arXiv:2304.03295v1 [cs.SD] 6 Apr 2023

# 通过耳戴式传感进行音乐反应自动检测

李义赫，韩国技术教育大学，大韩民国
Chulhong Min，诺基亚贝尔实验室，英国李宰承，
韩国技术教育大学，大韩民国柳真，韩国技术教育大
学，大韩民国姜胜宇， 韩国技术教育大学，大韩民国

我们提出了 GrooveMeter，这是一种新颖的系统，可通过耳戴式感测自动检测发声和动作反应，并支持音乐参与度感知应用。为此，我们使用智能耳塞式耳机作为感测设备；这些设备已广泛用于聆听音乐。我们利用耳塞式耳机上的惯性测量单元（IMU）和麦克风，设计了反应检测技术。为了探索日常听音乐情境中的反应，我们收集了首个此类数据集MusicReactionSet，其中包含来自 30 位参与者的 926 分钟惯性测量单元和音频数据。借助该数据集，我们发现了使用音频和运动感知准确且稳健地检测音乐聆听反应时所面临的一系列独特挑战。我们设计了复杂的处理流程，使反应检测更加准确且高效。我们进行了全面评估，以考察反应检测的性能和系统成本。结果表明，在留一参与者交叉验证中，GrooveMeter对发声反应和动作反应取得的宏 $F_1$ 分数分别为 0.89 和 0.81。更重要的是，与替代方法相比，GrooveMeter 展现出更高的准确性和稳健性。我们还表明，我们的过滤方法可减少50 %或更多的能耗开销。最后，我们通过一个案例研究展示了其潜在的使用场景。

CCS 概念：• **以人为中心的计算** → **普适与移动计算系统和工具**。

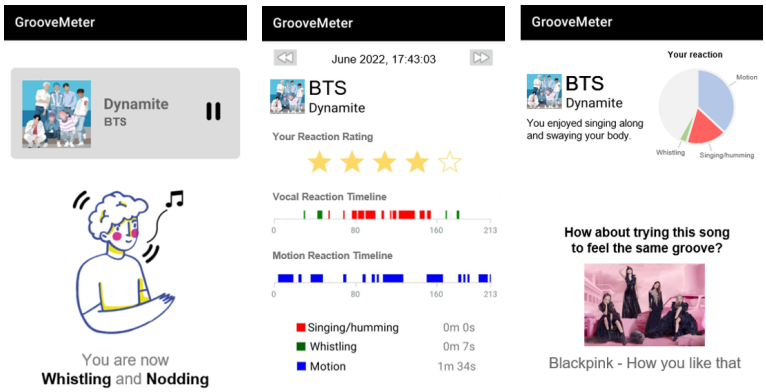
附加关键词和短语：听音乐、反应检测、耳戴设备

## 1 引言

听音乐是我们生活中不可或缺的一部分。根据一项发表于2021年的研究 [3], , 音乐消费者每天听音乐的时间超过2.63小时，相当于每天听53首歌曲。人们听歌曲或音乐时，常常会同时点头、跺脚、哼唱或跟着歌曲一起唱。这些都是人们听音乐时的自然反应 [13, 14, 38, 49], , 被视为体现他们对音乐参与度的特征 [6, 23]。这些反应为开发有趣的音乐参与度感知应用提供了有力依据。例如，音乐播放器应用（图1a）可以利用听众即时产生的反应，通过观察听众在聆听时经常对歌曲的哪些部分做出反应，提供感知参与度的自动音乐评分（图1b）和基于反应的音乐推荐（图1c）。

音乐心理学研究广泛探讨了聆听音乐时的反应。这些研究提供了深刻的见解，有助于理解聆听音乐时反应的特征，但几乎不可能将这些方法应用于现实生活场景，因为它们大多依赖自我报告，或是在受控环境中使用笨重的实验设备。例如，以往的研究使用正电子发射断层扫描（PET）或功能性磁共振成像（fMRI）测量脑部活动 [10, 36], 使用动作捕捉系统观察音乐诱发的动作 [13, 24, 40], 并测量心电图（ECG）和皮肤电反应（GSR）等生理反应 [41]。

本文提出*GrooveMeter*，一种新型移动系统，可跟踪对音乐聆听的反应，并支持音乐参与度感知应用。尽管存在各种类型的




(a) 音乐播放器 (b) 自动评分 (c) 推荐

图 1. 基于 GrooveMeter 构建的音乐参与度感知应用

作为对音乐反应的初步尝试，我们聚焦于外在且易于观察的身体反应，即身体反应，这是人们在听音乐时通常会体验到的 [25]。具体而言，我们针对跟唱、哼唱、吹口哨和头部运动，因为这些是我们在第 5节介绍的自然环境数据集中观察到的常见反应。为此，我们使用智能耳塞式耳机作为感知设备；耳塞式耳机已经广泛用于听音乐，我们通过利用耳塞式耳机上的惯性测量单元（IMU）和麦克风，设计了反应检测技术。

几十年来，研究人员为识别人类活动和手势付出了大量努力，但利用配备传感器的可穿戴设备检测音乐聆听反应，至今仍未得到研究。通过广泛观察和分析，我们发现，要利用音频和动作感知准确、稳健地检测音乐聆听反应，面临着一系列独特挑战。首先，常会出现与反应无关、却具有相似信号特征的事件，从而导致假阳性错误（例如自言自语、喝咖啡，或在显示器和键盘之间来回张望）。其次，听音乐通常是次要活动，因此音频和动作信号可能分别受到背景噪声（例如附近人交谈的声音，或咖啡馆里的背景音乐）和其他运动伪影的影响。在实验室环境数据上训练的模型，在各种日常听音乐情境中的性能会严重下降。第三，在移动设备上运行声音和动作分类模型，会产生相当大的处理开销和能耗，难以持续运行。

为应对这些挑战，我们设计了用于语音和动作反应检测的复杂处理流程，具备三个主要特征。首先，我们研究了能够被确定地 标记为无反应的数据片段的信号特征，并设计了一种过滤方法，在处理流程开始阶段有效过滤掉这些片段，不仅节省成本，还能提高稳健性。其次，我们详细设计了多步骤反应检测流程，以反映音乐聆听反应的独特模式。第三，我们利用听众的传感器数据与从歌曲中检索到的音乐结构信息 之间的语义相似度。其直觉是，音乐聆听反应与正在播放的歌曲高度相关。例如，听众的哼唱自然会遵循歌曲的节奏和音高模式。为此，我们在最后阶段根据韵律相似度，对标记含糊的音频片段进行修正。

为了构建和评估 GrooveMeter，我们收集了首个此类数据集，称为*MusicReactionSet*，其中包含30名参与者在四种情境下进行的240次音乐聆听会话（在休息室休息、

在办公室工作、乘车以及在咖啡馆放松）。其中包含时长926分钟的惯性测量单元和音频数据，以及经过人工标注的准确注释。

我们使用 MusicReactionSet 进行了全面评估，结果表明，GrooveMeter 在留一被试交叉验证中，对人声反应和动作反应检测的宏观$F_1$得分分别达到 0.89 和 0.81。更重要的是，与其他方法相比，GrooveMeter 展现出更高的准确率和鲁棒性。尤其是在嘈杂的情境下，例如在咖啡馆放松时，我们观察到性能显著提升。此外，过滤操作将测量带来的能耗开销降低了 50% 或更多。为展示 GrooveMeter 的实用性，我们还在 Android 手机和智能耳塞式耳机上制作了 GrooveMeter 原型，并基于 GrooveMeter 进一步开发了一套音乐参与度感知应用（见图 1）。我们通过应用案例研究展示了自动音乐评分、音乐熟悉度检测和基于反应的音乐推荐的可行性。

本文的贡献总结如下。

• 我们收集了首个此类数据集 MusicReactionSet，其中包含 30 位参与者在日常听音乐情境中的 926 分钟惯性测量单元和音频数据，用于研究人声反应和动作反应。

• 我们开发了 GrooveMeter，这是一种通过耳戴式感测检测音乐聆听反应的新系统。据我们所知，这是首个提出专用于自动检测音乐发声和动作反应的耳戴式感测解决方案的系统。

