arXiv:2301.00508v2 [cs.SD] 6 Apr 2023

EMOGATOR：采用基线机器学习分类方法的全新开源声音爆发数据集

Fred W.Buhl佛罗里达大学
fredbuhl@ufl.edu

2023年4月7日

摘要

声音爆发——笑声、哭声、叹息声、呻吟声和哼声等传达情绪的简短非言语发声，是语音情感识别中一个常被忽视的方面，却也是人类声音交流的重要组成部分。研究这些有趣发声的一大障碍是缺乏大型数据集。很高兴向大家介绍 EmoGator 数据集，其中包含来自 357 位说话者的 32,130 个样本，共计 16.9654 小时的音频；每个样本均由说话者归入 30 种不同的情绪类别之一。本文将讨论几种构建分类器以识别情绪类别的方法，并提出未来研究方向。数据集可从 https://github.com/fredbuhl/EmoGator 下载。

关键词 语音情绪识别；声音爆发；情感爆发；非言语发声；情感计算；机器学习；数据集

1 引言

情绪是人类经验的核心——它们激励并影响着我们的许多行为。识别他人的情绪一直是人们长期关注的领域。或许，最早的情绪识别科学研究是 Duchenne [1] 于1862年开展的工作；他收集了通过电刺激面部肌肉诱发的面部表情照片。

情绪有多少种，这个问题仍无定论。Duchenne通过面部表情识别出13种基本情绪和60种组合。Cowen和Keltner最近的一项研究发现，人类能够可靠地从面部表情中识别出28种不同情绪 [2]。同一团队的另一项近期研究 [3] 表明，人们自我报告的不同情绪多达27种；这些反馈来自受试者观看短视频片段后的反应。研究将情绪类别呈现为连续变化的梯度，它们有时会与其他情绪类别重叠；某个特定刺激物会在不同程度上引发多种情绪。

人类常通过改变语音的**韵律**来用声音表达情绪——即言语的音频特征。一项研究 [4]发现，人们能够通过言语韵律识别12种不同情绪，而且这一发现适用于两种文化；此前一项研究 [5] 发现，来自五个国家的受试者能够跨文化识别情绪，不过也观察到了群体内优势。

人类还会通过一种称为**声音爆发**的短促非语音声音来表达情绪，这种声音也称为“情感爆发”“情绪性发声”或“非言语发声”，例如笑声、哭声、叹息声、呻吟声和哼声——这些发声并非语音，从进化角度来看，很可能早于语音出现。研究 [6] 发现，人类能够从这些声音爆发中分辨出14种情绪状态。Cowen、Keltner等人的一篇近期论文 [7] 则表明，人们能够从这些短促发声中分辨出24种情绪状态。

通过人类发声检测和表达情绪的能力在人类发展早期便已出现 [8, 9, 10, 11, 12]。这对语言和社会发展至关重要；由于无法辨别他人情绪而遇到困难的人，包括因脑损伤或自闭症谱系障碍等状况导致此类困难的人，

在交流时会遇到困难。患有听觉情感失认症的人 [13] 无法辨别语音中的情感线索，但仍能理解词语；而患有韵律失调的人 [14] 说话单调，没有语调或情感色彩；帕金森病患者 [15]也可能出现这种情况。这些能力的任何受损都会严重影响与他人的交流和社会交往，凸显了唤起和理解情感表达的重要性。

1.1 问题所在

如今，通过“智能音箱”及其配套的虚拟助手（如 Siri、Alexa 和 Google Assistant）进行语音识别，与计算机交互已十分普遍。目前，这些系统都无法从语音音频信号中检测情绪；它们通过语音转文本深度学习模型将信号转换为文本（有时会产生滑稽的结果），却忽略了语音韵律中可能包含的情感内容。对于某些应用而言，一个词的说法可能与说了哪个词同样重要，甚至更加重要，因此这可能构成严重的局限。此外，由于声音爆发并非语音，这些系统会完全忽略它们。

能够从语音中进行情绪识别的计算机有许多应用，例如让电子游戏中的非玩家角色做出更逼真的回应。在幼儿教育中，了解年幼用户的情绪状态有助于判断其兴趣、受挫感或厌倦感；这类系统还可用于评估并提升儿童的情绪智力（或“情商”） [16]。检测情绪的能力也有助于发现孤独、焦躁或抑郁症的迹象 [17], ，这对居家养老的老年人等孤立人群尤其重要。社交机器人——旨在与人类密切互动的机器人——也能从情绪识别中受益 [18]；这类系统甚至可用于评估机器人对人类用户的吸引力 [19]。有人认为，在实现人类水平的语音情绪识别之前，我们永远无法声称语音识别达到了人类水平的性能，因为人类两者都能做到[20]。（需要指出的是，这一领域只是 Rosalind Picard 开创的情感计算这一更广阔领域的一个方面 [21], ；该领域不仅涉及情绪识别，还包括情感表达和情感感知决策。）

尽管当前商业产品存在局限，语音情感识别（SER）长期以来一直是计算机科学领域备受关注的研究方向 [22]。1996 年，Cowie 等人 [23] 开发了一种技术，可自动检测语音信号中的显著特征并收集摘要统计数据，再据此量化语音特征，将其归入四种情绪类别。多年来，语音情感识别领域采用了多种方法 [24]——梅尔频率倒谱系数（MFCC）、高斯混合模型（GMM）、支持向量机（SVM）、隐马尔可夫模型（HMM），以及长短期记忆网络（LSTM）等神经网络技术 [25] ；近来，深度学习神经网络也已得到应用。

本文所述研究借助新近收集的数据集，探讨了长期以来鲜少受到关注的声音爆发领域。研究将探索多种机器学习技术，比较其不同的性能表现，并提出未来研究方向。

这项工作的主要灵感源自 [7]；作者慷慨地向其他研究人员提供的声音爆发数据集，在发布时是规模最大的一套声音爆发数据集。该数据集包含 2,032 个声音爆发样本，涵盖 30 种情绪类别；如前所述，人类能够可靠地区分其中 24 种。当前研究的核心问题是：如果人类能够从声音爆发中区分 24 种情绪类别，机器是否也能做到?

