超越言语的情绪：基于边缘计算的非语音音频情绪识别

易卜拉欣·马利克[1], 西迪克·拉蒂夫[2], 萨纳乌拉·曼佐尔[3],穆罕默德·乌萨马[4],朱奈德·卡迪尔[5],和拉贾·朱尔达克[2]

[1]EmulationAI[2]澳大利亚布里斯班昆士兰科技大学（QUT）[3]英国西苏格兰大学[4]巴基斯坦国家计算机与新兴科学大学（NUCES）[5]卡塔尔多哈卡塔尔大学

arXiv:2305.00725v1 [cs.SD] 1 May 2023

摘要——**非语音情绪识别具有广泛的应用场景，包括医疗保健、犯罪防控、救援和娱乐等。利用边缘计算实现这些应用具有巨大潜力。然而，近期研究主要聚焦于使用复杂架构的语音情绪识别。本文提出了一种基于非语音的情绪识别系统，可依托边缘计算分析通过尖叫、哭泣等非语音表达传达的情绪。具体而言，我们探索了知识蒸馏，以设计一种计算高效的系统，使其能够部署在资源有限的边缘设备上，同时不会显著降低性能。我们使用两个公开可用数据集对所提出的框架进行了全面评估，并通过将结果与广为人知的 MobileNet 模型进行比较，突出了其有效性。结果表明，利用边缘计算进行非语音情绪检测具有可行性和有效性，并有望改进依赖通信网络中情绪检测的应用。据我们所知，这是首个用于检测非语音音频中情绪的边缘计算框架，为未来研究提供了有前景的方向。**

索引词—**非语音情绪识别、边缘计算、知识蒸馏和计算效率。**

I. 引言

THE 物联网（IoT）时代已经到来。物联网设备数量的激增，以及科技制造商之间争夺市场份额的竞争，已经发展到通信系统难以满足服务质量和用户体验要求的程度。近年来，人工智能（AI）与物联网的融合催生了大量实际应用。这些应用涵盖广泛的领域，从图像分类到稳定扩散 [1] ，从语音识别到实时语音生成 [2], [3]。医疗保健行业在疾病检测方面取得了显著进展，人工智能在疾病早期检测方面的表现已经超过人类医生 [4]。数据收集与清理以及城市计算 [5],也受益于这种结合。此外，语音助手和自适应情绪识别 [6] 只是人工智能与物联网融合所催生的众多其他应用中的一小部分。这些发展对数据存储和计算能力提出了前所未有的要求。传统的通信系统设计已不足以满足这些数据密集型和计算密集型应用的需求。这促成了云计算 [7],的兴起；云计算是人工智能赋能物联网应用的支柱，而物联网应用的广泛采用也得益于云计算技术。

边缘计算是一种分布式计算范式，通过将企业应用部署在靠近数据源的位置（例如物联网设备或边缘服务器），降低向云端传输数据的负载。靠近数据源能够带来显著的业务优势，包括更快的响应时间、更充足的带宽、更迅速的决策以及更好的隐私保护。图形处理器、张量处理器等计算技术的发展，使得将部分计算任务卸载到性能强大的边缘服务器成为可能。对于实时服务/应用（例如交通监控系统、人脸识别和控制系统应用），延迟、服务质量和用户体验变得愈发重要。尤其是在情绪识别应用中，低延迟至关重要，因为计算设备需要根据给定的音频或视觉输入数据，针对特定应用对用户的情绪状态进行分类。在危及生命的紧急情况下，实时情绪识别更是至关重要。在这类场景中，边缘计算有望满足延迟要求。本文提出了一种基于边缘计算的非语音情绪检测系统。

情绪识别系统日益受到关注，其性能也因采用基于尖端深度学习的面部、语音、语言和心理信号模型而大幅提升。如图所示，大多数情绪感知服务采用这样的系统范式：通过物联网传感器采集的原始数据被传输至远程服务器进行处理和决策

电子邮箱：siddique.latif@qut.edu.au

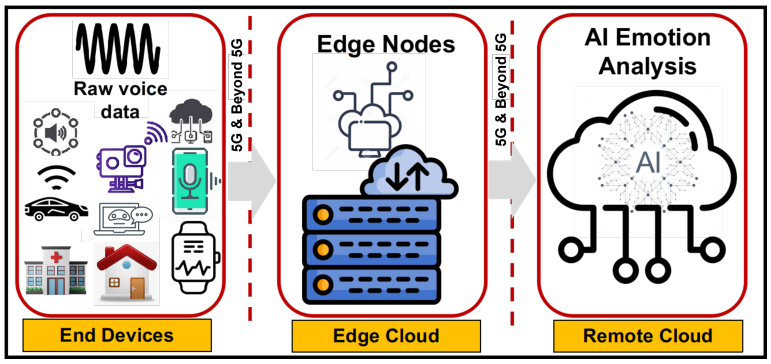

图 1：最先进的非语音情绪感知系统通过通信网络传输原始语音信号，以进行情绪分析。

见图 1。由于情绪识别系统旨在实时检测和分类情绪，因此，构建一个端到端延迟处于可接受水平的系统至关重要，其延迟范围从数据采集到情绪分类。

现有关于基于语音的情绪检测的研究主要集中于提高系统的准确率，以实现其实时应用 [8]。在这些系统中，通常将音频对话输入不同的深度学习模型，以预测不同的情绪 [9]。这些系统所使用的音频对话包含脚本化语音数据集。然而，情绪并不总是存在于语音中。尖叫声等非语音信号同样包含丰富的情绪信息。及时识别尖叫声在公共场所、医疗保健[10],老年护理、救援服务、犯罪防控和游戏等领域具有广泛的应用价值。分类延迟可能导致人员死亡，因此延迟问题变得更加令人担忧。此外，最先进的情绪感知应用遵循这样一种系统模型：将原始语音传输至远程服务器进行处理和决策。这类系统在现实场景中取得了成功，但它们需要通过通信网络完整地共享语音数据，这可能对人们的隐私造成不利影响 [11], [12]。边缘计算通过在联邦环境下利用边缘服务器处理数据，解决了延迟和隐私相关的问题 [6]。