• 我们提出了一种新技术，通过过滤掉与反应无关的数据片段，并利用从歌曲中获取的音乐信息，使反应检测更加高效和稳健；同时，我们使用 MusicReactionSet 进行了全面评估。

## 2 相关工作

**内容消费中的反应感测：** 已有多项工作尝试监测消费者对多媒体内容或表演艺术的反应，例如，通过感测用户观看电影时的隐性反应 [7], , 通过监测自发的身体动作来测量现场演出中的观众反应 [43], , 量化观众在戏剧中的体验 [60], , 推断观众对连环漫画的幽默评价 [8], , 基于生理感测检测观众在音乐演出期间产生的激动战栗 [26], , 以及通过眼神注视和注视手势跟踪，从在线讲座中估计学习者的注意力水平 [31]。这些工作与我们的研究具有相同的高层目标，旨在实时检测内容消费者的反应。然而，由于依赖内容的反应具有不同特征，所需的传感器模态、设备和技术也应有所不同。我们的目标是通过反映反应独特的信号特征来检测音乐聆听反应。我们还发现了一个机会，可以使检测在噪声条件下更加稳健。因此，我们开发了一种新技术，用于挖掘传感器数据与音乐信息之间的语义相似度。需要注意的是，先前的工作并未利用内容的特征进行反应感测，而是仅依赖传感器数据。

**使用耳戴设备进行人体感知：** 近期研究尝试使用配备传感器的耳塞式耳机感知各种人体情境。例如，惯性测量单元可用于检测身体活动 [44],头部运动 [19], 面部表情 [37, 58], 呼吸频率 [53], 下颌动作 [33], 步态姿势 [29], 以及自由重量训练中的错误 [52]。由于距离嘴部较近，耳塞式耳机中的麦克风能够采集佩戴者更高质量的语音 [32]。它还可以检测嘴部周围发生的有趣音频事件，例如进食活动 [44]。此外，近期研究展示了声学动作追踪 [15], 步态感知 [18], 人体活动识别 [42], 用户身份验证 [61, 62], 舌头与下颌动作识别 [16], 以及面部表情追踪 [39]。配备生理传感器的耳周设备也已针对各种用途开发，例如，

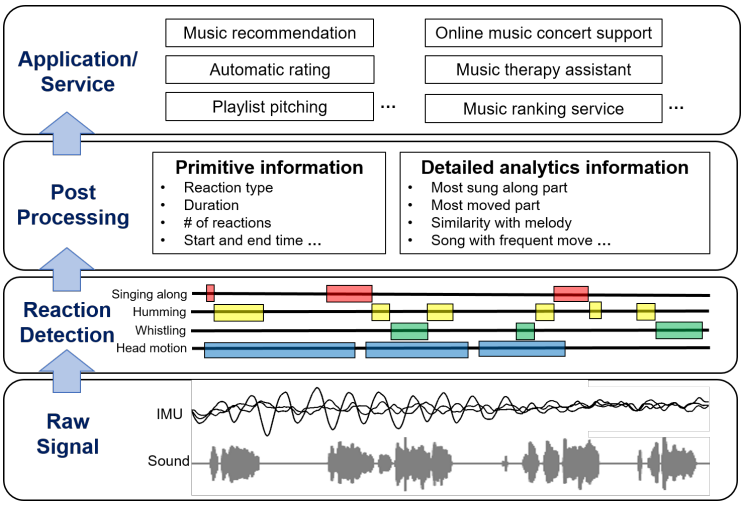



图2. 支持音乐参与度感知应用的高层流程。

面部表情检测 [17], 血压监测 [12], 微睡眠检测 [51]。近期的一项工作提出了一种基于耳机的声学感知系统，用于检测耳部疾病 [30]。此外，MusicalHeart 提出了一种利用耳机中的麦克风和惯性测量单元持续监测用户心率和活动水平的系统 [46]。在这项工作中，我们专注于音乐聆听反应，并提出了用于检测听音乐时通常产生的反应的新型传感技术。

**理解音乐聆听行为与情境：**此前一些研究尝试了解音乐聆听行为与情境。2002年，[47] 开展了最早尝试之一，运用移动技术研究这一问题。作者使用短信分析人们听过的音乐。 [63], 中，作者开发了一款基于智能手机的工具，用于收集音乐聆听行为，例如周围情境、用户活动和情绪。针对情境音乐推荐，Volokhin等人尝试了解用户听音乐的意图及其与日常活动的关系，但研究依赖于在线调查 [59]。我们的工作与这些研究有两点不同。现有研究大多关注用户的行为和情境，即人们听音乐的内容、时间、原因和地点；而我们的工作关注的是人们如何对音乐产生反应，这一问题在移动计算研究中鲜少得到探讨。其次，我们设计了新的感知和处理方法来进行反应检测，从而能够在现实场景中大规模收集数据。我们希望本研究能对现有研究有所补充。

## 3 GrooveMeter 概述

### 3.1 GrooveMeter 的应用示例

在不受约束的移动环境中实时监测音乐聆听反应，开拓了广泛的应用空间。图2 展示了 GrooveMeter 如何支持音乐参与度感知应用的整体流程。首先，作为任何应用的共同基础，GrooveMeter专注于利用惯性测量单元和音频信号实时检测发声和动作反应，即识别做出了哪种反应，以及反应发生的时间和持续时长。通过结合这些基础信息，GrooveMeter进一步为音乐参与度感知应用提供高保真信息，例如听众最常跟唱或随之动作的歌曲部分。在本研究中，我们将跟唱、哼唱和吹口哨作为发声反应，并将头部运动作为

动作反应。请注意，这些是我们在真实环境数据集（包含30名参与者）中常见的反应，详见§5。

在本小节的其余部分，我们将通过展示音乐参与度感知应用的潜在场景，说明 GrooveMeter 提供的反应信息有何用处。

**自动且细粒度的音乐评分：** 一个直接的应用场景是自动且细粒度的音乐评分。如今，音乐评分主要依赖用户的手动输入以及播放次数等简单统计数据，来了解用户对歌曲的偏好。基于人们会根据自己对正在播放的歌曲的喜爱程度而产生不同反应模式这一直觉，我们设想可以利用反应信息预测评分。我们使用来自GrooveMeter的反应输出开发了评分预测模型，并在 §6.4 中展示了其可行性。

**基于反应的音乐推荐与播放列表：** 生成定制化和个性化的播放列表，对于音乐流媒体服务提供商吸引更多消费者，以及帮助消费者从数百万首歌曲中轻松发现符合自己品味的音乐，都至关重要。Spotify 是最受欢迎的音乐流媒体服务之一，它结合协同过滤、自然语言处理和音频分析模型提供推荐服务 [2, 4]。我们设想，反应信息可以通过补充用户数据来扩展当前策略，展现听众如何与正在聆听的音乐互动，例如哪些歌曲让听众跟着唱或随节奏摇头等。我们并非认为基于反应的音乐推荐比 Spotify 的推荐服务更先进或效果更好。我们使用 GrooveMeter 为基于反应的音乐推荐开发了原型应用，并在 §6.4 中开展了一项案例研究。

**丰富粉丝与音乐家之间的远程互动：** 最近，由于 COVID-19 [1]，许多音乐家在线上举办了音乐会。我们设想，将GrooveMeter 集成到在线音乐会平台中，可以丰富音乐家与远程粉丝之间的互动。它能够检测并收集在线观看现场演出的粉丝的反应。收集到的反应可以用于生成某种集体声音化或可视化效果，并呈现给音乐家。例如，可以根据人声反应生成欢呼和跟唱的合成音效，以传达粉丝的热情。这些反应及其效果还可以在粉丝之间共享，从而帮助他们更有参与感和沉浸感地享受音乐会。