尽管 Cowen 等人的数据集在当时规模最大，但它仍然相对较小，而且各类别的样本数量并不均衡；大多数机器学习方法都能从更多样本和均衡类别中获益良多。作者认为，需要收集一个规模更大的数据集，并评估几种不同的方法，以找到性能最佳的情绪分类器。

2 数据集，以及用于分类的深度学习与其他方法

2.1 数据集

EmoGator 数据集由 357 名说话者产生的 32,130 个声音爆发组成，共提供 16.9654 小时的音频；样本平均长度为 1.901 秒。每位说话者针对 30 个情绪类别中的每一类录制 3 个样本，共为每位说话者提供 90 个样本——这使每个类别的样本数量相等，并确保每位说话者在数据集中的代表性均等。这些情绪类别与 [7] 中使用的 30 个类别相同：崇拜、愉悦、愤怒、敬畏、困惑、轻蔑、满足、渴望、失望、厌恶、痛苦、狂喜、兴高采烈、尴尬、恐惧、内疚、兴趣、中性、疼痛、自豪、恍然大悟、宽慰、浪漫之爱、悲伤、宁静、羞耻、负面惊讶、正面惊讶、同情和胜利。说话者获得了文本

提示语和场景，以帮助引发情绪反应；所使用的提示语是 [7]，使用的版本的修改和扩展版，详见在线补充材料[1]。

数据收集自无偿志愿者，也通过 Mechanical Turk 向众包工作者收集；研究人员创建了一个网站，供说话者使用自己的电脑或移动设备录制并回放自己的样本。
音频文件最初根据参与者的硬件，以 44100 或 48000 Hz 录制，并存储为 MP3文件。每个单独的录音文件都以一个六位非连续用户 ID、一个两位数的情绪ID（1-30）和一个一位数的录音编号（1、2、3）命名。由于文件按用户 ID 标记，研究人员可以按说话者划分训练集、测试集或验证集，确保同一说话者的提交只出现在其中一个数据集中。（研究人员已尽力避免说话者提供多个样本，但无法百分之百保证。）所有参与者均提供了知情同意，且研究程序和设计的所有方面均经佛罗里达大学机构审查委员会（IRB）批准。

数据收集过程中，质量保证占据了重要部分；有些提交的内容完全是静音录音，或只包含随机的背景噪声。部分贡献者似乎误解了任务要求，录下自己朗读类别名称或与类别相关短语的声音。许多说话者提供了大量高质量样本，但也提交了一些有问题的样本，通常是由于音频问题，例如背景噪声（如手机提示音或背景交通声）；另一个问题是麦克风收录到了过多的呼吸噪声。遇到这些情况时，会要求说话者重新录制有问题的样本，以确保每位说话者提交的样本数量相同。

此外，有些说话者似乎无法根据提示说出富有感染力的语音；他们的回应没有传达出鲜明的情绪。这最后一组说话者未被纳入数据集。因此，受上述因素影响，这个数据集几乎可以肯定会偏向北美英语使用者的情感表达，因为作者兼唯一评估者拥有相同的个人背景。

数据集　　将在以下网址公开提供：https://github.com/fredbuhl/EmoGator　　.
我们评估了多种数据预处理步骤。将每个音频样本的数值范围归一化至 [-1,1] 范围内，是训练、验证和测试中普遍采用的方法。此外，我们还评估了对音频文件进行去噪，以及裁剪音频文件开头和结尾的静音部分。我们还探索了通过为每个样本创建移调和变速版本来扩充数据。

在收集该数据集期间，一家名为 Hume AI 的公司也收集了自己的声音爆发数据集，其中一部分作为 Hume-VB 数据集，提供给 ICML 2022表达性发声研讨会与竞赛[26]使用。该数据集包含来自 1702 位说话者的 59,201 条发声，涵盖 10 种情绪类别（愉悦、敬畏、尴尬、痛苦、兴奋、恐惧、惊恐、悲伤、惊讶和胜利）。每个样本均由评审者评分，并为每个样本的每种情绪类别提供 [0:100] 强度评分。该 Hume-VB 数据集也用于 ACII 2022情感声音爆发研讨会与竞赛[27]
EmoGator 数据集与 Hume-VB 数据集之间有几处差异：

