoniou,nassos,thodoris,shri}@behavioralsignals.com

摘要

迫切需要制定指南并建立标准测试集，以便对语音情感识别（SER）进行直接、公平的比较。尽管交互式情感双人动作捕捉（IEMOCAP）数据库等资源已成为研究人员开发和测试 SER 模型时广泛采用的参考语料库，但已发表的研究显示，其使用方式存在多种假设和诸多差异，对可复现性和泛化构成了挑战。基于对采用 IEMOCAP 作为案例的 SER 最新进展进行的批判性评述，我们的工作旨在作出两项贡献：首先，通过分析近期文献，包括其中所作的假设和采用的指标，我们提出一套 SER 评估指南。其次，借助近期发表且提供开源实现的研究，我们重点开展 SER 的可复现性评估。

索引词——语音情感识别、情感评估、可复现性、IEMOCAP

1. 引言

仅通过处理语音信息来识别情感的系统，对于构建能够整合所感知情感表达的智能机器尤其有用。例如，推断所表达的人类情感状态可以帮助语音助手调整对用户的回应。语音情感识别（SER）属于计算副语言学中的一大类问题 [1]。

与此类语音机器学习应用的常见情况一样，进步的基石在于使用标注数据集。这些资源使我们能够构建系统，并评估系统将输入语音样本成功映射到预期输出的性能。为语音情感识别任务收集合适的输入—标签配对，是一个高度模糊的过程。模糊性在于，标签质量取决于标注人员对情感的细腻感知。此外，还存在一个固有的权衡：收集情感语音时，数据质量与便于标注的需求之间存在取舍。例如，自发语音片段的（音频）质量可能很高，但其中可能包含重叠或含义模糊的情感表达，这会使标注人员困惑，并加剧此类数据的标注难度 [2]。另一方面，强行规定某个语音片段与某种情感一一对应，可能会导致用于学习和推理的数据样本过于人工化。

语音情感识别领域已有多项研究使用交互式情感双人动作捕捉（IEMOCAP） [3] 数据库。本文以近期使用 IEMOCAP 进行语音情感识别的研究为案例，通过批判性回顾，探讨各研究所采用的评估方法如何相互比较，以及如何制定指南，以实现可比较的评估和可复现的研究结果。我们的贡献概述如下：

- 通过批判性回顾以往研究，我们发现所采用的评估方法存在三项方法论局限。分析表明，评估协议缺乏共识，导致研究结果的取值范围较窄，无法在该领域的相关研究中得到恰当的背景化解读与推广。基于这些发现，我们总结出一套建议性的评估指南（借鉴了此前多项研究的实践），旨在尽可能减少语音情感识别整体流程中潜在的盲点。

- 我们还报告了一项可复现性研究，评估提供开源代码实现的模型是否能够得到与已发表研究结果一致的结果。

2. 相关工作

深度学习方法引发了范式转变，包括语音情感识别（SER）的研究与发展。用于解决语音情感识别问题的系统已基本完成转型，

从基于手工设计特征应用传统机器学习模型（如隐马尔可夫模型或支持向量机），转向端到端的深度神经网络（DNN）学习。

在设计深度学习语音情感识别系统时，实践者需要面对一系列重要选择。首先是如何表征语音信息，例如使用梅尔频率倒谱系数、频谱图或其他声学特征。随后，确定一种能够高效编码这些信息的合适架构至关重要，最自然的候选方案包括卷积神经网络[4, 5, 6], 长短期记忆网络[7, 8], , 甚至是 [9, 10] 二者的组合。

与此相关的是，近年来人们对自监督学习（SSL）特征的兴趣日益浓厚。大量研究正在探索使用 wav2vec [11, 12] 或 HuBERT [13] 特征进行情感识别的可能性 [14, 15, 16, 17,18]。类似地，应用于副语言学的 Conformer（Conformer Applied to Paralinguistics, CAP） [19] 是一种基于 Conformer[20] 架构的表征，其训练方式与 wav2vec 2.0（w2v2）模型类似，并在非语义语音基准（NOSS） [21]上评估其下游性能。在[22], 中，作者将高效的 CAP 表征蒸馏到更轻量的架构中，旨在显著降低内存与计算开销，同时仅以略微牺牲下游性能为代价。

