FORUM ACUSTICUM
EURONOISE 2025

面向可解释的情绪识别：使用机器学习识别关键特征

亚库巴 · 卡洛加 * **伊纳 · 科达西**

通信信号处理组，瑞士 Idiap 研究 institute，Martigny，瑞士

arxiv:2508.04230v1[中译本]

ABSTRACT

无监督方法，如 wav2vec2 和 HuBERT，在音频任务中已达到最先进的性能，导致研究从可解释特征转向其他方向。然而，这些方法缺乏可解释性限制了它们在医学等关键领域中的应用，在这些领域中理解特征相关性至关重要。为了更好地了解无监督模型的特征，识别与给定任务相关的可解释特征仍然是至关重要的。在这项工作中，我们专注于情绪识别，并使用机器学习算法来识别和推广此任务中最重要且可解释的特征。虽然以前的研究探讨了情绪识别中的特征相关性，但它们往往受限于狭窄的背景并呈现不一致的结果。我们的方法旨在克服这些限制，为识别最重要且可解释的特征提供一个更广泛、更稳健的框架。

Keywords: 重要特征，情感识别，可解释性。

1. 介绍

有效沟通依赖于对话者之间对情感的准确相互感知。在对话中交换的信息有相当一部分是通过情绪暗示传达的——无论是视觉上，还是通过声音、语调和其他听觉特征。当情绪感知受损时，如某些病理情况 [1]，沟通变得特别具有挑战性，这强调了在听觉互动中保持这些线索的重要性。同时，我们音频互动中的很大一部分现在是通过电子系统进行的，包括远程会议、电话和助听器。虽然这些系统处理声音以优化信号质量、信号可理解性、能源效率或延迟等因素，但很少考虑情绪暗示的保存。通过人类主观测试评估各种处理方案对情绪暗示的影响是不切实际的，因为成本高且扩展性有限。使用机器学习模型进行语音情感识别（SER）的不同音频处理方案的自动化系统可以提供一种可行的替代方案。先前的研究表明人与机器在情绪感知上的相似之处，通常机器依赖于手工制作的声学特征来识别关键的情绪暗示 [2,3]。然后可以通过分析手工制作的声学特征是如何受到影响的 [4, 5] 或测量依赖这些特征的模型性能下降的程度 [4, 6] 来量化各种处理方案（或扰动）对情绪感知的影响。然而，这样的方法通常依赖于有限范围的声学特征和数据集数量，缺乏通用性，并突显了需要一个更广泛的情绪相关声学特征集的需求，这些特征可以在不同场景中应用。

我们的工作旨在使用各种自动分类器在不同的数据集上识别 SER 的相关声学特征。更具体地说，我们采用了六个分类模型，这比通常在 SER 文献中看到的数量要显著多得多。此外，我们在四种不同语言的六种不同数据集上进行了实验，每种数据集使用不同的分割多次训练了模型。利用多种模型（对同一数

*Corresponding author: yacouba.kaloga@idiap.ch.

Copyright: ©2025 Kaloga et al. This is an open-access article distributed under the terms of the Creative Commons Attribution 3.0 Unported License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

据的不同分割进行多次训练）和数据集提供了几个优势。首先，在多个模型（分别在多个数据集）中提取相关声学特征可以增加稳健性。如果某个特定特征在不同模型（分别在不同数据集）上一直很重要，这表明其可靠性并减少了单一模型（或数据集）特性的影响力。此外，使用多种模型有助于减轻表现不佳的模型的影响，这些模型的重要性不准确是因为性能差，以及依赖于高度特定于某一数据集的功能而表现出色的模型。其次，在多个模型和不同的数据集上测试特征重要性算法使我们能够评估在不同学习范式和数据集上的特征排名的一般化能力。最后，多次运行分析可以帮助减轻相关特征带来的挑战。具体而言，我们的贡献如下：

- 我们推导了在不同分类模型和数据集上最相关的声学特征用于情感识别。
- 我们提出了一种方法，将每个单独模型和数据集的相关特征结合起来，从而产生一个单一的、稳健且具有泛化能力的关键声学特征列表。
- 我们实验性地展示了我们方法在建立关键声学特征方面的优势，表明它在稳健性和泛化能力方面优于使用单一分类器或单一数据集。

2. 相关工作

语音情感感知的研究主要探讨了关键声学特征及其对人类感知的影响。在本节中，我们回顾这些研究及其局限性。此外，我们考察机器学习如何通过将其性能与人类感知进行比较来提供更普遍的发现。最后，我们讨论当前利用这种联系的研究为何仍然有限。