**音乐治疗辅助工具：** 我们设想，除了丰富音乐播放器之外，GrooveMe- ter 还可以作为基础，用于开发各种以音乐为媒介的领域中的辅助工具，例如音乐治疗、音乐心理学和教育。在临床音乐治疗会话期间，治疗师和来访者会进行聆听音乐、唱歌、演奏乐器以及创作音乐等音乐体验 [11]。为了探索 GrooveMeter 在音乐治疗中的潜力，我们就当前临床外音乐治疗的实践及其局限性咨询了一位音乐治疗师。该治疗师提到，有时会给来访者布置任务，使他们能够将临床会话中进行的活动应用到日常生活中。然而，由于目前治疗师依赖来访者填写的清单或口头报告，因此很难完全检查来访者是否遵从。治疗师提出，GrooveMeter可以通过在家监测和分析来访者的音乐聆听会话，为治疗师提供客观测量工具。GrooveMeter 还可以进一步扩展，以采用其他生理传感器来分析来访者的心理状态，例如情绪和压力。

### 3.2 设计考量

我们提出 GrooveMeter，这是一套利用耳戴式感测追踪用户聆听音乐时反应的系统，可为音乐参与度感知应用提供支持。我们介绍开发 GrooveMeter 时的三项主要设计考量。

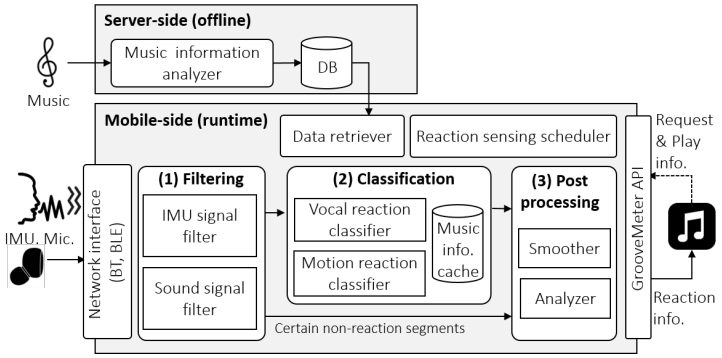



图 3. GrooveMeter架构

**无感感知：**为了支持现实场景中的音乐参与度感知应用，重要的是追踪用户对音乐的反应，而不依赖部署在基础设施中的传感器或过多的穿戴式传感器。

**准确且稳健的检测：**与日常人类行为情境相比，音乐聆听反应具有独特的语义特征，但它们有时会表现出相似的信号特征。例如，用户的点头可能是对正在播放的歌曲所做出的反应，但也可能是用户与附近的人进行面对面互动时做出的回应。用户在听音乐时跟着哼唱，或自言自语时，都可能发出喃喃声。由于听音乐通常作为次要活动进行，这类与反应无关的行为、背景噪声以及其他运动伪影可能会降低性能。在日常听音乐情境中，确保反应检测准确且稳健至关重要。

**低开销：**尽管GrooveMeter仅在用户聆听音乐时运行，但保持低开销仍然十分重要。人们通常已经会听相当长时间的音乐（在 2021 [3] 中每天 2.63 小时），他们也不希望消耗手机的大量电池。

### 3.3 GrooveMeter架构

图3展示了GrooveMeter架构。它获取歌曲的音乐信息，并将其存储在服务器上的数据库中。当用户开始听一首歌曲时，GrooveMeter会在耳塞式耳机上启用音频和运动感知，并将传感器数据流发送到智能手机；由于耳塞式耳机的处理和编程能力有限，目前我们仅将其用作传感器数据流提供方。所有处理均在用户的智能手机本地完成，并经过用户授权，以避免隐私顾虑。GrooveMeter的主要操作如下。

(1) 首先，它研究信号特征，并筛除能够被确定标记为无反应的数据片段（§4.1.3 和 §4.2.3）。(2) 对于不确定的数据片段，它通过分类模型识别反应事件（§4.1.4和§4.2.5）。(3) GrooveMeter 利用从歌曲中检索到的音乐信息来提升分类性能。GrooveMeter计算感知信号与音乐信息之间的相似度，并根据相似度修正标签（§4.1.5）。(4) 基于检测到的反应事件，GrooveMeter执行后处理操作，以提供高保真信息（§4.1.6和 §4.2.6）。GrooveMeter 还在 人声反应检测的后处理操作中采用输出平滑（§4.1.6）。

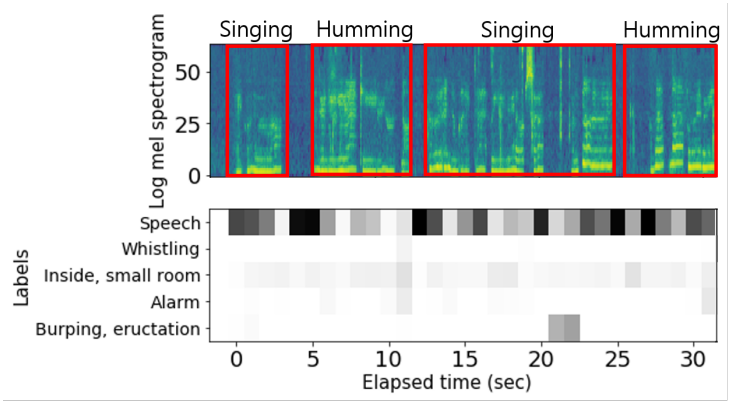



图 4. YAMNet 对人声反应的检测结果

## 4 反应感测

### 4.1 人声反应检测

*4.1.1*挑战。 在此，我们介绍了人声反应检测需要解决的若干技术挑战。

**发声反应事件：**检测人声反应等声音事件的一种直接方法是开发音频模型或使用预训练模型。然而，开发自定义模型需要投入大量精力进行大规模现实场景数据收集。此外，我们发现，直接采用现有预训练模型并不适合我们的用途。近年来，许多声音分类模型相继发布，这些模型使用海量数据进行预训练，能够预测大量现实场景中的声音事件。一个具有代表性的例子是 Google 于 2019 年发布的 YAMNet [28], 。该模型使用超过 200 万个 YouTube 视频进行训练，可对 521 种音频事件进行分类。

虽然这类预训练模型在日常事件上的表现令人满意，但据我们所知，目前没有任何模型包含我们的目标事件，即听音乐时的人声反应，因此它们对这些反应事件的检测性能较差。例如，YAMNet 有唱歌标签，但相应的音频数据大多取自播放歌曲并伴有乐器声的视频片段，例如音乐视频或乐队现场演出片段。然而，听众通过耳塞式耳机听音乐时做出唱歌反应，所录制的音频信号只包含听众的唱歌声，不含背景音乐。图 4 展示了我们使用 YAMNet 进行的初步研究：我们测试了该模型对听众跟着歌曲唱歌时录制的音频片段的检测性能；图中显示的是 softmax 层的输出。结果表明，YAMNet 几乎不会选择唱歌标签，而是大多将唱歌反应分类为语音标签。吹口哨的音频片段不会被标记为语音，但有时会被标记为与动物有关的类别，例如鸟鸣。使用新收集的数据开发自定义模型或许可以解决这些问题。然而，为了反映用户差异，并避免现实场景中各种噪声导致的推断错误，仍需收集大规模的现实场景声音事件数据，且范围不能仅限于反应事件。

**混有背景噪声：** 背景噪声（例如附近人们交谈的声音或咖啡馆中的背景音乐）使得正确检测人声反应变得更加复杂。我们观察到，YAMNet 经常将咖啡馆中混有各种噪声的唱歌和哼唱反应分类为音乐 标签。图5 展示了这样一个例子。此外，在噪声条件下，YAMNet对与反应相关标签的softmax分数往往相对较低，有时甚至低于与反应无关的标签。根据我们在咖啡馆等嘈杂场所收集的数据