大多数情绪感知服务都遵循这样的系统模型：将原始语音传输到远程服务器进行处理和决策。如图 1 所示。这类系统在现实生活中取得了成功，但它们需要通过通信网络完整地共享语音，这可能会对人们的隐私 [12], [13]造成不利影响。语音信号包含有关消息、说话人、性别、语言等的敏感信息，窃听对手可能会在未经用户同意的情况下滥用这些信息 [11]。

本文提出了一种面向物联网的边缘计算非语音情绪识别系统框架。我们的贡献如下：

1）我们提出利用边缘计算，为资源受限设备设计一种低延迟的基于非言语声音的情绪识别系统。2）我们利用知识蒸馏开发了一种计算高效的非言语情绪检测系统。3）我们详细讨论了将非言语情绪应用于医疗保健、救援服务等各种场景的潜力。4）我们使用两个公开可用数据集对该系统进行评估，以展示所提出框架的有效性。结果表明，与广为人知的 MobileNetV3-small 模型相比，我们提出的模型能够实现更好的性能 [14] 并提供更高的计算效率。

II. 尖叫识别的应用

在本节中，我们将简要介绍文献中现有的尖叫检测系统。这里的目的在于列出一些基于非语音情绪识别的系统相关研究，但并非详尽无遗。

A. 医疗保健与救援服务

理解非语言情绪是医疗保健研究中一个日益受到关注的领域。在最近的疫情期间，许多研究人员通过理解咳嗽声，研究了对 COVID- 19 进行早期预测的可能性。同样，许多老年患者护理研究人员通过检测和解读咳嗽声与尖叫声，试图了解患者需求 [15], [16]。心理学家也在努力检测并理解非语言情绪活动，以识别各种精神障碍。近年来，我们还见证了面向病患和老年人的辅助技术显著发展。其中一些技术采用腕带/手环和置于人体表面及体内的轻型传感器形式。这些辅助技术将尖端物联网技术与广泛的学习算法相结合，从而识别人类的非语音情绪及其他人类能力。Alam 等人 [16]设计了一款用于痴呆症患者尖叫检测的便携式手环。该手环价格低廉且易于使用（即 HealthBand），可用于监测痴呆症患者的活动以及相关人员在各种困难情况下的警觉性。

非言语语音检测的另一项应用是救援服务，且与医疗保健有部分关联。在地震、野火等灾难中，尖叫检测技术对于定位受害者（人类或动物）发挥着至关重要的作用。由于救援相关行动对时间十分敏感且需要保持警觉，使用人工智能/机器学习技术进行尖叫检测，有助于在瓦砾下和燃烧现场救出受困受害者。Saeed 等人 [17]设计了一种搭载于小型自动驾驶车辆上的人工智能/机器学习尖叫检测系统，可帮助救援燃烧现场的受害者。该系统中的尖叫检测模型基于支持向量机（SVM）和长短期记忆网络（LSTM）。由于职业本身具有动态性，流动工作者始终面临受伤风险；在发生紧急情况时，尖叫检测器可以协助救援人员 [17], [18]。

B. 犯罪防控

随着人工智能和先进通信技术的出现，犯罪检测正成为一个蓬勃发展的研究方向。尖叫检测与暴力犯罪直接相关，而利用人工智能/机器学习技术，并结合部署在城市空间各处的多个传感器所收集的数据，是一项颇具意义的应用。Laffitte 等人 [19], [20] 提出了一种基于机器学习的尖叫/喊叫检测机制。Marteau 等人 [21] 提出了基于深度学习的方法，用于识别尖叫声、玻璃破碎声、枪声和喷雾声等音频事件。不幸的是，种族主义、骚扰和强奸犯罪在社会中日益增多 [22]–[24]。城市空间对女性、跨性别者及其他性别群体而言正变得越来越不安全。将尖叫检测与其他监控技术结合使用，有助于保护人们免受这些犯罪侵害。2018年，韩国首尔地铁在地铁站的女性洗手间安装了尖叫检测器，以保障女性安全[1]。同样，巴黎地铁公司也在考虑在地铁中采用支持人工智能的尖叫检测技术，以防止异常情况发生 [19], [25]。我们坚信，尖叫检测技术有助于减少大量犯罪。

C. 家庭应用

尖叫检测正逐渐成为家庭、安防应用、养老院等场景中视觉监控的重要辅助工具。例如，Huang 等人 [26] 提出了一种基于能量连续性的方法，用于从家庭音频录音中提取特征，随后使用支持向量机（SVM）识别录音数据中的尖叫声。O'Donovan 等人 [27] 提出并评估了一种基于机器学习的方法，用于在公开可用的家庭语音录音数据集中检测尖叫（行为障碍）。他们使用预训练卷积神经网络，从音频数据集中学习尖叫检测。为进行验证，他们选择了著名电视节目《超级保姆》中的数据集，因为它与临床数据相似。这些结果表明，利用公共数据集学习行为障碍（尖叫和发脾气），再将其用于临床录音，比收集昂贵的私有敏感临床数据语料库来训练行为障碍检测模型更为合适。

近年来，家庭暴力和暴力关系显著增加 [28]。尖叫检测技术与物联网语音辅助技术相结合，可以识别这些令人发指的罪行，并有望保护许多人免受伤害。Fleury等 [29]使用放置在公寓内的八个麦克风录音，并运用语音识别算法来确定人类语音的各种要素。这种自主语音检测的理念可以进一步扩展，用于识别尖叫声，并帮助大量遭受家庭暴力和虐待性关系伤害的人 [30]。尽管在技术上可行，但开发这些安全与监控应用仍面临重大的伦理和隐私问题。

D. 游戏应用

尖叫声是语音的重要组成部分，理解尖叫声所关联的情绪对于语音检测和翻译系统至关重要。传统的语音对话数据集并未包含足够多的尖叫声，无法用于充分学习和研究尖叫声。Mori等 [31] 使用战斗游戏录制了包含更多尖叫声样本的对话语料库，并用该语料库分析社交尖叫声的特性。虚拟现实和基于增强现实的游戏日益流行，可用于培训救援人员、应急响应人员、普通人和儿童应对紧急情况。尖叫检测系统有望在应对紧急情况的游戏化训练中发挥重要作用[32]。