2.1 声学线索与人类情绪感知

众多研究表明特定的声学特征与人类对情绪的感知之间存在强烈联系。音高变化扮演着关键角色，研究显示较高的平均音高值与愤怒、恐惧和兴奋等高度唤醒的情绪相关联，而较低的音高值则与无聊等低度唤醒的情绪有关 [7–11]。时间方面如语速和节奏以及频谱测量包括共振峰和频谱通量也是传达情绪的关键因素 [2, 12]。其他声学线索如韵律轮廓、强度和能量相关特征进一步丰富了言语的情感表达能力 [13, 14]。重要的是，这些声学特征并非孤立地发挥作用，而是以复杂的方式相互作用来塑造情感感知。尽管有这些见解，现有的研究仍面临限制。大多数实验依赖于小样本量，因为基于参与者的评估需要较高的成本和时间投入。这导致了高度依赖具体情境的研究结果，难以进行概括 [9]。此外，一些研究报告了相互矛盾的结果，可能是由于方法论、数据集构成以及说话人差异所引起（例如，[15] 在悲伤方面与 [16–19] 存在不一致的发现）。鉴于这些限制，机器学习为研究情感感知提供了有希望的研究途径，因为它允许分析大规模的数据集并识别出人类有限研究难以察觉的潜在模式。

2.2 机器基于的 SER 和人类情感感知

最近的研究直接将机器学习模型与人类情绪感知进行了比较。例如，[4] 评估了分类模型（即多层感知器 (MLP) 和支持向量机 (SVM)）在干净和嘈杂条件下的人类感知情况。他们的研究引入了情绪干扰因素，即测试集中存在但在训练期间未见过的情绪，以确保模型进行的是真正的识别而非简单的区分。他们的结果显示，人类与机器的感知之间存在着显著相似性，包括整体 SER 性能可比、不同情绪下的相对准确性相似、在嘈杂环境中性能一致下降以及相似的混淆矩阵，特别是对于定义明确的情绪如愤怒和悲伤而言更是如此。这些发现与早期文献 [3, 6] 相吻合，该文献也报告了人类与机器 SER 之间的强烈相似性。此外，预测效价/唤醒位置的回归模型表明，基于人类的情绪定位可以有效预测在同一任务上的机器性能 [2] 。这表明，无论是人类还是机器学习模型，在解读情绪时都依赖于类似的声学线索 [3–5, 20] 。

2.3 声学线索与情感识别

据我们所知，只有有限数量的研究定量探讨了机器学习模型中 SER 最重要的声学线索。这项工作主要使用回归和固定效应回归 [2, 5, 15] 进行。由于特征重要性结果通常取决于使用的分类器/回归器，因此其结论的普适性受到限制。此外，考虑的特征数量相

对较少，这些分析可能会忽略一些关键特征。研究背景往往局限于少数说话人、几个数据集和受限的语言环境。在 [20] 中，强调了模型在跨数据集设置中的表现不如人类，突出了这些受约束环境中缺乏普适性的问题。此外，由于个别作者的选择——例如使用不同的情绪或声学线索——汇总结果以得出更广泛的结论仍然具有挑战性。鉴于这些限制，我们的目标是推导出能够广泛适用于多个代表性数据集和模型的最重要的声学线索。

3. 提议的方法

我们考虑 D 个数据集，每个包含 n_d 个语句 $\boldsymbol{x}_i$，$i = 1, \ldots, n_d$。每个话语 $\boldsymbol{x}_i$ 都与一个情感标签 y_i 相关联，例如幸福、愤怒、悲伤等。SER 任务是一个监督任务，在该任务中，我们训练模型学习预测输入话语 $\boldsymbol{x}$ 的正确情感标签 y。在本节中，我们提出了一种方法，用于使用各种数据集和不同模型识别最适合 SER 的相关声学特征。

3.1 特征集

每个话语 $\boldsymbol{x}_i$ 由一个固定大小的特征列表 $f = [f_1(\boldsymbol{x}_i), \ldots, f_Q(\boldsymbol{x}_i)] \in \mathbb{R}^Q$ 表示，旨在捕捉多种声学特性。为了构建这个特征列表，我们使用了 ComParE_2016 特征集 [21]，这是一个大型特征集合 ($Q = 6373$)，这些特征并非专门为 SER 设计，但已知可以捕获广泛范围的声学信息。该集合基于提取 65 低级描述符（LLD）例如音调、频谱能量和听觉频谱，以及另外 65 个作为它们时间差值派生出的 LLD。由于 LLD 随时间变化且话语长度不同，随后应用了一系列表征功能到这些 LLD 上（如各种分布矩、局部极值和不同类型分位数），以构建固定大小的特征列表。