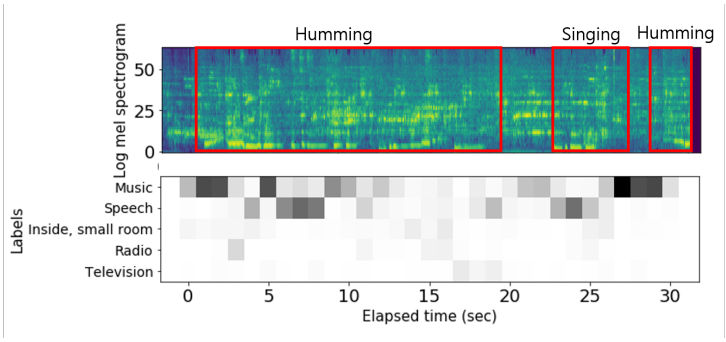



图 5. 背景噪声下的 YAMNet 结果

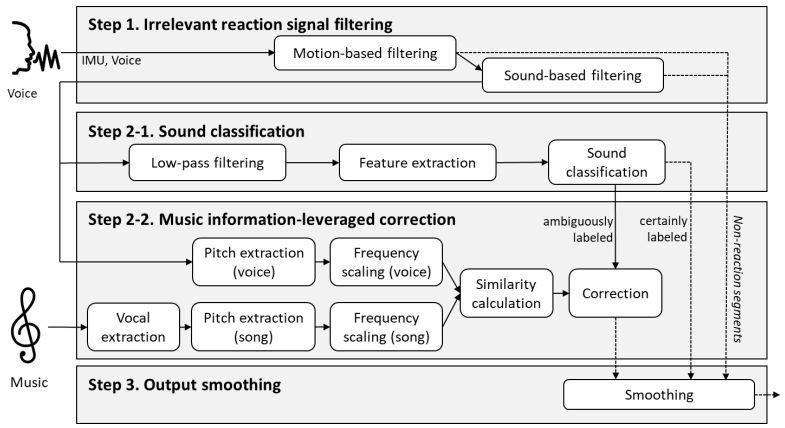


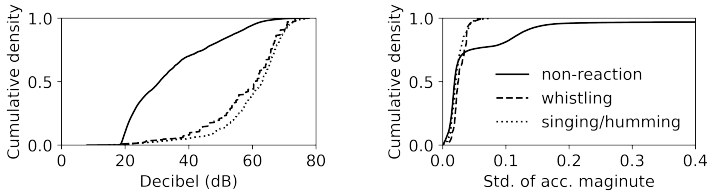

图 6. 人声反应检测流程

以及汽车，只有5.3%的歌唱/哼唱片段的YAMNet softmax 分数大于 0.95；在噪声较小的场所（如休息室），这一比例为 11.7%。

**间歇性与交替性：**人声反应并不会在一次活动中持续不断地发生，而是间歇出现，有时还会与其他反应交替发生。例如，当听众跟唱时，我们会观察到他们在唱歌过程中会短暂停顿以吸气。此外，听众经常交替做出不同类型的人声反应。例如，听众跟唱时，如果不熟悉歌词，往往会暂时改为哼唱或吹口哨，等到熟悉歌词的部分再继续跟唱。

**处理成本：**声音分类通常涉及计算密集型操作，例如 MFCC 计算和深度神经网络。尽管如今的模型已针对设备端处理进行优化，例如 YAMNet 的 MobileNet 架构，但持续运行仍会带来显著的处理和能耗成本。例如，在 Galaxy S21 上播放歌曲时，持续使用 YAMNet 进行音频分类，一小时会导致电池电量下降 3%。

*4.1.2* 概述。为应对上述挑战，我们设计了一种新颖的流程，能够高效、可靠地检测人声反应。图 6 展示了该流程的概况，包含三个主要操作。首先，为了节省成本，我们采用早期过滤操作。其核心思路是分析音频和动作信号的特征，并过滤掉能够确定标记为无反应的数据片段（§4.1.3）。其次，我们使用 YAMNet 模型对声音事件进行初步分类（§4.1.4）




(a) 声音强度　　(b) 运动程度

图 7. 声音强度和运动程度的累积分布函数

并利用从正在播放的歌曲中提取的音乐信息来修正模糊标签（§4.1.5）。具体而言，该方法会提取音频信号和歌曲的时序音高模式，并通过比较二者来修正标签。最后，它会平滑最终输出，以应对偶尔出现的短暂、间歇性事件，并提供高保真信息（§4.1.6。）

*4.1.3* 确定无反应信号的过滤。为了节省成本，我们采用早期过滤操作。该操作可识别能够确定标记为无反应的数据片段，并避免对这些片段执行计算量大的分类操作。识别逻辑基于两个观察结果。首先，人声反应会产生超过一定音量的声音事件，尤其是因为音源（耳塞佩戴者的嘴）与耳塞麦克风之间距离很近。这意味着，低于某个音量阈值的声音，例如静音和背景噪声，都可以确定标记为无反应。图 7a 展示了反应声音和无反应声音每秒分贝数的累积分布函数，验证了我们的假设。其次，更值得关注的是，人声反应还会使耳塞产生一定程度的动作。这是因为佩戴者做出人声反应时的嘴部动作会激活位于嘴和耳朵之间的颧大肌，从而在耳塞的惯性信号中引发脉冲响应。近期研究还表明，耳塞能够捕捉由面部表情 [37] 和无声发音 [33]产生的动作信号。与第一个观察结果类似，这也意味着，如果未检测到动作，相应的音频信号极不可能属于反应标签。图 7b 展示了一秒内加速度计幅值标准差的累积分布函数。有趣的是，大幅度动作也与无反应标签相关，因为听众在走路、跑步和锻炼时几乎不会做出人声反应。

基于这些发现，我们设计了一个两步过滤组件。首先，它监测动作水平（定义为加速度计幅值的标准差），并过滤掉超出阈值范围的数据片段（根据 §5 中对数据集的分析，我们将范围分别设为 0.0104 和 0.12）。其次，对于未过滤的片段，它监测相应音频信号的分贝值，并过滤掉分贝值低于阈值的数据片段（在我们的实现中，阈值为 49 分贝）。我们将基于动作的过滤器置于基于声音的过滤器之前，因为前者更加轻量。根据我们的测量，基于动作的过滤器在 GalaxyS21 上消耗 113 mW，而基于声音的过滤器消耗 134 mW。它将过滤后的片段标记为 无反应 ，并将其传递给后处理操作，而不执行后续操作。

*4.1.4* 声音事件分类。 **目标事件：** 我们针对三类发声反应事件；唱歌（跟唱）/哼唱、吹口哨和无反应。如上所述，由于即使在单个反应时段内也经常交替出现唱歌和哼唱，我们将二者合并为同一类别。

**预处理：** 来自耳塞式耳机的音频数据以16kHz重采样，并划分为1秒长的数据片段。然后，我们对这些数据片段应用1阶低通滤波器（LPF）（截止频率为2 kHz）

Tab表 1：从 YAMNet 到 GrooveMeter 的映射 l 标签

| YAMNet | GrooveMeter |
|---|---|
| 哼唱，唱歌 | 歌唱/哼唱 |
| 吹口哨、口哨 | 吹口哨 |
| 语音、音乐 | 含糊不清（可能是歌唱/哼唱或无反应） |
| 其他 | 无反应 |

以减少频率高于目标人声反应主要频率范围的噪声信号。

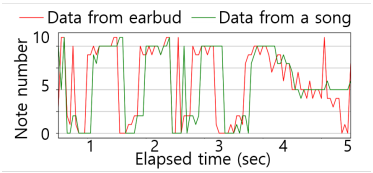
**音频特征提取：** 随后，使用短时傅里叶变换和周期性 Hann 窗将预处理后的片段转换为频谱图；我们将窗口大小和窗口步长分别设为 25 ms 和 10 ms。接着，将频谱图映射到 64 个梅尔频带，频率范围为 125 至 7,500 Hz，并通过取对数计算对数梅尔频谱图。最后，将特征帧化为一个矩阵，其中包含 $96 \times 64$ （96 帧，每帧 10 ms；每帧包含 64 个梅尔频带）。

**分类与标签标注：** 我们使用 YAMNet [28] 作为声音事件分类的基础组件。然而，由于 YAMNet 的标签分类体系与我们的目标类别并不完全匹配，我们构建了从 YAMNet 标签到 GrooveMeter 标签的映射，如表 1 所示。如前所述，正如图 4 和图 5 所示，YAMNet 难以区分唱歌 反应与语音和音乐信号。因此，我们将 YAMNet 分类输出中的语音和音乐标签映射为模糊。对于模糊片段，我们通过后续的校正操作进行进一步调查。其他标签则直接送入后处理操作。