1. EmoGator 有 30 个不同的情绪类别，每个样本都属于一个由说话者意图决定的类别。Hume-VB 的评审者会为每个样本的 10 个类别分别提供 0–100 分的评分——这反映的是听者的解读，在某些情况下可能与说话者意图大相径庭。2. EmoGator 的贡献者会收到描述特定情境的文本提示，这些情境旨在引发某一类别的发声爆发。Hume-VB 的贡献者则会收到用于模仿的“种子”发声爆发音频样本，这可能会缩小特定类别的表达范围。3. EmoGator 只允许每位说话者提交一次，每次 90 个样本；Hume-VB 则允许每位说话者多次提交。4. EmoGator 的类别分布均衡；每个情绪类别恰好有 1,071 个样本。Hume-VB 中的数量则有所不同；例如，“能够明确传达胜利的样本较少” [26, p. 2]5. 尽管 Hume-VB 的样本数量几乎是 EmoGator 的两倍，但该数据集仅供两项赞助竞赛使用，且需要签署最终用户许可协议（EULA）[2]；EmoGator 则依据开源许可证免费提供。

截至发表时间，EmoGator 似乎是目前公开可用的最大声音爆发数据集。

[1] https://supp.apa.org/psycarticles/supplemental/amp0000399/amp0000399_补充材料.docx [2] https://www.competitions.hume.ai/exvo2022

2.2 分类方法

语音情感识别、声音分类及其他领域中使用的多种不同技术，已应用于这类音频分类问题。

2.3 频谱图方法

一些音频分类方法会将音频信号转换为时频频谱图（或类似频谱图）的表示形式，这种表示可以通过多种方式生成。通常使用短时傅里叶变换（STFT） [28] ，以提供不同频率随时间变化的振幅；其变体梅尔频谱图会调整频率，使其符合梅尔刻度 [29], ，从而更贴近人类对音高差异的感知。梅尔频率倒谱系数（MFCC）提供一种类似频谱的“倒谱” [30], ：虽然采用梅尔频率，但它呈现的是相移下以分贝表示的振幅对数，而非频谱图所用的时域信息。生成的频谱图或倒谱图会作为其他机器学习方法的特征。

2.4 基于原始波形训练一维卷积神经网络

在 [31], 中，Dai 等人采用直接的声音分类方法：使用处理原始输入波形的一维卷积神经网络，而不使用频谱图或其他表示形式作为中间特征检测器。为此，他们使用了由多层一维卷积神经网络（1D CNN） [32] 构成的网络。 [31] 研究采用了 UrbanSound8k 数据集 [33], ，该数据集包含 10 个类别和 8,732 个样本，规模略小于 EmoGator 数据集。在测试了多种架构后，他们报告称，18 层模型的准确率最高达 71.68%，与使用同一数据集频谱图的卷积神经网络相比具有竞争力。对于 EmoGator 数据集，我们开发了一个与[31],相同的 18 层网络，并在每个一维卷积之后添加丢弃层，以帮助防止过拟合。

2.5 随机森林

我们也探索了随机森林分类器 [34] 。随机森林通过生成多个随机决策树构建而成，每棵树都基于数据集的一个随机子集构建，并使用每个样本的特征中的一个随机子集。构建完成后，森林中的每棵树都会为某个类别投一票，得票最多的类别即为胜者。这种方法可用于原始数据，也可用于类似频谱图的表征。

2.6 大型预训练语音模型

2022 年 ICML 表达性发声研讨会与竞赛中的多个团队使用了大型预训练语音模型 [35], [36], [37], [38],[39],[40]。其中两种模型得到了广泛使用：WavLM [41] 和 HuBERT [42]。这两种模型都是自监督语音表征模型，采用 Transformer 架构 [43]构建；Transformer 已成功应用于众多领域——它们通常是规模非常大的模型，使用大型数据集进行了长时间训练。借助这些预训练模型，可以取得比单独训练其他（通常规模较小的）数据集更好的结果。

WavLM 是一种大规模自监督预训练语音模型。WavLM 的“Large”版本使用 9.4 万小时的语音进行训练，拥有 3.1662 亿个参数。HuBERT 是一种类似的模型，其“large”版本拥有 3.17 亿个参数，并使用 128 个图形处理器（GPU）在 6 万小时的音频上进行训练。WavLM 和 HuBERT 都基于 wav2vec 2.0 [44], a 这一“对比学习”自监督语音模型构建；wav2vec 本身使用 64 个 GPU 进行训练，其输出被用作 HuBERT 或 WavLM 的输入，为它们提供更高层次的特征，以供构建和训练。

WavLM 实验首先将 EmoGator 的训练数据、验证数据和测试数据输入经过预训练的 WavLM 模型，并将最后一个隐藏层保存为每个样本的新表征，采用 70% / 15% / 15% 的训练集、验证集和测试集划分。随后，将训练数据的隐藏层用作输入来训练一个全连接网络，并使用验证数据确定合适的停止点；确定理想模型后，再在测试数据上运行。HuBERT 模型采用相同的方法——使用 HuBERT 模型的最后一个隐藏层，而非 WavLM 的隐藏层，作为全连接层的输入。

本研究能够采用 WavLM 和 HuBERT，很大程度上得益于 HuggingFace Transformer 库 [45],。这些库最初涵盖自然语言处理，如今已扩展到许多其他领域。只需几行代码就能集成大型预训练语言模型，这一优势怎么强调都不为过。

2.7 集成方法

集成方法通过结合多个模型的输出，并进行适当的训练和加权，来提升性能；集成结果通常优于单个模型。针对 EmoGator 数据使用了两种方法：**集成 A** 将 WavLM 与 HuBERT 单层模型生成的长度为 n 的输出（其中 n 为情绪类别数）取平均值，并根据所得平均值选出最可能的情绪类别。**集成 B** 将 WavLM 和 HuBERT 的最后隐藏层拼接起来，然后用这些输入训练一个全连接层。