另一方面，将我们的工作与先前的出版物联系起来至关重要，例如 Musgrave 等人 [23], 的研究。这些研究旨在提高人们对机器学习论文中方法学缺陷的认识，有助于提升该领域的实验严谨性，并在认识各方法报告结果可能存在的局限性方面发挥关键作用。

3. 重新审视 IEMOCAP

3.1. 数据集详情

IEMOCAP [3] 数据库由五组双人互动环节组成，每组环节均由一名男性说话人和一名女性说话人参与，共有 10 名说话人。每组环节中的对话按说话人轮次分段，这些对话片段均标注了感知与表达的情绪。每个片段由 3 名不同的标注人员进行标注，他们既赋予片段一个离散的类别标签（例如快乐、中性、悲伤等），也赋予一个连续值标签，用于评估效价、支配度和激活度维度。此外，这些对话融合了剧本式和即兴语音互动。

尽管该数据库中的标签信息多样，使得我们能够从多个方面研究语音情感识别（SER），但本文聚焦于文献中最常见的设置。在这种设置下，语音情感识别被视为四分类问题，所有离散标签不属于以下集合{中性、快乐、悲伤、愤怒、兴奋}的对话片段均被剔除。随后，考虑到快乐与兴奋这两个标签类别在表达上的相似性，将二者合并。经过这一处理，得到一个包含 $N = 5531$ 个样本的数据集，其中 {1708, 1636, 1103, 1084}个样本分别属于 {中性、快乐、悲伤、愤怒} 类别。经过筛选后，数据集中的音频片段平均时长为 ~ 4.5 秒，共计7小时语音数据，来自10位不同的说话人。

由此得到的数据集面临两项主要挑战：首先，与语音情感识别中使用的其他数据集相比，它包含的语音数据量有限，例如，MSP-Podcast [24] 有 27 小时的语音。其次，经验标签分布不平衡，大多数样本来自中性类和快乐类。这两项挑战都会以各种方式影响机器学习方法；前者意味着模型容量增大时容易过拟合，因此应采用强正则化加以应对，例如多任务学习 [9, 18, 15]。若未能妥善解决后一个问题，系统输出很可能会产生非预期偏差。近期研究提出了复杂的数据增强技术 [25, 26] ，或采用不同的损失函数，例如 Aftab 等人 [4], 为应对类别不平衡而使用的焦点损失。

3.2. 评估协议

本节重点介绍研究人员在评估语音情感识别模型时通常采用的三个主要假设。这导致评估协议种类繁多，最终结果难以与文献中以往方法的结果进行轻松、公平的比较。随后，我们提出一套评估指南，希望借此明确评估流程并减少其中的错误。

3.2.1. 说话人相关评估

在涉及多个说话人的语音相关应用中，必须确保测试数据中的说话人不出现在训练集中。否则，系统可能会看似表现优异，而实际上可能只是以最优方式学会利用虚假相关性。因此，评估过程应遵循说话人独立性（SI）这一属性。

与 IEMOCAP 相关的一项重要挑战是，它没有既定的名义测试集；因此，在每项研究中，研究人员都可以自行决定如何进行评估。在 IEMOCAP 上进行说话人无关评估有两种常见方式：(1) 五折交叉验证（或留一会话法），每轮使用四个会话（共八位说话人）作为训练数据，另一个会话（含两位说话人）作为测试数据；(2) 十折交叉验证（或留一说话人法），保留一位说话人用于测试，其余九位用于训练集。通常，在五折交叉验证中会预留两位说话人用于验证集（十折情况下则预留一位）。每折的准确率指标通过加权准确率（WA）和非加权准确率（UA）进行汇总。WA 衡量的是