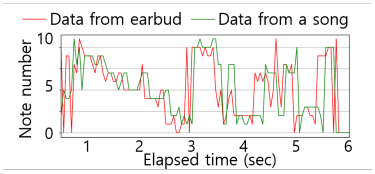
**秩约束松弛：**即使应用标签映射，仅依赖 YAMNet 的分类结果也不足以准确检测人声反应。由于背景噪声以及人声反应的特性，在我们的数据集中，YAMNet 仅将 57.8% 的歌唱/哼唱片段的语音或音乐识别为排名第一的分类标签。若将排名较低的语音或音乐标签也纳入，比例便会提高，例如排名前 2 为 65.4%、前 3 为 70.1%、前 4 为 73.3%、前 5 为 75.7%，依此类推。吹口哨片段也呈现出类似的特征。

为了解决这个问题，我们采用秩约束松弛策略，允许部分未被 YAMNet 归为目标标签之一的片段进入修正步骤。在此需要取得平衡，既要避免修正步骤产生不必要的成本，也要防止由于额外的无反应片段可能被错误分类为人声反应标签而导致假阳性错误增加。如果 YAMNet 的分类输出质量足够高，我们就不进行松弛。否则，我们会检查排名较低的输出中是否包含人声反应标签。如果包含，我们就将其视为需要通过修正步骤进一步调查的不确定标签。

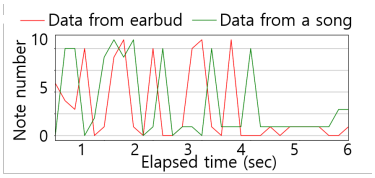
为了量化分类输出的质量，我们采用了主动学习领域提出的不确定性采样策略。已经开发出多种技术来计算推理实例的不确定性，例如最小间隔 [55]和最高熵 [56]。目前，我们选择最小间隔，通过计算排名前两位输出类别的置信度值之差来衡量不确定性。如果某个片段的间隔低于阈值，我们会检查 YAMNet的前k个输出标签。如果其中包含我们的目标标签，即语音、音乐、哼唱、唱歌、吹口哨和哨声，则该片段被视为带有不确定标签，例如不确定的哼唱，并会被转发至校正步骤。在当前实现中，我们通过经验将间隔阈值和 k 分别设为 0.9 和 5。



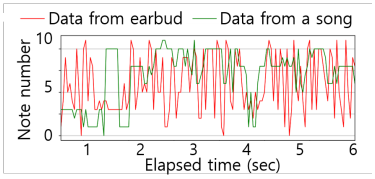

(a) 跟唱




(b) 哼唱

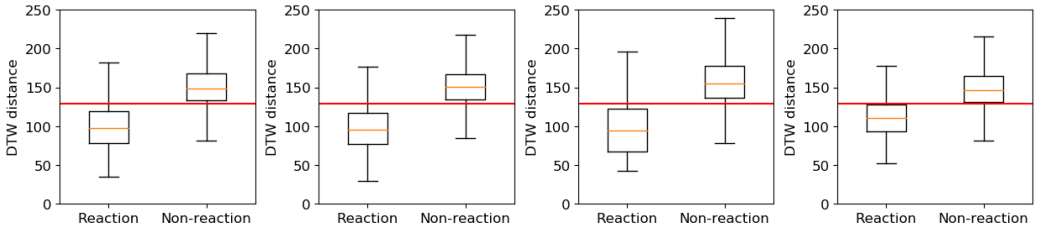



(c) 吹口哨




(d) 无反应

图 8。 半音阶的音符




(a) 休息室 (b) 办公室 (c) 汽车 (d) 咖啡馆

图 9. 相似度差异

*4.1.5* 利用音乐信息的校正 。我们通过利用正在播放的歌曲的音乐信息，最终确定上一步中存在歧义或不确定的片段。基于音频信号（来自耳塞）与歌曲之间的韵律相似度，我们将带有语音或音乐标签的歧义片段校正为歌唱/哼唱 或无反应。对于不确定的片段，我们也采用同样的处理方式。

**韵律相似度计算：**为了衡量人声信号与歌曲之间的韵律相似度，我们考察旋律，即一系列线性排列的音乐音调。我们的核心直觉是，人声反应会跟随正在播放的歌曲中的音符序列，而与反应无关的语音信号则不会。图 6 展示了详细流程（见步骤 2-2）。为了提取音符序列，我们首先使用 CREPE [35],（一种先进的音高跟踪器）提取音高信息；也就是说，每隔 0.1 秒提取一次频率信息。然后，我们将音高（频率）转换为带八度编号的音符。接着，我们将其转换为 12 音半音阶，即去掉八度编号，因为我们观察到，人声反应的音高通常比正在播放的歌曲高或低一个八度。对于歌曲音频文件，我们会在提取音高之前先提取人声，以聚焦于音乐中占主导地位的旋律线。这一发现也源于我们的观察：人声反应大多会跟随人声（唱歌声），而不是乐器。我们使用 Spleeter [27] 从歌曲中分离人声声源。

我们计算两个音符序列之间的相似度（一个来自用户的声音信号，另一个来自歌曲），并作出最终判断。我们将 12 个音符（C、C#、D、D#、E、F、F#、G、G#，

A、A#、B）映射为 12 个整数值（0 到 11）。图 8a、8b 和 8c 显示，音符模式与正在播放的歌曲高度相关，而无反应部分则没有这种相关性（图 8d）。我们采用动态时间规整（DTW）作为相似度度量函数，因为两个音符模式的速度可能不同。

如果 DTW 距离大于阈值，我们就将该片段标记为无反应。否则，我们会应用标签映射并确定最终标签，即将 speech 和 music 确认为歌唱/哼唱，将 whistling 和 whistle 确认为吹口哨，将 humming 和 singing 确认为歌唱/哼唱。在我们的实现中，我们通过经验将阈值设为 130。图 9 展示了数据集中反应（即歌唱/哼唱）片段和无反应片段的 DTW 距离值分布。有趣的是，无论噪声条件如何，距离的分布都相似。在咖啡馆这一数据集中噪声最大的情形下，反应片段的距离中位数略高于其他情形。尽管如此，大多数反应片段仍可与无反应片段区分开来。

*4.1.6* 后处理。**平滑：**我们使用隐马尔可夫模型（HMM）对分类输出进行平滑。其核心思想是根据训练数据集的分类输出序列训练 HMM 模型，并使用训练好的模型进行输出平滑。我们将观测序列定义为分类输出序列，并通过估计最优隐状态序列来进行平滑，该序列可映射为平滑后的反应事件序列。具体而言，对于给定的时间 $t$、$O^t$ =（$o_1$，..... $o_t$）处的分类输出序列，我们提取从时间 1 到 $t-1$ 的最大概率隐状态序列 $S^t$ =（$s_1$，.....$s_{t-1}$）。然后，时间 $t$ 处的平滑值 $\hat{s}_t$ 可按如下方式得到：

$$\hat{s}_t = \underset{s_t}{\operatorname{argmax}} p(s_t|O^{t+1}, \lambda) \tag{1}$$

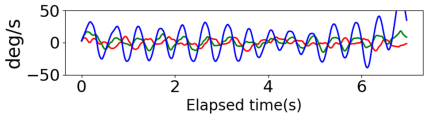
我们应用维特比算法 [22] 以高效计算最大概率，并将长度为 6 秒的窗口作为输入序列，即最近 6 个分类输出的序列。如果应用更倾向于为交互式服务提供实时输出，则可以省略平滑操作。

**高保真信息：**GrooveMeter 提供的原始信息是发声反应事件的发生情况，即反应类型以及开始和结束时间。我们可以通过汇总事件进一步提供高保真信息，例如找出跟唱最多的主歌；也可以通过分析音频信号提供高保真信息，例如根据韵律相似度为人声反应评分。此类信息的定义及其可用性评估需要根据应用需求区别考虑。我们将其留待未来工作。