III. 提出的系统

在本节中，我们将讨论所提出的基于边缘的非语音情绪识别系统的详细内容。本研究提出的框架（图2）在5G网络架构内提供数据收集和分析支持，根据3GPP标准，该架构通常被称为网络数据分析框架（NWDA）。基本上，所提出的框架可以调用NWDA功能，以提供以下两项核心功能：首先，从网络功能（NFs）收集数据，即进行本地数据处理；其次，进行数据分析和非语音情绪识别。具体而言，所提出的框架包含以下关键层：感知层、边缘计算层和决策层。

A. 非语音信号感知层

为了实现基于边缘的非语音情绪识别，我们设计了一个专用的语音感知层。非语音信号感知层使终端用户设备能够从信息物理空间采集非语音数据。如今，智能边缘设备已大量涌现，涵盖

[1]首尔地铁在女性洗手间安装尖叫检测系统，《韩国先驱报》。https://www.koreaherald.com/view.php? ud=20180323000689。访问日期：2023年4月15日。

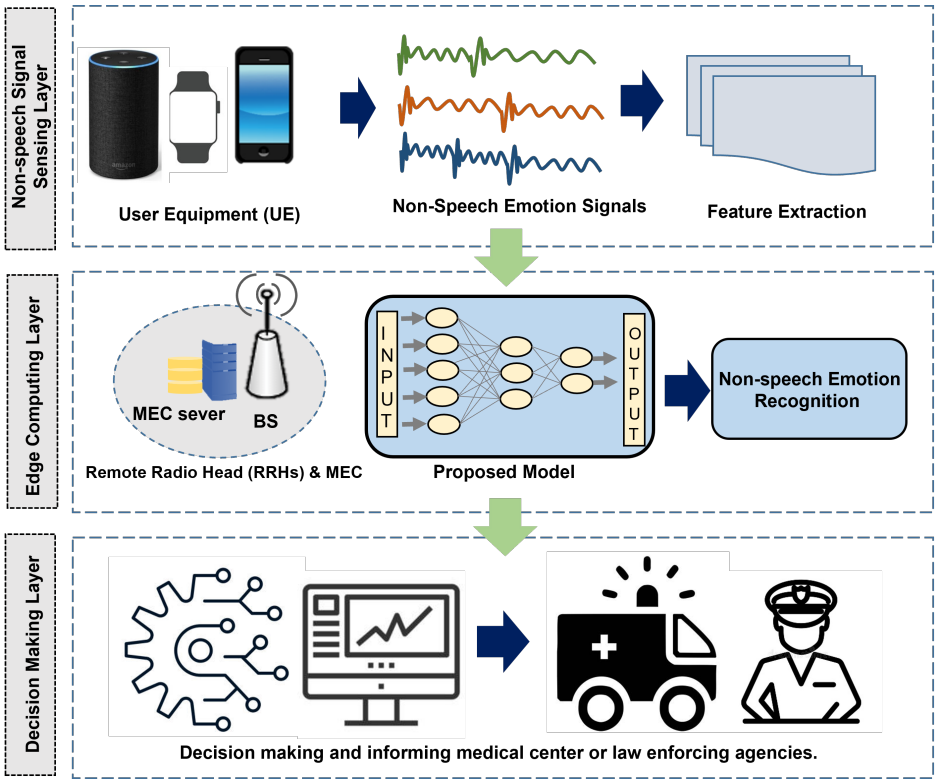

图 2：所提出系统的组成部分：(1) 非语音信号感知层，用于从手机、个人助理或智能手表收集语音数据，并将其转换为语音特征；(2) 边缘计算层，利用这些特征训练用于非语音情绪识别的深度学习模型；(3) 决策层，用于分析非语音情绪，并根据其积极或消极倾向作出决策，还可能向救护车或警察发送警报。

智能手机、智能手表以及 Alexa Home、Google Assistant、Siri 等个人助理。这些边缘设备，尤其是个人语音助理，会接收语音输入并执行特定操作，例如开灯、关闭电器或播放点播音乐。在大多数此类应用中，输入语音会被转换为文本，再通过自然语言处理确定用户意图以及需要执行的具体操作。在我们提出的框架中，非语音信号感知层接收原始语音信号，并进行必要的预处理以减轻背景噪声的影响。该层处理输入信号并将其转换为特征，即梅尔频谱图。如图 2 所示，在非语音信号感知层中，音频特征会被传输至网络，而不是将原始信号直接发送到云服务器。

B. 边缘计算层

在我们提出的模型中，系统分为边缘层和核心层。边缘层由终端设备（即手机、平板电脑等）以及部署在基站（BS）附近的边缘服务器组成，后者作为移动边缘计算（MEC）服务器 [33]。假设 MEC 服务器能够处理边缘信号，也能够对非语音情感数据进行分析 [34]。同时假设边缘服务器具有足够的计算资源来执行数据密集型机器学习任务。所提出的框架采用第三代合作伙伴计划支持基于机器学习的数据驱动优化的 3GPP 第 16 版 [35]。

边缘服务器是我们所提出架构的重要组成部分。边缘服务器利用低功耗的计算和存储硬件，例如边缘云端小站 [36]，并部署在靠近终端用户的无线接入网络（RAN）中 [33]。在我们提出的架构中（如图 2 所示），边缘服务器不仅执行流量聚合网关和网络服务控制功能，还充当智能边缘服务器，负责根据给定的语音特征识别尖叫声。

在边缘计算设备上部署深度学习模型是一个活跃的研究领域，已有许多技术被提出，用于提升这些模型的延迟表现和性能。一些重要技术包括剪枝、量化、知识蒸馏，以及训练计算效率高的模型。在我们的实验中，我们使用知识蒸馏 [37]来训练一个小巧高效的模型，使其能够部署在边缘设备上。

1）知识蒸馏：顾名思义，知识蒸馏是指将知识从计算开销较大的大型模型迁移到相对较小模型中的方法。较大的模型和较小的模型分别称为教师模型和学生模型。因此，知识蒸馏包含三个主要组成部分：（1）知识；（2）蒸馏算法；（3）教师模型与学生模型架构。尽管如今已有多种蒸馏算法，我们选择了

基于响应的方法。如图3所示，其假设是学生模型将学会模仿教师模型的预测结果。这可以通过使用一种损失函数来实现，该函数称为蒸馏损失，用于捕捉学生模型与教师模型各自的 Logits 之间的差异。