3.2 导出关键声学特征

为了识别 SER 中最重要声学特征，我们使用了能够提供特征重要性评分的分类模型，量化每个特征对模型决策的贡献。要可靠地利用这些重要性评分来推导出最重要的声学特征，达到强大的分类性能是必要的。此外，必须特别考虑相关特征。在许多特征重要性模型中，一个变量可能被赋予高重要性，而其相关的同类特征尽管同样相关，可能会被忽视。相反，重要性可能会分散在一个相关变量簇上，导致个别特征缺乏重要性的误导印象。解决这一挑战是特征重要性分析的关键方面。这也是为什么我们的方法涉及多个模型、数据集和实验重复的原因。如果两个特征传达相同的信息，在所有这些实验条件下，它们应该表现出相似的平均或中位数重要性。

为了识别 SER 中最重要的声学特征，我们使用了 M 分类模型和 K 数据集。虽然在第 5.2节中使用了 $M = 4$ 和 $K = 6$，但所提出的方法适用于任何数量的模型 M 和数据集数量 K。每次在数据集 d 上使用 $m = 1, \ldots, M$ 和 $d = 1, \ldots, D$ 训练模型 m 时，结果就是分类性能 p_d^m 和每个特征的重要性得分 $s_{q,d}^m$。初步实验表明，在一步中汇总所有特征重要性得分（例如使用 $s_q = \frac{1}{DM} \sum s_{q,d}^m$）是低效的。具体来说，所得的特征选择不如使用单一模型中的特征选择在第 4.5 节介绍的评估指标方面有效。相反，我们建议通过以下步骤找到 SER 的关键声学特征。

首先，我们将所有考虑模型中每个独立数据集的归一化特征重要性分数进行聚合。聚合的重要性分数 $s_{q,d}$ 计算如下：

$$s_{q,d} = \text{median} \left\{ \frac{s_{q,d}^m}{\max\limits_{q=1}^{Q} \{s_{q,d}^m\}} \right\}_{m=1}^{M}. \tag{1}$$

然后，对于每个数据集 d，我们根据分数 $s_{q,d}$ 对 f 中的特征按重要性从高到低排序，构建一个新的特征列表 f_d^{order}。每个模型然后使用顶级特征的子集重新训练，变化范围从 0.5%到 20%，增量为 0.5%。一旦模型达到或超过使用所有特征获得的原始性能 p_d^m，此过程即停止。设 $pt(m)$ 表示模型 m 至少达到同等性能所需的顶级特征的最小百分比。然后我们定义多重集 M（允许重复元素）包含所有用于达到阈值

$$M = \{f_{q,d}^{\text{order}} \mid \forall q \in [\![1, Q]\!], \forall m \in [\![1, \mathrm{M}]\!], q \leq pt(m)\}. \tag{2}$$

的特征

由于这些特征对应于低级别描述符的统计信息，我们通过计算每种低级别描述符统计在多重集 M 中出现的次数来识别对情感识别最重要的声学线索，并根据频率对其进行排序。

4. 实验设置

在本节中，我们介绍了 SER 数据集、分类模型以及所使用的特征重要性算法。最后，本节以对模型训练和评估的详细描述作为结论。

4.1 数据集和情感

为了推导出一个通用的重要声学线索列表，我们使用了六种在语音情感识别（SER）文献中常用的数据集，即*咖啡* [22], *情感移动* [23], *情感数据库* [24], *演示* [25], *苔丝* [26] 和杰梅普-5 emo [27]。如表 1所示，这些数据集涵盖了四种语言——意大利语、德语、法语和英语。根据表 2所示的情绪映射，我们将标记为*中性*的陈述以及代表六种核心情绪的陈述考虑在内，即*幸福*、（热）*怒气*、*恐惧*、*悲伤*、*惊喜*和*厌恶*。

4.2 性能评估

在 SER 文献中使用的标准评估指标是未加权平均召回率（UAR），它是从每种情感的个别召回率推导出来的。模型对于特定情感的召回率是通过将该情感正确识别的数量除以数据集中该情感的总数量来计算的。这个指标比传统的准确性指标更受欢迎，因为它侧重于模型识别特定情感的能力，而不强调与其他情感之间的区分。UAR 仅仅是所有情感 [1] 召回率的算术平均值。