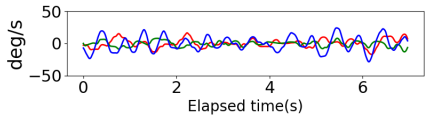
### 4.2 动作 反应检测

*4.2.1* 挑战。我们介绍了在设计动作反应检测流程时所考虑的挑战。

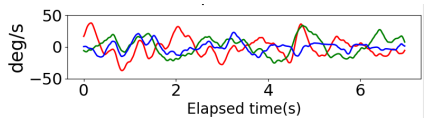
**具有周期性和重复性，但多样化的运动轨迹：**根据我们的观察，我们发现动作反应通常呈现重复且有周期性的模式。例如，点头（图 10a 和 10b）会以规律的模式持续一段时间，从而产生具有一定周期性的信号波形。有人可能会认为，这类重复模式可以通过典型的身体活动识别流程轻松捕捉。然而，我们发现，正如 §6.2 所示，使用广泛采用的、用于表征信号周期性或统计特征的特征的分类器，对于真实环境中的反应数据并不奏效。这主要是因为，与遵循典型运动轨迹、定义明确的活动或手势不同，现实生活中的动作反应所表现出的运动行为往往有所变化。例如，人们听音乐时通常会上下摆动头部，但有时也会左右移动或倾斜头部。即使是同一用户或同一场聆听活动，动作反应的幅度和速度也会有所不同。图 10a



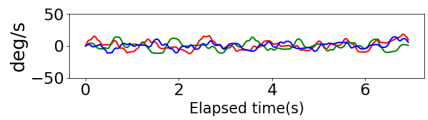

（a）点头——案例 1




（b）点头——案例 2

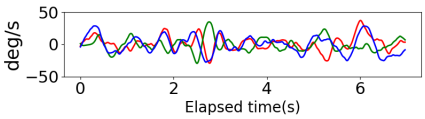



（c）伴随身体摇摆的头部运动——案例 1

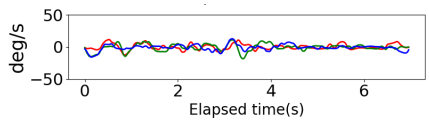



（d）伴随身体摇摆的头部运动——案例 2

图 10. 多种动作反应模式




(a) 乘车




(b) 咀嚼饼干

图 11. 重复且与反应无关的动作

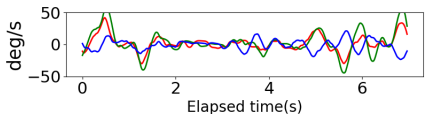



图 12. 在进行其他活动时点头

以及10b 展示了同一位参与者的两种不同点头模式，图10c 和图 10d 展示了另一位参与者的头部运动以及身体摇摆。

**与反应无关的动作产生的运动轨迹令人困惑：** 根据我们使用真实场景数据进行的初步研究，我们发现，有几种与反应无关的动作会呈现重复的信号模式，因此可能导致反应检测中的分类错误。例如，当一个人乘车时，即使没有做出任何其他明显动作，汽车的运动和振动也会导致耳塞式耳机上的惯性测量单元信号重复出现（见图 11a）。同样，图 11b 显示，咀嚼饼干也会产生一定程度的重复模式。

**受其他运动伪影影响：** 人们通常会将音乐作为次要活动来听，这意味着他们的主要任务（例如在办公室工作、在咖啡馆放松）以及相应的动作（例如敲击键盘、咀嚼饼干）都可能影响惯性测量单元信号。图 12 展示了用户在进行其他活动时随节奏点头的陀螺仪数据示例。如图所示，与图 10a 和图 10b 相比，点头动作所产生的周期性表现得不那么明显。

*4.2.2*概述。我们设计了一种新颖的动作反应检测流程，以应对上述挑战。图 13展示了该流程的概览，其中包含两个主要操作：预处理和反应检测。在预处理阶段，我们采用了一个简单的滤波器，以避免不必要的

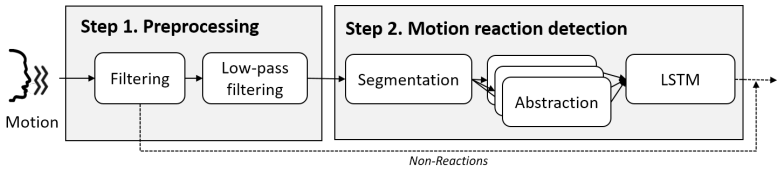



图 13. 运动反应检测流程。

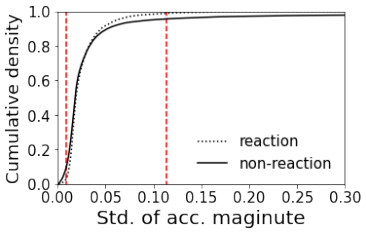



图 14。运动反应的运动程度累积分布函数。

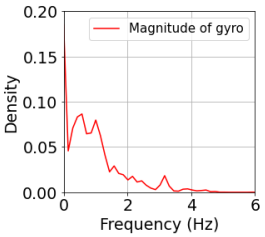



图 15.反应数据的主导频率分布。

用于分类操作的处理。它会预先将不太可能是反应的输入数据的输出标签判定为无反应，例如，没有动作或动作幅度过大（§4.2.3）。随后，我们采用低通滤波器去除由其他运动伪影造成的噪声（§4.2.4）。其次，我们对原始 IMU 信号进行抽象，并提取一系列运动单元，以表示用户动作模式的抽象。借助这一序列，我们使用长短期记忆网络检测动作反应，执行二元分类（头部运动与无反应）（§4.2.5）。

*4.2.3* 与反应无关的运动过滤。我们根据观察设计了一个简单的基于阈值的滤波器。它通过查看运动程度来筛除与反应无关的数据。我们将运动程度定义为一秒钟加速度计信号片段的标准差。与人声反应检测（§4.1.3）中的动作过滤器类似，显然，没有动作意味着无反应。较大的动作也与无反应标签相关，因为听众在锻炼、步行或跑步时点头的可能性很低。

在实现过程中，我们仔细选择了一个阈值范围，以便在不遗漏反应案例的情况下，尽可能筛选出更多非反应案例。为此，我们研究了数据集的动作程度。如图 14所示，我们检查了动作程度的累积分布函数（CDF），并根据经验将低阈值和高阈值分别设为 0.0092 g 和 0.114 g。

*4.2.4* 噪声去除。下一步是去除由其他运动伪影引起的运动噪声。这一想法背后的直觉是，听者的动作反应往往会跟随歌曲的节拍模式。因此，考虑到常见音乐流派的典型速度介于每分钟 60 至 180 拍之间，我们可以预期动作反应通常表现为低频动作。图 15 展示了从我们的反应数据中提取的主频分布。几乎所有主频都低于 4 Hz。因此，我们使用截止频率为 5 Hz 的一阶低通滤波器处理以 70 Hz 采样的原始惯性测量单元数据，以留出一定余量。

*4.2.5*动作反应分类。分类操作包括两个主要步骤。第一步是提取运动抽象序列。利用IMU数据，我们提取时间运动模式。如前所述，动作反应并不表现出定义明确的典型动作

表 2。音乐聆听情境的特征 tions

| 情况 | 与反应无关的动作 | 背景噪声 |
|---|---|---|
| 在休息室休息 | 坐在椅子上时的随机动作 | 空气净化器发出的噪声，<br>外面的低声交谈声，..... |
| 在办公室工作 | 网页搜索和文字处理时的动作 | 键盘打字，<br>鼠标点击声，... |
| 乘车 | 沿着道路蹦跳前行 | 各种驾驶汽车的噪声 |
| 在咖啡馆放松 | 饮用咖啡,咀嚼一块饼干 | 背景音乐，<br>附近的交谈、咀嚼声，..... |

轨迹和持续时间。我们设计了一种方法，以简单的方式抽象反映人体运动的 IMU信号，并据此检测动作反应。为进行抽象，我们定义了一个运动单元，用于表示从一段较短时间间隔的 IMU数据中提取的一组特征。我们从 IMU信号中提取运动单元，并生成运动单元序列。该序列可视为对用户运动模式的抽象。它可以表示对音乐的动作反应、与反应无关的随机身体动作，或静止状态。为提取运动单元，我们将经低通滤波器处理的 IMU数据按 100 毫秒的长度分段。对于每个片段，我们计算陀螺仪各轴的一组统计特征，即最大值、最小值、平均值、极差、标准差和均方根。