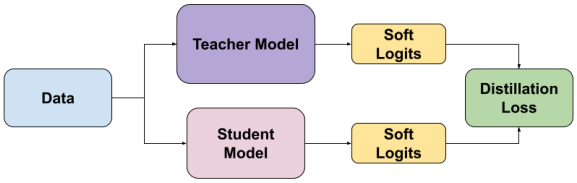

图 3：基于响应的知识蒸馏。学生模型和教师模型的输出 Logits 用于计算学生与教师之间的蒸馏损失。

由于这种损失可以减少过度训练，学生模型将更善于做出与教师相同的预测。在离线训练方案中，首先训练教师模型，然后冻结其权重。接下来，使用蒸馏损失和教师模型的 Logits 作为目标来训练学生模型。以下是蒸馏损失的公式。

$$L_d = \alpha T^2 \cdot KL\Big(\text{softmax}(T^{-1} \cdot f(T,x)), \text{softmax}(T^{-1} \cdot g(T,x))\Big),$$

其中：

L_d：用于知识蒸馏的损失函数α：控制分类损失与蒸馏损失之间权衡的超参数

T：用于软化教师模型和学生模型 Logits（Softmax 之前最后一层的输出）的温度超参数

KL：KL散度，用于衡量两个概率分布之间的差异

Softmax：将 Logits 转换为概率的函数

$f(T,\ x)$：输入 x时教师模型的 Logits；$g(T,\ x)$：输入 x时学生模型的 Logits

2）教师模型：通常会选择规模更大、层数更深的网络作为教师模型，使其在当前任务中表现出色。我们选择 ResNet18 [38]作为教师模型。ResNet18 包含 18 个堆叠在一起的残差块，可缓解网络退化和梯度消失问题。图4展示了单个残差层，其中前一层的输出会与下一层的输出相加；图5则展示了教师模型所用 ResNet18 的完整架构。为确保相加前尺寸一致，输入可能需要经过某种操作，使其与输出维度对齐。通常采用卷积来完成这一操作。

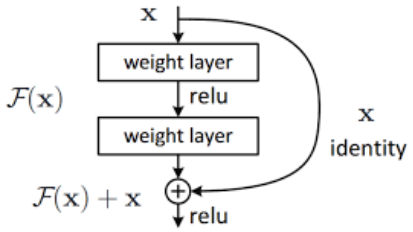

图4：残差层。输入的权重与后续卷积层的输出相加。

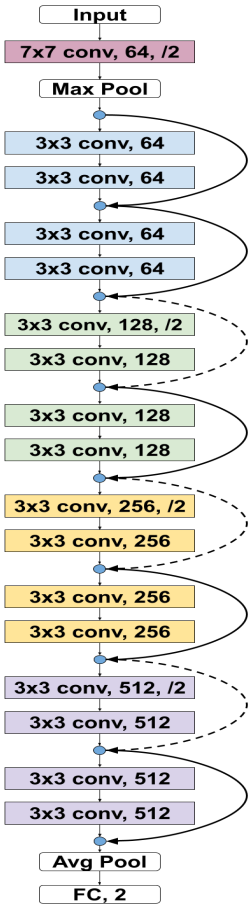

图5：ResNet18/教师模型架构。彩色方块表示卷积层及其对应的卷积核尺寸、输出滤波器和输入尺寸缩减。卷积层之间的连线表示残差连接；虚线则表示残差输入会经过卷积，以确保维度一致性。最后的全连接层是稠密层。

3）学生模型：与教师模型不同，学生模型规模更小、层数更浅，因此计算效率更高。我们提出的学生网络仅由3个卷积层和随后连接的3个全连接层组成。图6详细展示了这一相对浅层的学生模型。第一个卷积层的卷积核大小为 7×7，其余两个卷积层的卷积核大小均为 5×5。各层的滤波器数量分别为6、16和32，并且在每个卷积层之后，我们都应用一个窗口大小为 2×2 的最大池化层。全连接层的输出依次为：128、64、2。对于正则化，

我们在每个卷积层和全连接层之后都添加一个丢弃层，丢弃概率为 10%。层与层之间选用的非线性激活函数为修正线性单元，通常简称为 ReLU。

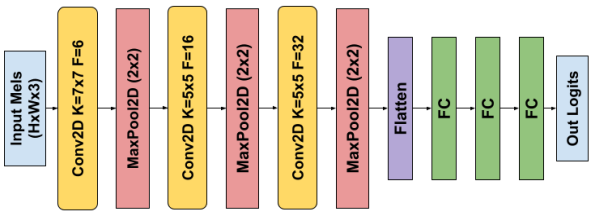

图 6：学生模型。每个彩色模块代表一个卷积层，其中 K 表示卷积核大小， F 表示输出滤波器的数量。最后的全连接层为稠密层。

C. 决策层

我们在边缘设备上训练尖叫检测模型。边缘设备通过蜂窝基础设施与云服务器通信，并共享模型结果。云服务器负责尖叫分析、决策和存储服务。我们将所提出的分类器部署到边缘设备上，以执行识别任务。模型的输出会发送到决策层，由决策层根据具体情况采取必要行动。例如，如果尖叫声所表达的情绪被分类为负面情绪，即此人正遭受疼痛或悲伤，云系统就会向医疗中心或警察发送警报，因为此人可能受伤或被他人伤害。如果尖叫声所表达的情绪被分类为正面情绪，例如喜悦的尖叫声，云系统则不会发送任何警报。

IV. 实验设置

本节预先介绍数据集、输入特征和训练方案的详细信息。

A. 实验中使用的数据集

- *ASVP-ESD*： 音频、语音与视觉处理实验室情感声音数据库（ASVP-ESD）是一个包含语音和非语音话语的数据集。该数据集共有 12625 个音频样本，采集自各种来源。样本包括男性和女性说话人，情绪包括无聊（叹气、打哈欠）、中性、快乐（笑、咯咯笑）、悲伤（哭泣）、愤怒、恐惧（尖叫、恐慌）、惊奇（惊奇、倒吸一口气）、厌恶（轻蔑）、兴奋（胜利、欢欣）、愉悦（渴望）、疼痛（呻吟）和失望。