2.8 平台与硬件要求

本项目的大部分工作是在佛罗里达大学的 HiperGator-AI 集群上完成的，该集群使用 80G A100 GPU；一块 A100 GPU 应足以运行所包含的所有模型，但除非对批量大小等参数进行调整，否则代码可能无法直接在显存较小的 GPU 系统上运行。

3 3. 结果

3.1 在原始波形上训练一维卷积神经网络

对于一维卷积神经网络，在完整数据集上取得的最佳结果来自70% / 15% / 15%的训练集/验证集/测试集划分，使用基于 [31], 的18层一维卷积神经网络，但在每个卷积层后都加入了丢弃层。0.07的较低丢弃率效果最佳。所有实验的批量大小均为128，并使用学习率为0.001的Adam优化器。我们计算了多项统计指标；对于完整的30类别数据集，平均F1分数为0.270。各类别的F1分数和其他准确率指标见表1；图1给出了F1分数最高的那次运行所对应的混淆矩阵。
以上实验均使用归一化的音频数据，但未对音频信号进行去噪，也未裁剪开头和结尾处的静音；此前采用70%/30%训练集/测试集划分的实验表明，对音频信号进行去噪或裁剪会降低性能。

我们也探索了数据增强：通过对音频样本分别进行独立的音高变换和节奏变换，生成了规模为原训练集（70% / 15% / 15%）两到三倍的“拉伸”版本；然而，尽管调整了音高和节奏缩放的幅度，拉伸后的训练集性能仍不如原始训练集。

回顾这些结果可以发现，有些类别要难以（或容易）得多；例如，表现最差的类别尴尬，其 F1分数（0.056）远低于表现最好的类别愉悦（0.627）。混淆矩阵很好地说明了这个问题：某些类型的声音爆发确实很难归入正确类别。根据混淆矩阵，尴尬（仅有 7 个样本被正确识别）更容易被判定为羞耻（16 个）或内疚（10 个）；这些概念彼此密切相关，且都可能产生相似的发声。这是一个本质上很难解决的问题，也有助于解释为什么人类只能可靠地区分 24 种《情绪》类别 [7]。

通过有选择地移除表现不佳的情绪类别，预期整体性能会有所提升。以 F1分数为指标，移除了得分最低的类别，创建了包含 24 类、16 类和 10 类的数据集子集。有趣的是，为创建 24 类子集而移除的六个得分最低的类别中，有三个也无法被人类在 [7] 中识别；而在 16 类子集中，又移除了另外两个无法被人类识别的类别。这表明两个数据集存在一定共性，也说明即使在不同研究中，人类和算法都难以处理某些情绪类别。

使用了相同的一维卷积神经网络模型架构、超参数和验证方法。结果见表 2；随着更易产生歧义的类别被逐步剔除，我们确实观察到性能有所提升。
通过创建二元一维卷积神经网络分类器（每种可能的情绪类别配对各设一个分类器），我们可以展示哪些类别配对最容易区分。采用相同的模型架构和 70%/15%/15% 的划分方式，并以 F1分数作为相似度指标（在 [0,1] 刻度上，1 表示最不相似），根据 30 个类别的 435 种排列创建了相似度矩阵，并根据该矩阵生成了展示各类别间关系的树状图（图 2）。树状图展示了最容易混淆或区分的类别。例如，它显示愉悦类别很容易与所有其他类别区分开来，也显示恍然大悟和轻蔑是最相似、因而最容易混淆的类别，尽管它们代表的情绪截然不同。

表1 ：带有丢弃层的18层一维卷积神经网络最佳运行结果的精确率、召回率和F1分数 s.

	精确率	召回率	F1分数	支持度
崇拜	0.407	0.488	0.444	162
愉悦	0.561	0.710	0.627	162
愤怒	0.405	0.327	0.362	162
Awe	0.220	0.296	0.253	162
困惑	0.354	0.574	0.438	162
轻蔑	0.236	0.296	0.263	162
满足	0.193	0.272	0.226	162
渴望	0.253	0.309	0.278	162
失望	0.144	0.093	0.113	162
厌恶	0.376	0.580	0.456	162
痛苦	0.243	0.111	0.153	162
狂喜	0.187	0.123	0.149	162
兴高采烈	0.190	0.074	0.107	162
尴尬	0.078	0.043	0.056	162
Fear	0.341	0.179	0.235	162
内疚	0.175	0.105	0.131	162
兴趣	0.288	0.420	0.342	162
中性	0.397	0.568	0.467	162
Pain	0.276	0.438	0.339	162
自豪	0.175	0.086	0.116	162
恍然大悟	0.351	0.241	0.286	162
宽慰	0.294	0.432	0.350	162
浪漫之爱	0.121	0.074	0.092	162
悲伤	0.355	0.302	0.327	162
宁静	0.209	0.191	0.200	162
羞耻	0.197	0.154	0.173	162
负面惊讶	0.296	0.364	0.327	162
正面惊讶	0.248	0.198	0.220	162
同情	0.233	0.370	0.286	162
胜利	0.378	0.228	0.285	162
准确性			0.288	4860
宏平均	0.273	0.288	0.270	4860
加权平均	0.273	0.288	0.270	4860