#	出版物	Code	声学特征	非加权准确率 (%)	加权准确率 (%)	交叉验证	评估说明
1	Feng 等人 [8]	Yes	梅尔频率倒谱系数	76.4	75.5	五折	-
2	Zhu 和 Li [5]	Yes	梅尔频率倒谱系数	73.90	73.70	五折	兴奋视为快乐
3	Xu 等人 [27]	Yes	频谱图	67.94	67.28	五折	兴奋视为快乐
4	Gat 等人 [15]	No	HuBERT	-	81.0	五折	-
5	Wang 等人 [14]	Yes	HuBERT	-	79.58	五折	-
6			w2v2	-	77.47		-
7	Peng等人 [28]	No	梅尔频率倒谱系数	79.1	78.0	10 折	-
8	Xu 等人 [6]	Yes	频谱图	77.54	79.34	五折	仅即兴表演，除快乐外
9	Xu 等人 [27]	Yes	频谱图	76.36 (63.92)	76.18 (65.90)	五折	仅即兴（剧本化），除快乐外
10	Zhu 和 Li [5]	Yes	梅尔频率倒谱系数	79.25 (70.39)	81.18 (71.44)	五折	仅即兴（剧本化），除快乐外
11	Liu 和 Wang [29]	No	梅尔频率倒谱系数	78.30	79.52	五折	仅即兴表演
12	Moine等人 [30]	No	频谱图	77.22	-	五折	仅即兴表演

表1：采用说话人相关（SD）评估的方法比较。第8–12行仅对即兴数据进行随机5折交叉验证。

#	发表情况	Code	声学特征	非加权准确率 (%)	加权准确率 (%)	交叉验证
1	Pepino 等人 [16]	Yes	w2v2	67.2	-	五折
2	Yang 等人 [17]	Yes	HuBERT	67.62	-	五折
3	Gat 等人 [15]	No	HuBERT	-	74.2	五折
4	Santoso 等人 [7]	No	梅尔频率倒谱系数+常数 Q 变换+基频	75.9	76.1	五折
5	Li 等人 [9]	No	频谱图	82.8	81.6	五折
6	Zou 等人 [10]	Yes	梅尔频率倒谱系数,频谱, w2v2	71.05	69.80	五折
7				72.70	71.64	10 折
8	Wang 等人 [14]	Yes	HuBERT	-	73.01	10 折
9			w2v2	-	70.99	
10	Aftab 等人 [4]	Yes	梅尔频率倒谱系数	70.76	70.23	10 折
11	Feng 等人 [8]	Yes	梅尔频率倒谱系数	69.67	68.63	10 折
12	Cai et al. [18]	Yes	w2v2	-	78.15	10 折
13	Shor 等人 [19]	No	CAP	-	79.2	第 05 场次
14	Shor 和 Venugopalan [22]	是†	TRILLSSON	-	73.2	第 05 场次

表 2：采用说话人无关（SI）评估的方法比较。第 13、14 行的论文仅使用第 05 场次（即说话人“05M”和“05F”）作为测试集。（†）：这些论文仅提供了特征提取器训练后的权重。

预测正确的百分比，而 UA 则对每个类别的召回率指标取平均。

耐人寻味的是，社区至今仍未明确认识到进行说话人无关评估的必要性，因为许多近期研究会随机划分训练集、验证集和测试集，导致不同集合中的说话人身份有所重叠。与说话人无关评估相对，这种做法称为说话人相关（SD）评估。除了明显违背说话人独立性之外，随机划分还使比较范围仅限于各研究所采用的特定基线，并限制了泛化能力。

3.2.2. 对即兴互动数据的评估

在此前的分析中，我们提到 IEMOCAP 包含来自即兴互动和剧本式互动的语音。另一种常见的评估趋势是只评估这两个部分之一的性能。这种情况经常出现在即兴部分，例如 [30, 6]。虽然这种选择可能与每篇论文的独特目标有关，但需要指出的是，这种做法也可能再次限制我们将结果与其他已发表方法进行对比分析的能力。解决这一问题的理想做法是，对即兴数据和脚本化数据都进行评估并公开评估结果，同时分别报告各部分数据的结果。