4.3 模型

用于计算 SER 中最重要的声学特征的分类器包括线性支持向量机（L-SVM）[28]、逻辑回归（LGRG）[29]、随机森林傅里叶（RFC）[30] 和轻型梯度提升机（LGBM）[31] 分类器。这些分类器因其固有的确定特征重要性的能力而被选中。然后使用两个额外的分类器验证结果，即径向基函数 SVM（RBF-SVM）[28] 和多层感知机（MLP）[32]（它们没有提供一种直接的方法来确定特征重要性）。这些分类器被选中是因为它们在 SER [4] 中表现出与人类感知的相似性。每个分类模型都有一组超参数，我们通过专注于关键参数的网格搜索对其进行优化，在训练效率和强大性能之间取得平衡（参见。第 4.4节）

4.4 超参数的选择

在本节中，我们概述了为每个考虑的模型选择最优超参数的过程。选择过程分为三个关键步骤，即：

1. *分割数据集。* 为了确保对说话者和性别特性的无偏评估，我们根据两个关键原则将每个数据集仔细划分为训练集和测试集。首先，来自特定说话者的所有发音都仅分配给训练集或测试集之一，防止任何说话者同时出现在两组中。其次，我们在测试集中尽可能争取男女平衡。遵循这些原则，划分策略如下：对于拥有超过 12 个说话者的数据集，20%的说话者被分配到测试集中。对于少于 12 个说话者的数据集，两名说话者被分配到测试集中。*情感移动力*（总计六名说话者）和*泰斯*（总计两名说话者）是例外情况，在这些情况下仅使用一名说话者进行测试。
2. *验证。* 为了选择最优超参数，我们使用分层的 5-折交叉验证。对于每个超参数配置 p，模型会在四个折叠上进行训练，并在剩余的一个折叠上进行评估。这个过程重复五次，确保每一个折叠都能一次作为验证集。这五次运行的结果会被平均计算，然后选择产生最高平均性能的配置作为最优超参数配置 p^*。这种交叉验证策略能够对不同的超参数组合进行全面评估，并保证每个模型的有效选择。对于较大的 DEMoS 数据集，由于其规模，仅使用一个折叠作为专用验证集进行一次评估。
3. *测试。* 最后，我们使用在步骤 2 中确定的超参数 p^* 对每个模型进行整个训练集上的训练。然

[1] 也称为宏召回率。

Table 1: 使用的表情识别数据集及其特征的总结。中性陈述表示具有情感上中性意义的陈述，片段是指根据说话人的句法或韵律从句子中提取的部分，而非理性陈述是没有实际含义但可能的语言声音序列。男性（女性）列汇总了每个数据集中男（女）发言者的人数。法国 C. 指加拿大法语。陈述列展示了每个数据集中的总陈述数量（#），在适用情况下不同情感下说出的不同句子的数量（# 句子），以及每个数据集中陈述的平均时长（μ）和标准差（σ）（以秒为单位）。

	数据集特征			发言者			话语		
数据集	**语言**	**动作**	**话语类型**	**# 男（F）**	**母语**	**演员**	**#**	**# 句子**	**持续时间 μ/σ(秒)**
咖啡	French	✓	Neutral	6 (6)	French C.	✓	936	6	4.4/0.8
演示	Italian	Induced	Chunks	45 (23)	Italian	✗	9697	-	2.9/1.3
情感数据库	German	✓	Neutral	5 (5)	German	✓	494	10	2.8/1.0
情感移动感	Italian	✓	Neutral	3 (3)	Italian	✓	588	14	3.1/1.4
盖梅普-5 情绪	French	✓	Non-sense	5 (5)	French	✓	50	2	2.6/1.2
蒂斯	English	✓	Neutral word	0 (2)	English	✓	2800	200	2.1/0.3

Table 2: 所考虑数据集中每种情绪的概述及该情绪的发言次数。

	情绪						
数据集	**幸福**	**愤怒**	**恐惧**	**悲伤**	**中性**	**厌恶**	**惊喜**
咖啡	120	120	120	120	60	120	120
演示	1395	1477	1156	1530	332	1678	1000
情感数据库	71	127	69	62	79	46	✗
情感移位	84	84	84	84	84	84	84
盖梅普-5 情绪	10	10	10	10	✗	✗	✗
蒂斯	400	400	400	400	400	400	✗
总计	2080	2218	1839	2206	955	2382	1204

后，我们在测试集上评估这个训练好的模型，以评估其在未见过的数据上的性能。

为了进一步减少偏差，此过程对每个数据集重复 3 次（即，使用 3 种不同的数据分割方式来划分训练集和测试集），并且这些三次运行的平均值对应于该模型在该特定数据集上的性能。