表 2：一维卷积神经网络在 EmoGator 数据集的 24、16 和 10 类别子集上的运行结果，与包含 30 个类别的完整数据集进行比较。

一维卷积神经网络数据集大小	F1分数（平均值）
30 类完整数据集	0.267
24 类子集	0.344
16 类子集	0.459
10 类子集	0.597

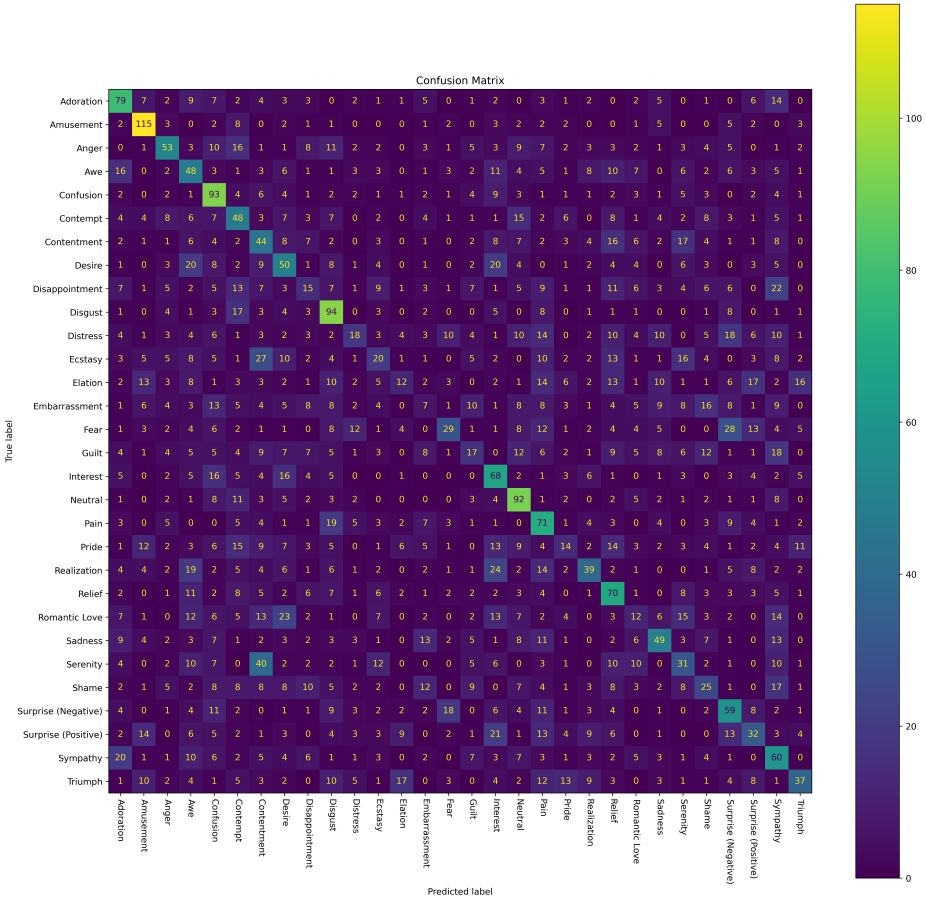

图 1：18层一维卷积神经网络通过丢弃层生成的混淆矩阵。

3.2 随机森林

如 [34], 所示，一种称为随机森林的方法已用于一些类别数量较少、样本数量也较少的数据集，这表明它可能适用于 EmoGator 数据集。该分类器（包含在 Scikit-learn 库中 [46]）使用音频数据的梅尔频率倒谱系数（MFCC）进行训练；实验分别针对完整的30类别数据集以及24、16和10类别子集进行。然而，所有结果都不如一维卷积神经网络的结果（见表3）。

3.3 大型预训练语音模型

结果是使用 WavLM 和 HuBERT 模型的最后一个隐藏层，并将其连接到一个全连接网络层后计算得出的。集成 B 的一个变体采用了两个全连接层（标记为“2 层全连接层”），取得了适度的提升。这些结果与其他结果一同列于表 4 中。

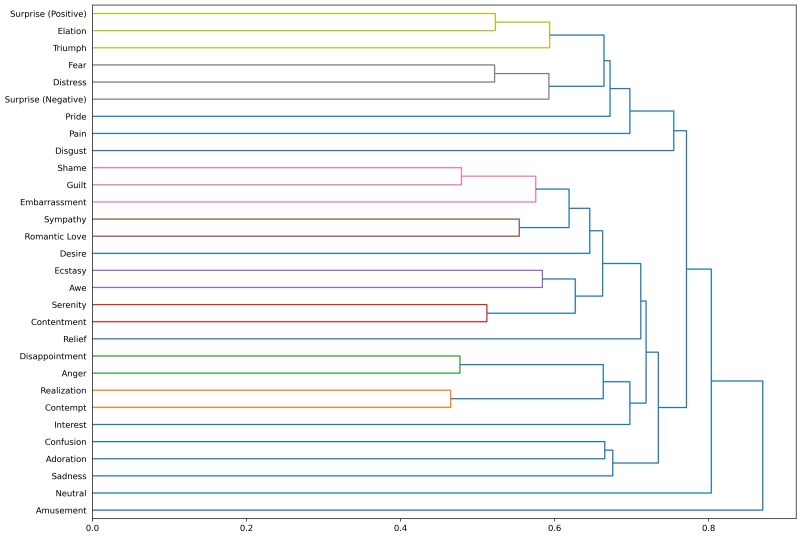

图 2：根据情绪类别两两之间的 F1分数（范围 [0,1]）生成的树状图。

表 3：随机森林　　使用 EmoGator 数据集的 24、16 和 10 类别子集进行的运行，与 30 类别完整数据集，使用 MFCCs.