下一步是将一系列运动单元分类为两个类别之一：头部运动和无反应。鉴于输入是时间序列，我们采用LSTM模型，该模型广泛用于预测时间序列等序列数据。我们根据实验结果选择LSTM，与RNN和门控循环单元等其他方法相比，LSTM表现更佳。我们构建了一个分类模型，由包含32个隐藏单元的LSTM层、丢弃率为0.5的丢弃层、ReLU层和softmax层组成。该模型使用Keras API实现，并使用Adam优化器训练最多300个周期。我们根据经验将分类的窗口大小设为7秒，此时 $F_1$ 分数开始趋于饱和。一个数据窗口被构造成一个 $70 \times 18$ 矩阵，输入模型进行处理。

*4.2.6*后处理。GrooveMeter将分类器输出和滤波结果结合起来，并将其作为无反应，以提供最终的推理输出。GrooveMeter 提供的基础信息是运动反应事件的发生情况。通过聚合事件信息，它进一步提供详细的分析结果，例如听众动作最多的歌曲部分，以及使听众频繁动作的歌曲。

## 5 音乐反应数据

为了构建和评估 GrooveMeter，我们创建了 MusicReactionSet，这是一个新颖的数据集，包含来自多种音乐聆听反应的音频和 IMU数据。据我们所知，这是首个针对音乐聆听情境中的反应的数据集。数据收集是在获得机构审查委员会批准后进行的。

**参与者：** 我们从一所大学校园招募了 30 名参与者（男：18 名，女：12 名），$P_A1$ 至 $P_A30$；他们的年龄在 20 至 26 岁之间（平均年龄为 22.7 岁）。我们获得了参与者的知情同意。所有人都表示，他们在日常生活中经常听音乐。完成数据收集后，他们获得了一张价值 18 美元的礼品卡作为补偿。

**自然场景数据采集：** 每位参与者被邀请到四个地点，在进行其他活动的同时听一组歌曲，即在休息室休息、在办公室工作、乘车以及在咖啡馆放松（见图 16）。我们认为，这些地点（a）反映了人们经常享受听音乐的多样现实场景，并且（b）有助于研究多样的音频噪声和运动噪声对反应检测的影响。表 2 展示了四种场景的特征。



(a) 休息室 (b) 办公室 (c) 汽车 (d) 咖啡馆

图 16. 在不同地点进行的自然场景数据采集。

请注意，汽车乘坐场景也反映了其他类似情况，例如乘坐出租车和公交车等公共交通工具。

具体而言，在每种情境下，参与者从音乐流媒体服务的排行榜前 50 名中，自由选择三首属于两种不同音乐流派的歌曲，即令人振奋的/快节奏歌曲和慢节奏/柔和歌曲。他们使用提供的耳塞设备和智能手机听这些歌曲。为了收集他们自然反应的数据，我们没有对他们如何做出反应作出任何指示，并让他们独自待在相应场所（乘车时除外）。请注意，我们未纳入第一首歌曲的数据，因为参与者换到新地点后通常会立即感到有些分心。最终，我们从 240 次音乐聆听会话（每位参与者 8 次）中共收集了 926 组时长一分钟的惯性测量单元和音频数据。其中一半对应令人振奋的/快节奏歌曲，其余则对应慢节奏/柔和歌曲。

**设备设置**：我们使用耳塞式耳机 进行音乐流媒体播放和 IMU/音频感知，并使用一部 Android 手机通过蓝牙连接耳塞式耳机并控制音乐播放。耳塞式耳机方面，我们使用了 Apple AirPods Pro，但也额外使用了 eSense [32] 来收集 IMU数据，即左耳佩戴一个 eSense 单元，右耳佩戴一个 AirPod Pro 单元；请注意，我们收集数据时，AirPods Pro 不允许开发者访问原始惯性测量单元数据，但最近已开始在 iOS 14.0 中提供访问权限。AirPods Pro 上麦克风和 eSense 上惯性测量单元的采样率分别设为 44.1 kHz 和 70 Hz。为了进行真实情况标注，我们使用一台带遮挡的摄像机记录数据收集过程，该摄像机在图 16 中以红色矩形标出。

请注意，我们还开展了一项小规模部署研究，参与者在不使用摄像头的情况下使用 GrooveMeter（见第 6.4 节）。

## 6 评估

为了进行评估，我们在两部手机上将 GrooveMeter 的原型实现为 Android 服务：GalaxyS21（Android 11.0）和 Galaxy S8+ （Android9.0）。我们使用 TensorFlow Lite [5] 运行基于 YAMNet 和长短期记忆网络的处理流程。我们还实现了一个音乐播放器应用原型，该应用提供反应摘要和音乐推荐，如图1所示。

### 6.1 人声反应检测

*6.1.1* 总体性能。我们展示了 GrooveMeter 人声反应检测的总体性能。验证时，我们使用 MusicReactionSet 数据集进行留一受试者交叉验证（LOSO CV）。请注意，我们在分类任务中使用了原始 YAMNet 模型 [28]，因此所有受试者的测试均使用同一模型。为避免过拟合问题，我们通过留一受试者交叉验证来确定人声反应检测流程中的阈值（动作和音频滤波器的阈值范围、音高相似度阈值）。

图 17 展示了 30 次验证中人声反应标签的平均精确率和召回率。实验结果表明，GrooveMeter 在

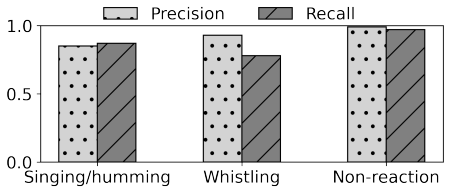



图 17. 总体性能

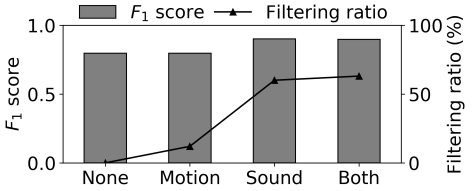



图 18. 过滤的效果

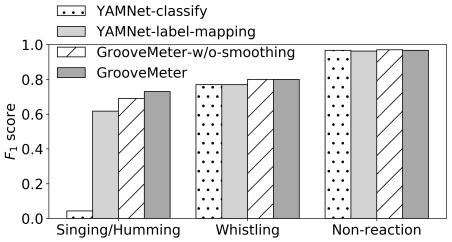



图 19. 校正和平滑的效果

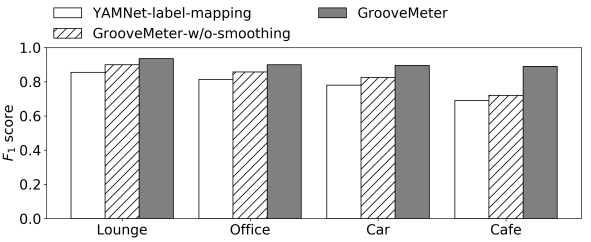



图20. 对噪声的鲁棒性

即使面对未见过的用户以及各种真实生活中的背景音频噪声类型时，仍能实现合理的人声反应检测性能；宏平均 $F_1$ 分数为 0.90。具体而言，它检测歌唱/哼唱 反应的精确率和召回率分别为 0.85 和 0.87，检测吹口哨 反应的精确率和召回率分别为 0.93 和 0.78。与其他标签相比，吹口哨的召回率相对较低，因为 YAMNet 会将一些声音微弱或与背景噪声混杂的吹口哨片段错误地判定。结果还表明，我们的方法能够正确识别无反应事件。无反应 标签的精确率和召回率分别为 0.99 和 0.97。

*6.1.2* 过滤的影响： 我们考察了早期过滤的影响。图 18 展示了不同过滤策略下的$F_1$ 分数和过滤比例。为此，我们开发了三种不同版本的人声反应检测流程：无、动作和声音。我们将过滤比例定义为已过滤片段数除以片段总数；也就是说，过滤比例越高，处理成本越低。无表示不进行任何过滤操作的人声反应检测流程，即过滤比例为 0%。动作和声音分别表示仅添加基于动作和基于声音的操作时的流程。