4.5 特征重要性评估

为了在实验部分验证上述过程的有效性，我们需要能够确定哪个特征重要性列表更好。为此，我们根据重要性顺序对两个列表中的特征进行排序。然后我们使用按重要性顺序递增的特征百分比重新训练模型。实现与使用所有特征相同性能但具有较小特征百分比的特征选择将被视为更好的选择。

5. 实验结果

在本节中，我们展示了之前描述的实验结果，并对最重要的特征进行了分析。请注意，在提及重要特征的选择时，我们的意思是选择具有最高重要性得分的特征，这些是从特征重要性系数排名列表中选出的。

5.1 所有模型在使用完整特征集的所有数据集上的性能

表 3总结了每个模型在所考虑的数据集上使用第 4.4节描述的训练和验证程序实现的未加权平均召回率（UAR）。

Table 3: 各考虑模型和数据集在三次运行中的 UAR 的均值和标准差。我们将每行中与最优值相差不超过 1%的所有数值用紫色突出显示。

数据集	RBF-SVM	多层感知器	请求评论	LGBM	L-支持向量机	LGRG	模型平均
咖啡	48.6±4.0	50.6±2.8	40.1±1.6	44.6±4.0	49.8±3.8	49.6±2.6	47.2±3.7
情感数据库	78.8±2.4	80.3±1.5	70.8±0.9	75.3±0.1	82.7±2.3	83.9±1.8	78.6±4.5
演示	74.2±0.1	74.9±0.8	51.6±0.6	63.8±0.8	65.9±0.3	66.0±0.3	66.1±7.7
情感移位	43.2±5.8	36.7±5.0	43.5±1.7	44.2±3.3	36.1±5.1	35.7±4.7	39.9±3.8
吉梅普-5 情绪	83.3±8.2	80.0±14.1	80.0±12.2	76.7±8.2	80.0±7.1	76.7±4.1	79.5±2.3
蒂斯	58.9±0.0	61.7±0.9	57.0±0.4	57.6±0.4	69.9±0.0	69.6±0.0	62.5±5.4
数据集上的平均值	64.5±15.2	64.0±16.2	57.2±14.2	60.4±13.0	64.1±16.5	63.6±16.3	62.5±15.5

从数据集级别的角度来看，很明显分类性能受到特定数据集的显著影响。值得注意的是，说话人的数量似乎会影响在三次运行中观察到的变化（每次运行涉及测试集中不同的说话人组合）。拥有较少说话人的数据集通常表现出更高的变化性（Tess 数据集是一个明显的例外）。这种增加的变异性可能源于说话人数目的有限，这限制了模型的泛化能力，并导致过度拟合训练数据中存在的特定于说话人的特征。这一观察强调了使用多样化和广泛的数据集的重要性，正如本研究中所做的那样。除了数据集规模外，没有发现分类性能与其他数据集特性（如表 1 所示的说话人的母语、所说语言或话语性质）之间的明确相关性。

从模型层面来看，RBF-SVM、L-SVM、MLP 和 LGRG 模型通常表现出最佳的整体性能。然而，在 EMOVO 数据集上，RFC 和 LGBM 模型显著优于这些模型。无论是由于过拟合，某些模型捕捉到了特定数据集的线索，还是仅仅因为某些模型更适合特定的数据集，这一观察结果突显了我们多模型方法的价值。其关键优势在于缓解任何单一模型可能表现出的局限性或过度专业化。

5.2 特征选择：多个模型 vs 单一模型

为了说明我们提出的方法在聚合特征重要性分数方面的优势，本节将比较两种条件下的模型训练性能：(i) 使用根据所有模型的聚合重要性评分得出的顶级特征，如第 3.2节所述，和 (ii) 仅基于每个单独模型的重要性的顶级特征。这些表现与使用所有特征时的基础线表现进行了比较。如第 4.4节所概述的那样，每次实验重复三次。图 1显示了每个数据集相对于基础线表现的差异，平均值是根据各种阈值计算得出的顶级特征的重要性的聚合评分或个别模型的重要性评分来考虑的所有已考虑模型 [2]。所示结果表明，在多个模型中使用聚合特征重要性分数的优势。具体而言，即使仅采用排名前 0.5%的特征，当基于聚合的重要性评分选择特征时，每个数据集的表现都高于使用个别模型排序的情况。此外，通过使用聚合的重要性和分选取排名前 6%的特征，所有数据集都能达到基础线表现。相比之下，当根据每个单独模型的个别重要性分数选取排名前 6%的特征时，只有三个数据集达到了基础线表现。