随机森林数据集大小	**F1分数（平均值）**
30 类完整数据集	0.146
24 类子集	0.180
16 类子集	0.256
10 类子集	0.345

3.4 集成方法

结果基于两种方式计算：一是对 WavLM 和 HuBERT 模型运行后接入的全连接层所产生的输出取平均（集成 A）；二是拼接两个模型最后一层隐藏层的输出（集成 B），然后用这些输出训练单个全连接层。为避免测试数据污染集成模型，采用了在验证数据集 上平均 F1分数最高的 WavLM 和 HuBERT 单个全连接层。

集成方法的结果见表 4，其中还列出了所有 EmoGator 实验的汇总数据。

4 讨论

回到我们的研究问题——机器是否能像人类一样可靠地识别 24 种情绪类别——结果似乎表明，24 类别实验所取得的成绩未达到人们假定的人类水平；在 24 类子集中，采用一维卷积神经网络方法取得的最高 F1分数仅为 0.344。24 类、16 类和 10 类子集的结果均优于完整的 30 类实验，其中 10 类实验的表现最好，同样采用一维卷积神经网络方法，得分为 0.597。（为了更直观地理解这些结果，在 24 类子集中，随机猜测的正确率仅为 4.2%；在 10 类子集中，随机猜测的正确率仅为 10%——因此，这些结果远胜于纯粹的偶然猜测。）

该数据集的一种潜在用途是衡量人类对发声爆发的判断准确度——听众能否正确识别说话者意图传达的类别。其他研究使用听众对各类别给出的梯度评分量表，但不一定会将其与

表 4：所用各种方法和数据集子集的全部结果。

方法	# 类别	F1分数
一维卷积神经网络	30	0.267
一维卷积神经网络	24	0.344
一维卷积神经网络	16	0.459
一维卷积神经网络	10	**0.597**
随机森林	30	0.146
随机森林	24	0.180
随机森林	16	0.256
随机森林	10	0.345
WavLM	30	0.255
WavLM	10	0.563
HuBERT	10	0.531
集成 A	10	0.571
集成 B	10	0.591
集成 B（两层全连接）	10	0.593

说话者意图的真实标签联系起来。另一个问题是，收集受基于文本的提示启发而产生的声音爆发，是否比尝试从录制的对话中“自然采集”，或通过其他类型的提示引发声音爆发更好。

毫无疑问，收集更多数据会改善这些结果；尽管这个声音爆发数据集目前是公开可用数据集中规模最大的，但按照机器学习的标准来看，它仍然很小。对数据集的子集进行评估会让情况变得更糟；例如，考察 10 类子集时，只使用了数据集的[13]。

采用更复杂的集成方法似乎是一个很有前景的方向；尽管这里的集成结果并未超过一维卷积神经网络的结果，但加入更多单个模型，或许能够将准确性提升到我们目前所能达到的水平之上。

这里没有探讨的一个主题是生成声音爆发；作者接下来将探索生成对抗网络（GAN）和 Stable Diffusion 模型等方法来生成声音爆发；理想情况下，只需提供少量音频样本，就能针对某位说话者进行定制（ICML 竞赛就将此作为挑战之一）。

更多数据会有所帮助，但仅凭音频数据可能不足以对声音爆发进行恰当分类。融合视频和音频数据的数据集与模型，不仅可以观察面部表情，还能捕捉可能引发声音爆发的视觉线索，从而提高准确性。发声者在声音爆发前后说的话，以及周围其他人说的话，也可能有助于识别。（不过，无论还能收集到多少额外信息，声音爆发分类或许都有其固有局限，难以做到完全确定——悲伤的哭声和愉悦的哭声听起来往往一样，而且人们有时会说，他们不确定“该笑还是该哭”。）

另一个值得探索的方面是说话者的人口统计特征；他们的年龄、性别、出生地和文化背景都可能影响对发声爆发的分类。这些人口统计方面的考量也适用于评估样本质量的人；理想情况下，评审者的人口统计特征应与提交者相匹配，以获得最佳质量。

除了人口统计特征之外，每个人独特的性格和个性当然也会影响他们产生发声爆发的方式。因此，事先了解发声者的经历可能是提高准确性的关键，尤其是在能够根据这些经历对模型权重进行微调的情况下。

希望 EmoGator 数据集能向研究人员介绍发声爆发这一引人入胜的领域；也希望其他研究人员未来能将该数据集纳入规模更大的集合中，并通过公开这些数据集来“传递善意”。