- *VIVAE*： 不同强度情感与情绪发声语料库（VIVAE）数据集[39] 由人类非言语情绪发声组成。完整数据集共包含来自 11 位说话人的 1085 个样本。这些发声被划分为三种积极情绪状态（成就/胜利、性愉悦和惊讶）和三种消极情绪状态（愤怒、恐惧和身体疼痛）。音频采样率为 44.1 kHz。

- *DEMAND*： 我们使用此数据集评估所提出的框架在噪声条件下的性能。多样化环境多通道声学噪声数据库（DEMAND）数据集 [40] 提供了录音，可用于评估算法在各种真实环境中采集的真实噪声下的表现。该数据集涵盖 6 个类别，其中 4 个类别属于室内环境，其余类别的样本采集自室外环境。数据集录音提供 48 kHz 和 16 kHz 两种采样率。音频最初录制了较长时间，之后经过剪辑，每段时长共计 300 秒。

B. 输入特征

在语音和音频研究中，梅尔频谱图是一种表示输入信号 [41] 的常用方法。同样地，我们选择使用短时傅里叶变换将音频样本表示为梅尔频谱图，变换大小为 1024，跳跃大小为 256，窗口大小为 1024。频率范围设置为 0–8kHz，共计算 128 个频带。每个梅尔频谱图都在 $[-1, 1]$ 范围内进行了归一化。由于样本话语的长度并不一致，我们将较长音频的截断长度设为 3 秒，并对较短音频使用零进行填充，从而得到长度一致的 3 秒音频。在将较高采样率的音频转换为梅尔频谱图之前，我们先将其重采样至 16kHz。整个实验和数据集中均保持使用这一采样率。

C. 训练方案

分类任务的训练使用配备 24 GB 显存的 Nvidia GeForce RTX 3090 GPU 完成，并选择 PyTorch 作为框架。模型以 64 的批量大小进行训练，采用二元交叉熵损失作为判据。每一层的权重均随机初始化，并使用 Adam 作为优化器，参数如下： β_1 =0.9, $\beta_2 = 0.9999$, $\epsilon = 1e^{-8}$。我们尝试了多种学习率，发现学习率为 $1e^{-5}$ 时效果更好，且训练时间更短。

所有实验均采用随机划分的数据集，其中 80% 用于训练，20% 用于测试。对于在 ASVP- ESD 数据集上进行的尖叫检测任务，由于尖叫话语总计有 1170 个样本，我们需要平衡尖叫话语和非尖叫话语的数量。为平衡数据集，我们随机选取了 1170 个非尖叫话语，从而得到 2340 个样本用于训练尖叫检测器。

在实验过程中，我们注意到模型会出现过拟合，导致训练准确率高而测试准确率低。这种高偏差可能是由数据集规模较小造成的。为缓解这一问题，我们对训练数据进行了数据增强，包括拉伸和压缩音频样本，以及添加低幅度的高斯噪声。此外，我们还随机遮蔽计算得到的梅尔频谱图的时间轴和频率轴。这种增强方案有助于提升模型的泛化能力和训练效果。

表 I：未添加噪声的实验分类结果。

模型	Task	准确率
教师	尖叫检测	85.47
	尖叫类型分类	73.80
MobileNetV3s	尖叫检测	76.92
	尖叫类型分类	65.80
学生模型（所提出的）	尖叫检测	**80.58**
	尖叫类型分类	**67.16**

V. 结果与讨论

该系统旨在利用边缘计算检测通信网络中的非语音情绪。我们主要聚焦于两类实验：尖叫检测和尖叫情绪检测。前者将尖叫声片段与非尖叫声片段区分开来，后者则判定一个人的尖叫是否发生在危险或受胁迫的情境中。这样做是因为并非所有尖叫声都需要调查或意味着发生了紧急情况。因此，将尖叫声归为积极和消极两类十分重要，以便向用户通知系统检测到的尖叫声类型。在每项实验中，我们都会比较教师、学生模型和 MobileNetV3-small 模型 [14]，——后者是边缘计算领域的热门选择。为简洁起见，在下文中我们将 MobileNetV3-small 简称为 MobileNetV3s。

A. 非言语情绪检测

本节介绍尖叫检测和尖叫类型分类任务的结果。前者在 ASVP-ESD 数据集上进行，后者在 VIVAE 数据集上进行。实验结果见表 I。可以看出，模型能够很好地区分尖叫样本和非尖叫样本。然而，模型难以将话语聚类为积极和消极类别。

我们可以通过使用 t-SNE 算法 [42]进行降维，观察样本可能形成的聚类。图 7 展示了使用数据集中的原始梅尔频谱图，以及教师和学生模型倒数第二层激活所绘制的 t 分布随机邻域嵌入（tSNE）图。图中显示，原始梅尔频谱图几乎没有形成聚类。在模型经过尖叫分类任务训练后，情况发生了变化。可以观察到，在尖叫检测任务中，教师模型和学生模型都形成了清晰的聚类。教师模型和学生模型在尖叫类型分类任务中的聚类较为稀疏，但仍优于原始梅尔频谱图的结果。这种稀疏性或许可以解释该任务的复杂性。

B. 噪声条件下的评估

在现实场景中，背景通常并非静态不变，因此环境中会包含噪声。为了模拟现实场景，我们将 DEMAND 数据集中的噪声添加到音频样本中，并在模拟的嘈杂现实环境中进行了评估测试。这些噪声样本随机选自以下环境：公交车、地铁、咖啡馆、厨房和办公室。在每个噪声样本中，我们随机截取了与输入音频样本等长的一段音频。表 II 总结了该实验的结果。为了测试模型性能的稳健性，我们仅在测试集划分中添加了噪声。结果表明，即使仅在评估阶段加入噪声，该模型仍具有良好的泛化能力。

表 II：噪声环境下实验的分类结果。

模型	Task	准确率
教师	尖叫检测	82.26
	尖叫类型分类	69.74
MobileNetV3s	尖叫检测	76.92
	尖叫类型分类	60.15
学生模型（所提出的）	尖叫检测	**78.60**
	尖叫类型分类	**66.42**

为了进一步测试学生模型在嘈杂环境下的性能，我们比较了之前使用的每种噪声类别对应的评估结果。表 III 显示，针对单种噪声的结果并未大幅下降，整体表现与前文讨论的结果相近。