5.3 最重要的声学线索用于情感识别

第 5.2节的结果表明，使用通过聚合特征重要性得分选出的前 6% 特征可达到与使用完整特征集相同或更好的性能（参见图 1）。保留这些前 6% 的特征，我们

[2] 请注意，MLP 和 RBF-SVM 模型也包含在右侧图形的平均值中。没有 MLP 和 SVM 时结果基本保持不变。显著性能提升并非归因于它们的包含。

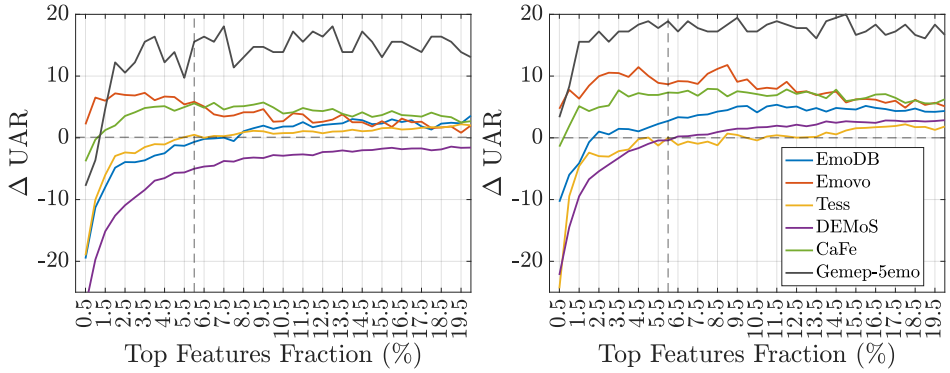

Figure 1: 所有模型在使用不同阈值的顶级特征重新训练时，相对于基线性能（即使用所有特征训练模型）在每个数据集上的平均 UAR 差异。左：基于每个模型自身的特征重要性得分选择的顶级特征。右：使用所有模型提出的聚合特征重要性得分选择的顶级特征。

得到一个包含 2,282 个特征的多重集，其中 1,687 个是不同的。如此多的不同特征突显了情感识别任务的复杂性，在这个任务中信息存在于多样且广泛的特征集合上。由于每个特征源自一个低级描述符（LLD），我们通过计算从它派生出的特征出现在顶级特征中的次数来确定该 LLD 的重要性（具有 ΔLLD 特征也归因于其提取自的具体 LLD）。图 2展示了这些 LLD 的标准化出现分布。可以观察到，少数几个 LLD，如 F0 最终和音频谱长度 L1 范数，具有很高的重要性。虽然整个 LLD 列表都是重要的，因为每个 LLD 都代表了 SER 中的一个重要声学线索，但该分布显示出急剧下降后又逐渐平稳的趋势。虚线表示的截断点代表着总出现次数的 50%，对应只有 16 个 LLD。尽管这个截断点用于说明目的并不具有内在意义，但它突出了特别重要的前 16 个 LLD。在本节的其余部分中，我们将提供对我们提出的方法确定为 SER 至关重要的这些前 16 个 LLD 的一些见解。

F0 最终。 基本频率，通常称为音高，在情感识别中被模型高度依赖作为 LLD。这一观察结果与现有的情感识别文献（见第 2.3节）一致，其中音高的重要性得到广泛认可。

音频谱。 听觉谱图是线性频率谱图的一种变换，考虑了人类听觉的非线性频率分辨率。由 LLDs 音频

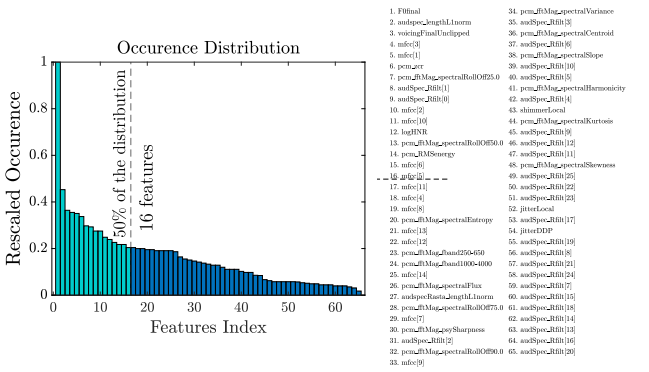

Figure 2: 标准化出现的低级描述符在排名前 6% 的特征中。

谱长度 L1 范数,audSpec_Rfilt[0] 和 audSpec_Rfilt[1] 捕获的听觉谱图的各种特征似乎是 SER 的重要声学线索。LLD 音频谱长度 L1 范数是听觉频谱的 l_1 范数的大小，大致对应于感知到的响度。低频带系数 audSpec_Rfilt[0] 和 audSpec_Rfilt[1] 是用于使听觉谱图更具抗噪性、抵抗不利条件和其他可能影响语音感知和分析因素的 Rasta 变换的第一组系数。