3.2.3. “兴奋”类别案例

这里，我们讨论另一种广泛采用的选择。这种选择催生了一个新的评估分支，限制了广泛比较的开展。早期有一些研究 [31, 32] 只考虑 {中性、悲伤、愤怒和兴奋} 类，并在研究中排除快乐类（而不是将其与兴奋类合并）。因此，随后的一些论文 [27, 6, 5] 也按照这一协议进行评估。与前述两项评估挑战类似，我们认为研究人员应遵循纳入合并的快乐类和兴奋类这一惯例，并进一步与采用特定基线的方法进行比较，例如仅使用兴奋类的方法。

3.2.4. 文献结果比较

我们汇总了近期多种基于 IEMOCAP 数据库开发的语音情感识别方法所报告的结果。表 1 和表 2 分别展示了采用说话人相关（SD）和说话人无关（SI）评估的研究所报告的性能指标。我们还提供了其他信息，例如各项研究中语音信号的表示方式、相应论文是否提供官方实现，以及评估过程中采用的其他选择。在说话人相关评估中，由于数据是随机划分的，结果无法公平比较（除非某项研究在完全相同的划分方式下评估其他方法，例如 [5] 优于 [27, 6]）。总体而言，显而易见的是，偏离（或缺乏）统一的评估设置会导致情况不明、难以判断各方法在已发表文献中的表现，也会阻碍整个研究群体取得进展。

3.3. 推荐的评估指南

我们对未来使用 IEMOCAP 进行语音情感识别的研究建议是，按照以下协议开展评估，以确保建立一个通用、最低可行且可比较的基线：

- 使用中性、悲伤、愤怒、快乐和+兴奋类，构成包含 $N = 5531$ 个样本的数据集，
- 进行 10 折说话人无关交叉验证：以一位说话人为测试集、八位为训练集、一位为验证集，

在报告采用这一设置所得的结果后，研究人员可以自行调整，例如报告 IEMOCAP 即兴部分或其他标签子集/合并类别的结果。已有众多方法 [10, 14, 4, 18] 按照此设置进行了评估。关于交叉验证的折数，我们指出，五折和 10 折评估都是合理的选择；后者每折使用更多训练数据，因此 WA 和 UA 指标会略高。

4. 检验可复现性的实验

本节中，我们对已公开实现代码的论文开展可复现性研究。我们的目标是依据第 3.3 节的评估指南，评估表 1 和表 2 中每种方法（其实现代码均已公开）。

首先需要指出的是，尽管研究人员上传了代码，但大多数人都没有提供训练后的模型检查点。值得注意的是，在我们考察的研究中，只有两项 [18,22]分享了训练后的权重，这使复现过程容易得多。不过， [22] 的作者只分享了骨干网络的权重，因此需要提取训练集、验证集和测试集的嵌入，并在其上训练一个线性层，才能复现结果。Feng 等人 [8] 上传了一个代码库但该代码库在执行时出现错误，因此我们无法继续。Aftab 等人 [4] 已经按照推荐的指南进行了评估；我们确认，重新从头训练他们的模型后，其报告的结果可以复现。同样，尽管 Zou 等人 [10] 和 Wang 等人 [14] 没有分享检查点，但从头训练后也验证了他们报告的结果。由于 Yang 等人 [17] 和 Pepino 等人 [16], 的研究与 [14]几乎完全相同，因此我们没有检查其结果的可复现性。

接下来，我们关注表 1 中的研究。Zhu 和 Li [5] 提出了 GLAM架构，但他们的公开代码不包含检查点。从头训练，并按照其协议（即剔除快乐类的样本），得到的结果与论文报告的结果几乎完全一致。然而，当我们尝试纳入与之密切相关的快乐类数据重新训练模型时，性能显著下降。这表明，一个简单的假设（此处指排除快乐类示例）可能导致对模型有效性的不完整认识。GLAM架构的表现优于 [27, 6], 中的两个模型；这两个模型遵循完全相同的非常规协议。 [27, 6]提供的代码无法运行，也未得到维护。