表 III：学生模型在各单独噪声类别下的分类结果。

Task	噪声类型	准确率
尖叫检测	Cafe	72.86
	厨房	78.85
	办公室	80.13
	地铁	73.50
尖叫类型分类	Cafe	65.68
	厨房	63.10
	办公室	63.84
	地铁	65.31

C. 计算复杂度

本节所列基准测试在一块 UpBoard（如图 8所示）上进行，该设备配备运行频率为 1.60GHz 的 Intel(R) Atom E3940 CPU，板载内存为 4GB。测试在标准版 Ubuntu 20.04 LTS 操作系统（未使用服务器/无头版本）和 PyTorch 1.13 上完成。结果未使用 PyTorch 提供的后处理功能。这些功能据称可以缩短推理时间，但通常需要作出一些权衡。

在生产环境中部署模型时，模型加载总时间和单次话语前向传播所需的时间是最受关注的两个指标。同样，我们也对教师、学生模型和 MobileNetV3s 模型的这些指标进行了基准测试。图9以图形方式比较了上述模型的基准测试结果，显示在延迟方面，学生模型的速度几乎是教师模型的两倍，比 MobileNetV3s 模型快约 20 毫秒。此外，在加载时间方面也存在显著差异，学生模型加载到内存所需的时间最短。表IV中列出的各模型参数总量和模型大小，或许能更好地解释这些指标之间的差异。

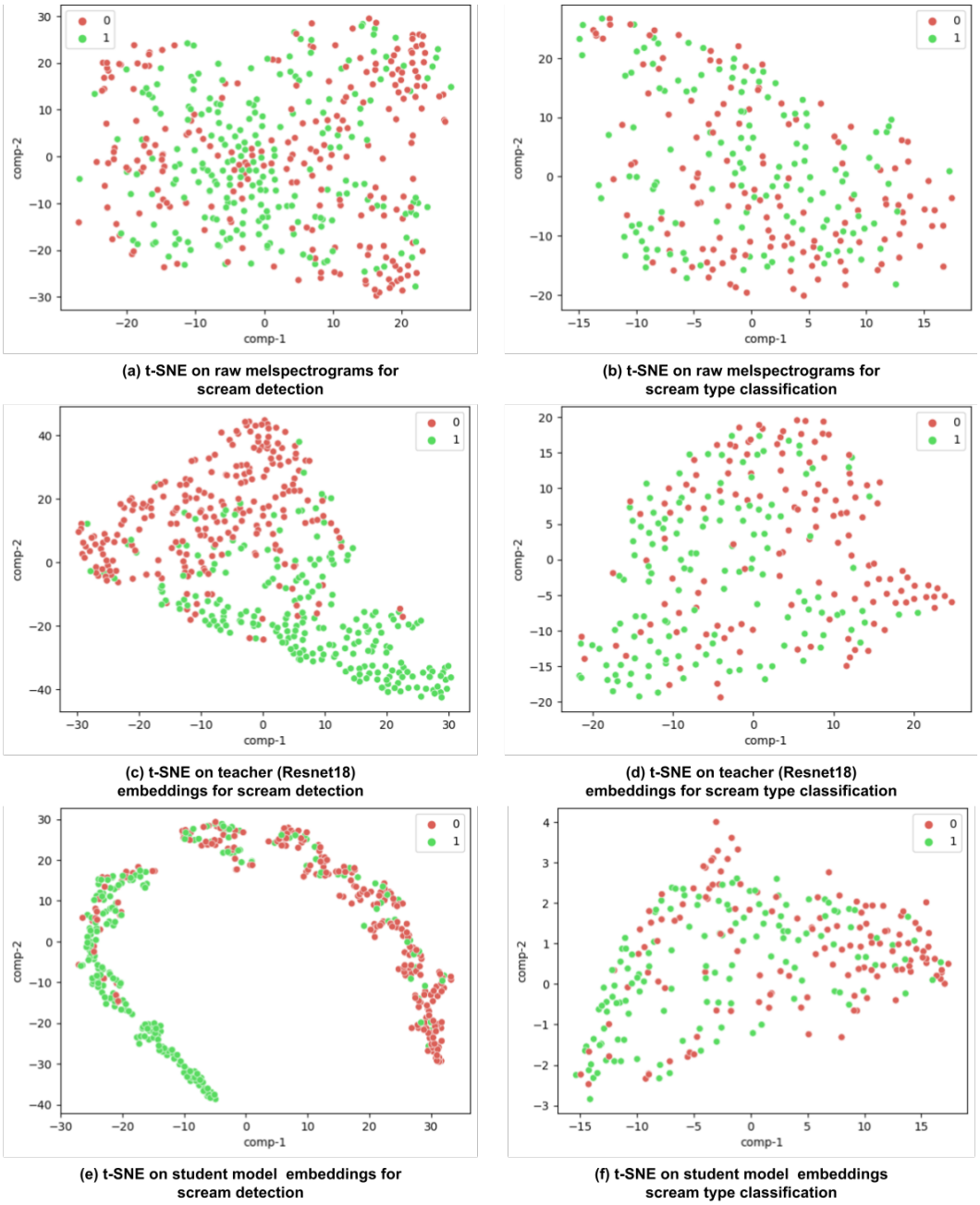

图 7：原始梅尔频谱图以及教师和学生模型倒数第二层激活（嵌入）的 t-SNE 图。原始梅尔频谱图中没有聚类现象；然而，使用训练好的模型后，t-SNE 图显示样本形成了聚类。