发音尾音未截断。 此 LLD 表示 F0Final 为浊音的概率。情感状态可以显著影响发声和声音强度，这两者都是声乐产生的重要因素。由于这些因素直接影

响 F0Final 是否为浊音，因此该变量预计会随说话人的情感参与程度而变化。

梅尔频率倒谱系数。 倒谱系数（MFCCs）是从应用于输入信号频谱的对数 Mel 尺度表示的余弦变换中导出的。这些系数提供了整体频谱轮廓的紧凑表示，并且对于录音条件和说话人特征的变化具有鲁棒性。较低的 MFCC 系数，对应于共振峰，比捕获更精细频谱细节和频谱快速变化的较高 MFCC 系数在情感识别（SER）中更为重要。

对数 HNR。 噪声与谐波比的对数表示浊音特征，因此也与说话人识别相关。

PCM. 除了上述特征外，重要特征的最后一类包括与信号或其频谱变化和能量相关的脉冲编码调制（PCM）特征。即使不结合人类听觉特异性，这些特征对于情感识别任务也是相关的。更具体地说：

- 该 pcm_RMS 能量度量与响度密切相关，而响度是 SER 的一个重要声学线索。

- 情绪对频谱能量分布有显著影响。例如，悲伤等情绪通常与高于 1kHz 的低谐波能量相关联，而愤怒、快乐或恐惧等情绪往往表现为该频率范围内的高谐波能量 [13]。因此，在情感识别（SER）中发现一些与此类频谱特征相关的特性是很常见的。这些包括 pcm_fft 频谱滚降 25.0，它表示信号能量的 25% 的频率界限，以及 pcm_fft 频谱滚降 50.0，它表示信号能量的 50% 的频率界限。

- (pcm_zc 率) 度量，即量化信号中符号交替次数的过零率，在关键特征中也占有重要地位。确实，与情感相关的语音表达、强度和音色的变化可能会影响过零率所捕捉到的时间变化。

这些是我们研究中强调的特征。这些 LLDs 的确切计算可以在 [33] 中找到。该项目的下一阶段将涉及开发一种方法，以明确或隐含地保留其中的一些特性，并确定这种方法的有效性。

6. 结论

在本文中，我们采用多种模型和数据集进行了全面分析，以识别 SER 任务中最重要且稳健的特征。据我们所知，这是首次利用如此多样化的数据集和模型来确定 SER 文献中的特征重要性。尽管 SER 内在复杂，但我们已经识别出几个在此背景下高度相关的关键特征。未来，我们将通过重复实验并将每种情绪的召回率作为评估指标来扩展我们的发现到特定的情绪上。此外，我们还将研究这些特征在背景噪声、信道变化或说话人多样性等不利条件下退化是否与自动模型性能下降和人类对情绪感知的变化相关。

7. 致谢

本工作部分得到瑞士 Stäfa 的 Sonova AG 的支持。

8. REFERENCES

[1] F. Y. Leung, V. Stojanovik, M. Micai, C. Jiang, and F. Liu, “Emotion recognition in autism spectrum disorder across age groups: A cross-sectional investigation of various visual and auditory communicative domains,” Autism Research, vol. 16, no. 4, pp. 783–801, 2023.

[2] C. Lima, T. Alves, S. Scott, and S. Castro, “In the ear of the beholder: How age shapes emotion processing in nonverbal vocalizations,” Emotion (Washington, D.C.), vol. 14, pp. 145–160, 11 2014.

[3] E. Coutinho and N. Dibben, “Psychoacoustic cues to emotion in speech prosody and music,” Cognition & emotion, vol. 27, 10 2012.

[4] E. Parada-Cabaleiro, A. Batliner, M. Schmitt, M. Schedl, G. Costantini, and B. Schuller, “Perception and classification of emotions in nonsense speech: Humans versus machines,” PLoS One, vol. 18, p. e0281079, Jan. 2023.

[5] J. Koemans, "Comparing cross-lingual automatic and human emotion recognition in background noise," Master's thesis, Radboud University, XXX, 2020.