致谢

感谢 Anand Rangarajan 与我们就该项目进行了富有助益的讨论。

参考文献

[1] G.B. Duchenne、G.B.D. de Boulogne、R.A. Cuthbertson、A.S.R. Manstead 和 K. Oatley。《人类面部表情的机制》。剑桥在线图书。剑桥大学出版社，1990年。[2] Alan S. Cowen 和 Dacher Keltner。面部所表达的内容：描绘自然表情所传达的28种情绪。《美国心理学家》，页码未指定–页码未指定，2019年。[3] Alan S. Cowen 和 Dacher Keltner。自我报告捕捉到27种不同的情绪类别，这些类别由连续梯度相互连接。《美国国家科学院院刊》，114(38):E7900–E7909，2017年9月。[4] Alan S. Cowen、Petri Laukka、Hillary Anger Elfenbein、Runjing Liu 和 Dacher Keltner。两种文化中，通过言语韵律识别12种情绪时类别的首要性。《自然·人类行为》，3(4):369–382，2019年4月。[5] Petri Laukka、Hillary Anger Elfenbein、Nutankumar S. Thingujam、Thomas Rockstuhl、Frederick K. Iraki、Wanda Chui 和 Jean Althoff。五个国家中声音里的情绪表达与识别：基于声学特征的透镜模型分析。《人格与社会心理学杂志》，111(5):686–705，2016年11月。[6] Emiliana R. Simon-Thomas、Dacher J. Keltner、Disa Sauter、Lara Sinicropi-Yao 和 Anna Abramson。声音传达特定情绪：发声爆发表现所提供的证据。《情绪》，9(6):838–846，2009年。[7] Alan S. Cowen、Hillary Anger Elfenbein、Petri Laukka 和 Petri Keltner。描绘人类短促发声所传达的24种情绪。《美国心理学家》，74(6):698，2019年。[8] Elena Lyakso 和 Olga Frolova。声音特征中的情绪状态表现：黑猩猩、人类婴儿、儿童、成年人。载于 Andrey Ronzhin、Rodmonga Potapova 和 Nikos Fakotakis 编，《语音与计算机》，《计算机科学讲义》，第201–208页，尚，2015年。施普林格国际出版。[9] Mariana Vaillant-Molina、Lorraine E. Bahrick 和 Ross Flom。低龄婴儿会匹配其他婴儿的面部和声音情绪表达。《婴儿期：国际婴儿研究学会官方期刊》，18(Suppl 1)，2013年8月。[10] Amaya Palama、Jennifer Malsert 和 Edouard Gentaz。6个月大的人类婴儿能否将情绪信息（快乐或愤怒）从声音迁移到面孔？一项眼动追踪研究。《公共科学图书馆·综合》，13(4):e0194579，2018年4月。[11] Lois Bloom 和 Richard Beckwith。《带着情感交谈：整合早期语言发展中的情感与语言表达》。《认知与情绪》，3(4):313–342，1989年10月。出版方：劳特利奇 _电子版：https://doi.org/10.1080/02699938908412711。[12] Yang Wu、Paul Muentener 和 Laura E. Schulz。1至4岁儿童会将各种积极情绪发声与其可能原因联系起来。《美国国家科学院院刊》，114(45):11896–11901，2017年11月。[13] K. M. Heilman、R. Scholes 和 R. T. Watson。听觉情感失认症：情感性言语理解受损。《神经病学、神经外科学与精神病学杂志》，38(1):69–72，1975年1月。出版方：英国医学杂志出版集团有限公司，栏目：研究论文。[14] G. H. Monrad-Krohn。韵律失调或“语言旋律”改变。《脑：神经病学杂志》，70:405–415，1947年。地点：英国；出版方：牛津大学出版社。[15]Sabine Skodda、Heiko Rinsche 和 Uwe Schlegel。帕金森病患者韵律失调随时间的进展——一项纵向研究。《运动障碍》，24(5):716–722，2009年。_电子版：https://movementdisorders.onlinelibrary.wiley.com/doi/pdf/10.1002/mds.22430。[16] Tsai-Hsuan Tsai、Hsien-Tsung Chang、Shin-Da Liao、Hui-Fang Chiu、Ko-Chun Hung、Chun-Yi Kuo 和 Chih-Wei Yang。在社交聊天机器人中采用基于语音的情绪识别功能，促进学龄前儿童的社会与情感学习。载于 Constantine Stephanidis 编，《2019年国际人机交互会议：最新研究论文》，《计算机科学讲义》，第341–356页，尚，2019年。施普林格国际出版。[17] Young-Shin Lee 和 Won-Hyung Park。基于 Fast R-CNN 的面部表情抑郁障碍诊断模型。《诊断学》，12(2):317，2022年1月。[18] Cynthia Breazeal。情绪与社交型人形机器人。《国际人机交互研究杂志》，59(1):119–155，2003年7月。[19] Jekaterina Novikova、Christian Dondrup、Ioannis Papaioannou 和 Oliver Lemon。同情始于微笑，智慧始于言语：口语人机交互中的多模态特征应用。arXiv:1706.02757v*1 [*cs*]*, 2017年6月。