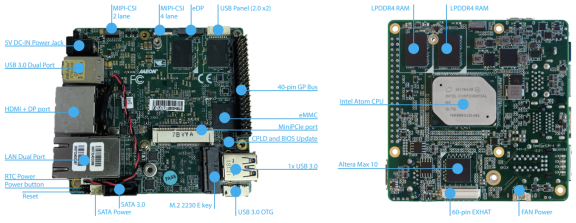

图 8：标注了 Intel(R) Atom E3940 CPU 的 UpBoard 示意图，来源 [43]。

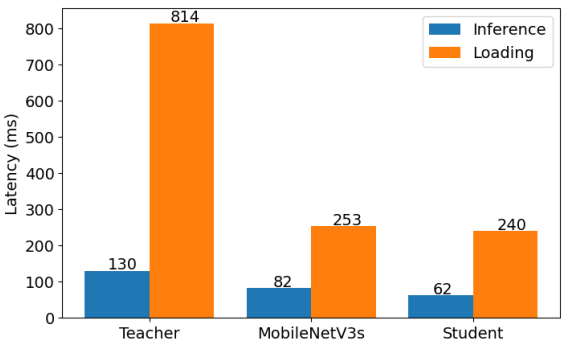

图 9：教师模型、学生模型（本文提出）和 MobileNetV3s 模型的延迟及加载时间（毫秒）对比。

模型中的参数数量越多，计算成本就越高。同样，较大的内存占用也会导致模型加载时间更长。

这些结果凸显了将所提出的系统部署到内存和计算资源受限的设备（如个人助理（Alexa Home, Google Assistant 和 Siri）。个人助理是

表IV：教师、学生模型和 MobileNetV3s 模型的大小及总参数量。

模型	总计参数	Size (MB)
教师	11177616	42.676
MobileNetV3s	1519906	5.844
学生模型（拟议）	**712778**	**2.719**

因此，该系统成为尖叫检测和分类任务的有效选择。根据尖叫检测结果是阳性还是阴性，这些个人助理可以向警察或医疗中心发送警报，以提供帮助。

VI. 结论

本文提出了一种基于知识蒸馏的边缘计算非语音情绪识别系统。我们介绍了非语音情绪识别的各种应用，并基于现实生活场景提供了一个案例研究。我们基于两个公开可用数据集评估了系统性能。我们设计了实验设置，以区分尖叫声与其他话语，并根据其情绪状态对非语音话语进行分类。为突出所提系统的稳健性，我们还通过向输入中加入典型的现实世界背景噪声来模拟现实世界场景，并对这些实验进行了评估。结果表明，与广为人知的 MobileNetV3 相比，所提框架具有更高的计算效率，并实现了更优的性能。这些结果展示了所提非语音情绪识别系统在通信网络中的可行性和有效性。未来，我们计划研究所提非语音情绪识别系统的能效。

参考文献

[1] F.-A. Croitoru、V. Hondru、R. T. Ionescu 和 M. Shah，“视觉中的扩散模型：综述”，*arXiv* 预印本 *arXiv:2209.04747*，2022年。[2] J. Li 等，“端到端自动语音识别的最新进展”，*APSIPA* 信号与信息处理汇刊，第 11 卷，第 1 期，2022年。[3] A. Wali、Z. Alamgir、S. Karim、A. Fawaz、M. B. Ali、M. Adan 和 M. Mujtaba，“用于语音处理的生成对抗网络：综述”，计算机语音与语言，第 72 卷，第 101308 页，2022年。[4] Z. Hu、J. Tang、Z. Wang、K. Zhang、L. Zhang 和 Q. Sun，“基于图像的癌症检测与诊断中的深度学习——综述”，模式识别，第 83 卷，第 134–149 页，2018年。[5] Y. Zheng、L. Capra、O. Wolfson 和 H. Yang，“城市计算：概念、方法论与应用”，*ACM* 智能系统与技术汇刊（*TIST*），第 5 卷，第 3 期，第 1–55 页，2014年。[6] S. Latif、H. S. Ali、M. Usama、R. Rana、B. Schuller 和 J. Qadir，“基于人工智能的情绪识别：前景、风险与促进亲社会发展的建议”，*arXiv* 预印本 *arXiv:2211.07290*，2022年。[7] T. Dillon、C. Wu 和 E. Chang，“云计算：问题与挑战”，载于*2010*年第*24*届*IEEE*高级信息网络与应用国际会议。IEEE，2010年，第 27–33 页。[8] S. Latif、R. Rana、S. Khalifa、R. Jurdak、J. Qadir 和 B. W. Schuller，“语音处理中的深度表征学习：挑战、最新进展与未来趋势”，*arXiv* 预印本 *arXiv:2001.00378*，2020年。[9] S. Latif、A. Zaidi、H. Cuayahuitl、F. Shamshad、M. Shoukat 和 J. Qadir，“语音处理中的 Transformer：综述”，*arXiv* 预印本 *arXiv:2303.11607*，2023年。