[6] E. Parada-Cabaleiro, M. Schmitt, A. Batliner, S. Hantke, G. Costantini, K. Scherer, and B. Schuller, "Identifying emotions in opera singing: Implications of adverse acoustic conditions," 09 2018.

[7] S. Mozziconacci, "Speech variability and emotion : production and perception," Transport Logistics - Transport Logist, 01 1998.

[8] M. Schroder, "Expressing degree of activation in synthetic speech," Audio, Speech, and Language Processing, IEEE Transactions on, vol. 14, pp. 1128 – 1136, 08 2006.

[9] X. Luo, Q. J. Fu, and J. J. Galvin, "Vocal emotion recognition by normal-hearing listeners and cochlear implant users [published correction appears in trends amplif," Trends Amplif, vol. 11, no. 3, pp. 301–315, 2007.

[10] K. Hammerschmidt and U. Jürgens, "Acoustical correlates of affective prosody," J. Voice, vol. 21, pp. 531–540, Sept. 2007.

[11] E. Rodero, "Intonation and emotion: influence of pitch levels and contour type on creating emotions," J. Voice, vol. 25, pp. e25–34, Jan. 2011.

[12] S. Mozziconacci and D. Hermes, "Expression of emotion and attitude through temporal speech variations," 11 2001.

[13] M. Guzman, S. Correa, D. Muñoz, and R. Mayerhoff, "Influence on spectral energy distribution of emotional expression," Journal of Voice, vol. 27, no. 1, pp. 129.e1–129.e10, 2013.

[14] C.-H. Wu, J.-F. Yeh, and Z.-J. Chuang, Emotion Perception and Recognition from Speech, pp. 93–110. 01 2009.

[15] O. Scharenborg, S. Kakouros, and J. Koemans, "The effect of noise on emotion perception in an unknown language," pp. 364–368, 06 2018.

[16] E. Parada-Cabaleiro, A. Baird, A. Batliner, S. Hantke, and B. Schuller, "The perception of emotions in noisified nonsense speech," pp. 3246–3250, 08 2017.

[17] W. Thompson and L.-L. Balkwill, "Decoding speech prosody in five languages," Semiotica, vol. 2006, 01 2006.

[18] P. Tsiakoulis, A. Potamianos, and D. Dimitriadis, "Spectral moment features augmented by low order cepstral coefficients for robust asr," IEEE Signal Processing Letters, vol. 17, no. 6, pp. 551–554, 2010.

[19] M. Pell, L. Monetta, S. Paulmann, and S. Kotz, "Recognizing emotions in a foreign language," Journal of Nonverbal Behavior, vol. 33, pp. 107–120, 06 2009.

[20] J. Jeon, D. Le, R. Xia, and Y. Liu, "A preliminary study of cross-lingual emotion recognition from speech: Automatic classification versus human perception," Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 2837–2840, 01 2013.

[21] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - the munich versatile and fast Open-Source audio feature extractor," in Proc. ACM Multimedia (MM), ACM, pp. 1459–1462, 2010.

[22] P. Gournay, O. Lahaie, and R. Lefebvre,

"A canadian french emotional speech dataset," pp. 399–402, 06 2018.

[23] G. Costantini, I. Iaderola, A. Paoloni, and M. Todisco, "EMOVO corpus: an Italian emotional speech database," in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), (Reykjavik, Iceland), pp. 3501–3504, European Language Resources Association (ELRA), May 2014.

[24] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," vol. 5, pp. 1517–1520, 09 2005.

[25] E. Parada-Cabaleiro, G. Costantini, A. Batliner, M. Schmitt, and B. Schuller, "DEMoS: an Italian emotional speech corpus. Elicitation methods, machine learning, and perception," Feb. 2019.

[26] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020.

[27] T. Bänziger, M. Mortillaro, and K. R. Scherer, "Introducing the geneva multimodal expression corpus for experimental research on emotion perception," Emotion, vol. 12, pp. 1161–1179, Oct. 2012.

[28] V. N. Vapnik, The nature of statistical learning theory. New York, USA: Springer-Verlag, 2000.

[29] D. H. Hosmer and S. Lemeshow, Applied logistic regression. New Jersey, USA: John Wiley & Sons, Inc., 2000.

[30] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," in Advances in Neural Information Processing Systems, vol. 20, 2007.

[31] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "Lightgbm: a highly efficient gradient boosting decision tree," in Proc. International Conference on Neural Information Processing Systems, p. 3149 – 3157, 2017.

[32] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986.

[33] F. Eyben, Real-time speech and music classification by large audio feature space extraction. Springer, 2015.