表 3 总结了我们可复现性研究的发现，其中报告了按照 3.3 节所述协议测得的 WA指标。

#	发表情况	预训练	可复现性	WA（%）
1	Cai et al. [18]	✓	✓	78.15
2	Shor and Ven. [22]	✓	✓	68.05
3	Aftab 等人 [4]	✗	✓	71.43
4	Zou 等人 [10]	✗	✓	69.2
5	Wang 等人 [14]	✗	✓	69.61
6	Zhu 和 Li [5]	✗	✗	63.82
7	Feng 等人 [8]	✗	✗	-
8	Xu 等人 [27]	✗	✗	-
9	Xu 等人 [6]	✗	✗	-

表 3：可复现实验。预训练列表示这些实现是否共享训练后的权重。

5. 结论

本文通过对基于 IEMOCAP 的语音情感识别系统已发表研究进行实证案例回顾与分析，从便于比较性能、泛化能力和可复现性的角度，考察了语音情感识别系统设计与评估中的各种选择。我们鼓励未来的研究按照通用协议（例如根据执行良好的研究所提出的协议）来设计和评估方法。缺少这类协议会妨碍我们公平比较和复现不同方法的结果，也会阻碍整个领域的共同进步。我们希望，未来这项案例研究能够促成一套定义明确的“良好实践”，用于语音情感识别的数据与评估基准的设计、构建和发布。

6. 参考文献

[1] B. Schuller、S. Steidl、A. Batliner、F. Burkhardt、L. Devillers、C. Muller 和 S. S. Narayanan，《语音与语言副语言学：最新进展与挑战》，《计算机、语音与语言》，第27卷，第1期，第4–39页，2013年1月。[2]E. Mower、A. Metallinou、C.-C. Lee、A. Kazemzadeh、C. Busso、S. Lee 和 S. S. Narayanan，《对模糊情感表达的解读》，载于《情感计算与智能交互国际会议论文集》（ACII），阿姆斯特丹，荷兰，2009年9月。[3] C. Busso 等人，《IEMOCAP：交互式情感双人动作捕捉数据库》，《语言资源与评估》，第42卷，第4期，第335–359页，2008年。[4] A. Aftab 等人，《Light-sernet：用于语音情感识别的轻量级全卷积神经网络》，载于 ICASSP 2022，IEEE，2022年，第6912–6916页。[5] W. Zhu 和 X. Li，《基于多尺度特征表示的全局感知融合语音情感识别》，载于 ICASSP 2022，IEEE，2022年，第6437–6441页。[6] M. Xu、F. Zhang、X. Cui 和 W. Zhang，《结合多尺度区域注意力与数据增强的语音情感识别》，载于 ICASSP 2021，IEEE，2021年，第6319–6323页。[7] J. Santoso 等人，《基于注意力权重校正与词级置信度度量的语音情感识别》，载于 Interspeech 2021 论文集，2021年，第1947–1951页。[8] H. Feng、S. Ueno 和 T. Kawahara，《结合声学到词语自动语音识别模型的端到端语音情感识别》，载于 Interspeech 2020 论文集，2020年，第501–505页。[9] Y. Li、T. Zhao 和 T. Kawahara，《利用自注意力机制与多任务学习改进端到端语音情感识别》，载于 Interspeech 2019 论文集，2019年，第2803–2807页。[10] H. Zou、Y. Si、C. Chen、D. Rajan 和 E. S. Chng，《基于协同注意力的多层次声学信息语音情感识别》，载于 ICASSP 2022，IEEE，2022年，第7367–7371页。[11] S. Schneider 等人，《wav2vec：用于语音识别的无监督预训练》，载于 Interspeech 2019 论文集，2019年，第3465–3469页。[12] A. Baevski 等人，《wav2vec 2.0：语音表示自监督学习框架》，《神经信息处理系统进展》，第33卷，第12449–12460页，2020年。[13] W.-N. Hsu 等人，《HuBERT：通过掩码预测隐藏单元进行自监督语音表示学习》，《IEEE/ACM音频、语音与语言处理汇刊》，第29卷，第3451–3460页，2021年。[14] Y. Wang、A. Boumadane 和 A. Heba，《用于语音情感识别、说话人验证和口语理解的微调后 wav2vec 2.0/HuBERT 基准》，arXiv 预印本 arXiv:2111.02735，2021年。[15] I. Gat、H. Aronowitz、W. Zhu、E. Morais 和 R. Hoory，《自监督语音情感识别中的说话人归一化》，载于 ICASSP 2022，IEEE，2022年，第7342–7346页。