[10] S. Latif、J. Qadir、A. Qayyum、M. Usama 和 S. Younis，“医疗保健语音技术：机遇、挑战与最新进展”，IEEE 生物医学工程评论，2020年。[11] S. Latif、S. Khalifa、R. Rana 和 R. Jurdak，“用于语音情绪识别应用的联邦学习”，载于 2020 年第 19 届 ACM/IEEE 传感器网络信息处理国际会议（IPSN）。IEEE，2020年，第 341–342 页。[12] S. Manzoor、A. N. Mian、A. Zoha 和 M. A. Imran，“联邦学习赋能的面向移动性感知的雾无线接入网络主动内容卸载框架”，Future Generation Computer Systems，第 133 卷，319，2022年 307页。–。[13] M. Jaiswal 和 E. M. Provost，“用于情绪识别的隐私增强型多模态神经表征。”载于 AAAI，2020年，第 7985–7993 页。[14] A. Howard、M. Sandler、G. Chu、L. Chen、B. Chen、M. Tan、W. Wang、Y. Zhu、R. Pang、V. Vasudevan、Q. V. Le 和 H. Adam，“搜索 MobileNetV3”，CoRR，第 abs/1905.02244 卷，2019年。 [在线]。可访问：http://arxiv.org/ abs/1905.02244[15] A. Imran、I. Posokhova、H. N. Qureshi、U. Masood、M. S. Riaz、K. Ali、C. N. John、M. I. Hussain 和 M. Nabeel，“AI4COVID-19：通过应用程序利用咳嗽样本对新冠肺炎进行人工智能初步诊断”，Informatics in medicine unlocked，第 20 卷，页码 100378，2020年。[16] Z. Alam、H. Samin 和 O. B. Samin，“面向痴呆症患者的 HealthBand：跌倒与尖叫检测器及护理人员助手”，载于 Journal of Physics: Conference Series，第 976 卷，第 1 期。IOP Publishing，2018年，页码 012015。[17] F. S. Saeed、A. A. Bashit、V. Viswanathan 和 D. Valles，“针对火灾现场受害者尖叫检测的初步机器学习分析”，Applied Sciences，第 11 卷，第 18 期，页码 8425，2021年。[18] V. Kalbag 和 A. Lerch，“重金属音乐中的尖叫检测”，arXiv 预印本 arXiv:2205.05580，2022年。[19] P. Laffitte、Y. Wang、D. Sodoyer 和 L. Girin，“评估不同神经网络架构对公共交通中尖叫声和喊叫声检测的性能”，Expert systems with applications，第 117 卷，第 29–41 页，2019年。[20] P. Laffitte，“地铁中尖叫声和喊叫声的自动检测”，博士学位论文，Lille 1，2017年。[21] T. Marteau、S. Afanou、D. Sodoyer、S. Ambellouis 和 F. Boukour，“嘈杂嵌入式铁路环境中的音频事件检测”，载于 Dependable Computing-EDCC 2020 Workshops：AI4RAILS、DREAMS、DSOGRI、SERENE 2020，德国慕尼黑，2020年 9 月 7 日，论文集。Springer，2020年，第 20–32 页。[22] J. M. Le´on-P´erez、J. Escart´ın 和 G. Giorgi，“全球职场霸凌与骚扰的现状”，Concepts, approaches and methods，第 55–86 页，2021年。[23] N. Borumandnia、N. Khadembashi、M. Tabatabaei 和 H. Alavi Majd，“全球性暴力发生率：趋势分析”，BMC public health，第 20 卷，第 1–7 页，2020年。[24] H. Walia，《边界与规则：全球移民、资本主义与种族主义民族主义的兴起》。Haymarket Books，2021年。[25] P. Laffitte、D. Sodoyer、C. Tatkeu 和 L. Girin，“地铁列车中尖叫声和喊叫语音的自动检测深度神经网络”，载于 2016 IEEE 声学、语音与信号处理国际会议（ICASSP）。IEEE，2016年，第 6460–6464 页。[26] W. Huang、T. K. Chiew、H. Li、T. S. Kok 和 J. Biswas，“家庭应用中的尖叫检测”，载于 2010 年第 5 届 IEEE 工业电子与应用会议。IEEE，2010年，第 2115–2120 页。[27] R. O’Donovan、E. Sezgin、S. Bambach、E. Butter、S. Lin 等，“从家庭音频录音中检测尖叫声以识别发脾气：使用迁移机器学习的探索性研究”，JMIR Formative Research，第 4 卷，18279，2020年 6，no. p. e 。[28] J. Usta、H. Murr 和 R. El-Jarrah，“新冠肺炎封控与针对女性的暴力增加：理解疫情期间的家庭暴力”，Violence and gender，第 8 卷，第 3 期，第 133–139 页，2021年。[29] A. Fleury、N. Noury、M. Vacher、H. Glasson 和 J.-F. Seri，“ 医疗智能家居中的声音与语音检测和分类 ”，载于 2008 年 IEEE 工程医学与生物学学会第 30 届年度国际会议。IEEE，2008年，第 4644–4647 页。[30] A. Chen、Q. He、X. Wang 和 Y. Li，“基于声学场景分析的家庭安全监控”，载于 2017 年第 10 届图像与信号处理国际大会，

生物医学工程与信息学（*CISP-BMEI*）。IEEE，2017年，第 1–5 页。[31] H. Mori 和 Y. Kikuchi，“用于研究社交尖叫声的游戏语料库。”载于 *INTERSPEECH*，2020年，第 3132–3135 页。[32] A. Lochmannov´a，“利用虚拟现实为急救人员应对紧急情况做好准备”，载于 *INTED2023* 论文集。IATED，2023年，第 3358–3363 页。[33] Y. Liu、M. Peng、G. Shou、Y. Chen 和 S. Chen，“迈向边缘智能：面向 5G 和物联网的多接入边缘计算”，*IEEE* 物联网杂志，第 7 卷，第 8 期，第 6722–6747 页，2020年。[34] G. Muhammad 和 M. S. Hossain，“利用深度学习进行认知边缘计算的情绪识别”，*IEEE* 物联网杂志，第 8 卷，16 901，2021年 23，16 894 no. pp. – 。[35] 3rd Generation Partnership Project，“3GPP，《支持网络数据分析服务的 5G 系统架构增强》，TS 23.288，2019年，版本 16.0.0。 [在线]。可访问：http://www.3gpp.org/DynaReport/23288.htm”，2019年 3 月。[36] T. Taleb、K. Samdanis、B. Mada、H. Flinck、S. Dutta 和 D. Sabella，“多接入边缘计算：新兴 5G 网络边缘云架构与编排综述”，*IEEE* 通信调查与教程，第 19 卷，第 3 期，第 1657–1681 页，2017年。[37] G. Hinton、O. Vinyals 和 J. Dean，“提炼神经网络中的知识”，*arXiv* 预印本 *arXiv:1503.02531*，2015年。[38] K. He、X. Zhang、S. Ren 和 J. Sun，“用于图像识别的深度残差学习”，*CoRR*，第 abs/1512.03385 卷，2015年。[39] [ ] 1512 03385 在线。可访问：http://arxiv.org/abs/ 。N. Holz、P. Larrouy-Maestri 和 D. Poeppel，“不同强度情感与情绪发声语料库（VIVAE）”，2020年 10 月。 [在线]。可访问：https://doi.org/10.5281/zenodo。4066235[40] J. Thiemann、N. Ito 和 E. Vincent，“DEMAND：多样环境中的声学噪声多通道录音集合”，2013年 6 月，由 Inria 在 VERSAMUS 联合团队计划支持。 [在线]。可访问：https://doi.org/10 5281 1227121 。/zenodo。S. Latif、R. Rana、S. Khalifa、R. Jurdak、J. Qadir 和 B. W. Schuller，“语音情绪识别的深度表征学习综述”，*IEEE* 情感计算汇刊，2021年。[42] L. van der Maaten 和 G. Hinton，“使用 t-SNE 可视化数据”，机器学习研究杂志，第 9 卷，第 86 期，第 2579–2605 页，2008年。 [在线]。可访问：http://jmlr.org/ papers/v9/vandermaaten08a.html [43] “Up2 规格，[在线]。可访问：https://up-board.org/wp-content/uploads/2021/06/up-squared-datasheet-v2.pdf。”