[20] D. O'Shaughnessy。《语音通信：人类与机器》。威利出版社，2000年。[21] Rosalind W. Picard。《情感计算》。载于《情感计算》。麻省理工学院出版社，2000年。[22] Shashidhar G. Koolagudi 和 K. Sreenivasa Rao。语音情感识别：综述。《国际语音技术期刊》，15(2):99–117，2012年6月。[23] R. Cowie 和 E. Douglas-Cowie。对语音信号及情感韵律标记进行自动统计分析。载于《第四届国际口语语言处理会议论文集》。ICSLP '96，第3卷，第1989–1992页，1996年10月。[24] Akanksha Gadikar、Omkar Gokhale、Subodh Wagh、Anjali Wankhede 和 P. Joshi。使用神经网络进行语音情感识别的调查。《国际研究与分析评论》，7(3)，2020年9月。[25] Sepp Hochreiter 和 Jürgen Schmidhuber。长短期记忆。《神经计算》，9(8):1735–1780，1997年。[26] Alice Baird、Panagiotis Tzirakis、Gauthier Gidel、Marco Jiralerspong、Eilif B. Muller、Kory Mathewson、Björn Schuller、Erik Cambria、Dacher Keltner 和 Alan Cowen。ICML 2022表达性发声研讨会与竞赛：声音爆发的识别、生成与个性化，2022年7月。arXiv:2205.01780 [cs, eess]。[27] Alice Baird、Panagiotis Tzirakis、Jeffrey A. Brooks、Christopher B. Gregory、Björn Schuller、Anton Batliner、Dacher Keltner 和 Alan Cowen。ACII 2022情感发声爆发研讨会与竞赛：理解一种长期未受充分研究的情感表达模态，2022年7月。arXiv:2207.03572 [cs, eess]。[28] E. Jacobsen 和 R. Lyons。滑动离散傅里叶变换。《IEEE信号处理杂志》，20(2):74–80，2003年3月。会议名称：IEEE信号处理杂志。[29] S. S. Stevens、J. Volkmann 和 E. B. Newman。心理学音高量值测量量表。《美国声学学会杂志》，8(3):185–190，1937年1月。出版方：美国声学学会。[30] B. Bogert。回声时间序列的倒频率分析：倒谱、伪自协方差、互倒谱与倒谱裂解。载于《时间序列分析研讨会论文集》，第209–243页，1963年。[31] Wei Dai、Chia Dai、Shuhui Qu、Juncheng Li 和 Samarjit Das。用于原始波形的超深卷积神经网络。arXiv:1610.*00087 [*cs*]*,2016年10月。arXiv: 1610.00087。[32] S. Kiranyaz、T. Ince、R. Hamila 和 M. Gabbouj。用于患者特异性心电图分类的卷积神经网络。载于2015年第37届IEEE医学与生物工程学会国际年会（EMBC），第2608–2611页，2015年8月。ISSN: 1558-4615。[33] Justin Salamon、Christopher Jacoby 和 Juan Pablo Bello。城市声音研究的数据集与分类体系。载于第22届ACM多媒体国际会议论文集，MM '14，第1041–1044页，美国佛罗里达州奥兰多，2014年11月。计算机协会。[34] Leo Breiman。随机森林。《机器学习》，45(1):5–32，2001年10月。[35] Detai Xin、Shinnosuke Takamichi 和 Hiroshi Saruwatari。探索自监督学习与分类器链在非言语发声情绪识别中的有效性，2022年6月。arXiv:2206.10695[cs, eess]。[36] Chin-Cheng Hsu。基于离散语音表征合成个性化非语音发声，2022年6月。arXiv:2206.12662 [cs, eess]。[37] Josh Belanich、Krishna Somandepalli、Brian Eoff 和 Brendan Jou。结合残差网络与预训练副语言Conformer模型的多任务发声爆发建模，2022年6月。arXiv:2206.12494 [cs, eess]。[38] Roshan Sharma、Tyler Vuong、Mark Lindsey、Hira Dhamyal、Rita Singh 和 Bhiksha Raj。用于年龄、情绪和国家预测的自监督与可学习时频感受野，2022年6月。arXiv:2206.12568 [cs, eess]。[39] Tilak Purohit、Imen Ben Mahmoud、Bogdan Vlasenko 和 Mathew Magimai Doss。比较ExVo多任务学习赛道中的监督式与自监督嵌入，2022年6月。arXiv:2206.11968 [cs, eess]。[40] Atijit Anuchitanukul 和 Lucia Specia。Burst2Vec：一种用于根据声音爆发预测情绪、年龄和来源的对抗式多任务方法，2022年6月。arXiv:2206.12469 [cs, eess]。[41] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka、Xiong Xiao、Jian Wu、Long Zhou、Shuo Ren、Yanmin Qian、Yao Qian、Jian Wu、Michael Zeng、Xiangzhan Yu 和 Furu Wei。WavLM：面向全栈语音处理的大规模自监督预训练，2022年6月。arXiv:2110.13900 [cs, eess]。

[42] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Ruslan Salakhutdinov 和 Abdelrahman Mohamed。HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习，2021年6月。arXiv:2106.07447 [cs, eess]。[43] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Lukasz Kaiser 和 Illia Polosukhin。Attention is All you Need。《第31届NIPS会议论文集》，2017年。[44] Alexei Baevski、Henry Zhou、Abdelrahman Mohamed 和 Michael Auli。wav2vec 2.0：一种用于语音表征自监督学习的框架。《style id='8'>arXiv:2006.*11477 [cs,eess]*,2020年6月。arXiv: 2006.11477。[45] Thomas Wolf、Lysandre Debut、Victor Sanh、Julien Chaumond、Clement Delangue、Anthony Moi、Pierric Cistac、Tim Rault、Rémi Louf、Morgan Funtowicz 和 Jamie Brew。HuggingFace的Transformers：先进的自然语言处理。《style id='15'>arXiv:1910.*03771 [cs]*,2019年10月。arXiv: 1910.03771。[46] Fabian Pedregosa、Gaël Varoquaux、Alexandre Gramfort、Vincent Michel、Bertrand Thirion、Olivier Grisel、Mathieu Blondel、Peter Prettenhofer、Ron Weiss、Vincent Dubourg、Jake Vanderplas、Alexandre Passos、David Cournapeau、Matthieu Brucher、Matthieu Perrot 和 Édouard Duchesnay。Scikit-learn：Python中的机器学习。《机器学习研究杂志》，12:2825–2830，2011年11月。