[16] L. Pepino、P. Riera 和 L. Ferrer，《使用 wav2vec 2.0 嵌入进行语音情感识别》，载于 Interspeech 2021 论文集，2021 年，第 3400–3404 页。[17] S. wen Yang 等人，《SUPERB：语音处理通用性能基准》，载于 Interspeech 2021 论文集，2021 年，第 1194–1198 页。[18] X. Cai、J. Yuan、R. Zheng、L. Huang 和 K. Church，《基于多任务学习的语音情感识别》，载于 Interspeech 2021 论文集，2021 年，第 4508–4512 页。[19] J. Shor、A. Jansen、W. Han、D. Park 和 Y. Zhang，《使用自监督Conformer进行通用韵律语音表征》，载于 ICASSP 2022。IEEE，2022 年，第 3169–3173 页。[20] A. Gulati 等人，《Conformer：用于语音识别的卷积增强 Transformer》，arXiv 预印本 arXiv:2005.08100，2020 年。[21] J. Shor 等人，《迈向通用非语义语音表征的学习》，载于 Interspeech 2020 论文集，2020 年，第 140–144 页。[22]J. Shor 和 S. Venugopalan，《Trillsson：蒸馏得到的通用韵律语音表征》，arXiv 预印本 arXiv:2203.00236，2022 年。[23] K. Musgrave、S. J. Belongie 和 S. Lim，《度量学习的现实检验》，载于 ECCV 2020。[24] R. Lotfian 和 C. Busso，《通过从现有播客录音中检索情感语音，构建自然情感平衡语音语料库》，《IEEE情感计算汇刊》，第 10 卷，第 4 期，第 471–483 页，2017 年。[25] A. Chatziagapi 等人，《使用生成对抗网络进行语音情感识别的数据增强》，载于 Interspeech 2019 论文集，2019 年，第 171–175 页。[26] R. Pappagari、J. Villalba、P. Zelasko、L. Moro-Velazquez 和 N. Dehak，《Copypaste：一种语音情感识别数据增强方法》，载于 ICASSP 2021。IEEE，2021 年，第 6324–6328 页。[27] M. Xu、F. Zhang 和 S. U. Khan，《通过注意力头融合提高语音情感识别准确率》，载于 2020 年第 10 届年度计算与通信研讨会及会议（CCWC），2020 年，第 1058–1064 页。[28] Z. Peng、Y. Lu、S. Pan 和 Y. Liu，《使用多尺度卷积神经网络和注意力机制实现高效语音情感识别》，载于 ICASSP 2021。IEEE，2021 年，第 3020–3024 页。[29] J. Liu 和 H. Wang，《一种提升混淆类别区分能力的语音情感识别框架》，载于 Interspeech 2021 论文集，2021 年，第 4483–4487 页。[30] C. L. Moine、N. Obin 和 A. Roebel，《说话人注意语音情感识别》，载于 Interspeech 2021 论文集，2021 年，第 2866–2870 页。[31] V. Chernykh 和 P. Prikhodko，《使用循环神经网络进行语音情感识别》，arXiv 预印本 arXiv:1701.08071，2017 年。[32] L. Tarantino、P. N. Garner 和 A. Lazaridis，《用于语音情感识别的自注意力》，载于 Interspeech 2019 论文集，2019 年，第 2578–2582 页。