有趣的是，过滤操作不仅能有效降低计算成本，还能通过有效过滤掉与反应无关的音频信号，提升人声反应检测性能。虽然*none* 在完全不过滤的情况下达到 0.8 的 $F_1$ score，但同时应用两种过滤操作*both*后， $F_1$ score 提高了 0.1。我们还观察到不同过滤模式的效果各不相同。基于动作和基于声音的过滤操作的过滤率分别为 12% 和 60%。简单的基于动作的过滤能减少相当数量待处理的非反应数据，同时不影响性能。基于声音的过滤则能减少更多数据。同时，它还能有效去除原始 YAMNet 因背景噪声导致的假阳性，从而提高 $F_1$ score。在 GrooveMeter 中同时使用这两种操作时，过滤率提高至 63%，这表明即使只对 37% 的片段进行分类，也能取得比*none*更好的性能。

*6.1.3* 校正和平滑的影响：我们进一步分析了其余操作——校正和平滑——的性能。为量化每种技术的影响，我们实现了该方法的三个变体：*YAMNet-classify*、*YAMNet-label-mapping* 和*GrooveMeter-*无平滑，分别依次应用每种技术，并比较它们的性能；此处，我们排除了早期过滤操作，以重点考察分类性能。作为基线，YAMNet-classify 直接从原始 YAMNet 模型输出反应标签，不进行任何后续操作。YAMNet-label-mapping 将 YAMNet 模型输出中的语音和音乐标签映射为歌唱/哼唱。GrooveMeter-无平滑则处理表 1 中通过标签映射获得的模糊标签和秩约束松弛，并执行利用音乐信息的校正，但不进行平滑。

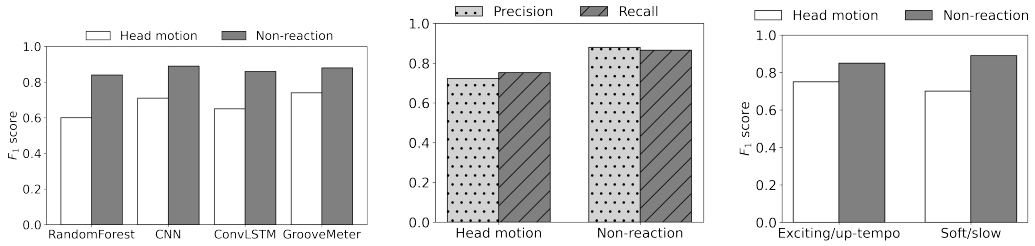
图 19 展示了各变体的 $F_1$ 分数。尽管使用了海量数据进行训练，我们仍观察到YAMNet分类在人声反应数据上的表现不佳，尤其是在歌唱/哼唱方面（ $F_1$ 分数为 0.04）。如 §4.1 所述，这主要是因为 YAMNet 模型极难区分跟唱数据与语音和音乐数据。YAMNet 模型仅将 1.8% 的歌唱/哼唱数据分类为唱歌或哼唱标签。应用标签映射后，YAMNet 标签映射在歌唱/哼唱数据上的表现有了明显提升，但仍出现相当多的假阳性错误，例如将背景音乐和附近人们的交谈误判为歌唱/哼唱。与 YAMNet 标签映射相比，我们的YAMNet输出松弛和利用音乐信息的校正提升了 $F_1$ 分数。尤其是，人声反应的分类性能有所提升（YAMNet 标签映射与 GrooveMeter无平滑之间的差异）。具体而言，歌唱/哼唱的 $F_1$ 分数从 0.61 提升至 0.68，吹口哨的分数从 0.77 提升至 0.8。我们还观察到，基于HMM的平滑显著提升了人声反应检测的性能（GrooveMeter无平滑与 GrooveMeter 之间的差异）。歌唱/哼唱的 $F_1$ 分数在 0.68 的基础上提高了 0.06，达到 0.74。请注意，吹口哨在我们的数据集中并不常见，因此平滑的效果有限。

*6.1.4* 现实场景中的声学噪声鲁棒性：我们进一步研究了我们技术在现实场景中对噪声的鲁棒性。如§5 所述，我们考虑了四个地点：休息室、办公室、汽车和咖啡馆，这些地点具有不同的噪声特征，如表 2 所示，并比较了在这些地点的检测性能。我们通过将GrooveMeter与上文提到的两个变体进行比较来分解性能：YAMNet标签映射和GrooveMeter-w/o- 平滑。在此，我们纳入早期过滤操作，以检验总体性能。

图20展示了人声反应和无反应的宏平均 $F_1$ 分数。结果表明，无论四个场所的噪声特征如何，GrooveMeter都具有稳健的检测性能。与YAMNet 标签映射（基于YAMNet的分类和标签映射）相比，通过采用利用音乐信息的校正（GrooveMeter无平滑）和平滑（GrooveMeter），GrooveMeter在休息室、办公室、汽车和咖啡馆中的 $F_1$ 分数分别提高了0.08、0.09、0.12和0.21。由于背景噪声导致的假阳性相对较多，校正操作确实对咖啡馆贡献较大（ $F_1$ 分数提高了0.04）。然而，有趣的是，通过利用反应标签的时间关联，平滑操作显示出显著的改进（ $F_1$ 分数进一步提高了0.17），从而实现了与其他场所相当的性能。

### 6.2 动作反应检测

我们展示了动作反应检测的性能。对于验证，我们使用 MusicReactionSet 数据集进行了留一受试者交叉验证。我们参考以往研究，实现了以下三种基线。




(a) 与基线的比较

(b) 精确率/召回率（GrooveMeter）

(c) 各音乐流派的结果（GrooveMeter）

图21. 动作反应检测性能

• 随机森林：它表示用于识别重复性和周期性身体活动（例如， [9, 34, 45]）的感知管线。这些管线通常由特征提取器和机器学习分类器组成。我们使用来自IMU数据的时间域和频域统计特征 [21] 以及基于自相关提取的特征 [9, 34] 。我们使用支持向量机（SVM）、逻辑回归、决策树和随机森林等常用分类器测试了性能，并选择了优于其余分类器的随机森林模型。

• 卷积神经网络：这是一种基于深度学习的人体活动识别方法，例如 [54], ，它使用来自惯性测量单元传感器数据流的三轴加速度计和三轴陀螺仪数据。我们构建了一个分类模型，包含3个带 ReLU 激活函数的卷积层、一个最大池化层、两个丢弃率为0.5的丢弃层，以及一个 softmax层。

• 卷积长短期记忆网络：这是一种基于深度学习的方法，将卷积层与长短期记忆网络循环层相结合，用于通过可穿戴设备进行活动识别，例如 [48]。它使用 convLSTM [57] ，并采用来自惯性测量单元传感器数据流的三轴加速度计和三轴陀螺仪数据。我们构建了一个分类模型，包含一个卷积长短期记忆网络层、一个丢弃率为0.5的丢弃层、一个 ReLU层和一个 softmax层。

*6.2.1* 总体性能。 图 21a 展示了 GrooveMeter 与各基线方法的性能对比。结果表明，GrooveMeter 检测头部运动的准确度高于各基线方法。尽管无反应类别的 $F_1$ 分数差异不大，但与各基线方法相比，GrooveMeter 的头部运动类别 $F_1$ 分数平均提高了 0.09。对于头部运动，GrooveMeter 的 $F_1$ 分数为 0.74，而随机森林、卷积神经网络和卷积长短期记忆网络的分数分别为 0.60、0.71 和 0.65。有人可能会认为，GrooveMeter 相较于卷积神经网络的性能提升（提高 0.03）微乎其微。然而，我们发现 GrooveMeter 在动作噪声环境中更加稳健。我们将在 §6.2.2 中深入分析。

随后，我们进一步深入考察GrooveMeter的性能。图 21b 显示，即使对于一名未见过的用户，我们的方法在动作反应检测方面仍表现合理。具体而言，它以 0.72 的精确率和 0.75 的召回率检测 头部运动 。对于无反应，其精确率达到 0.88，召回率达到 0.87。

我们研究了不同音乐流派下的结果（见图21c）。请注意，我们将舞曲、快节奏嘻哈/说唱和快节奏摇滚归类为快节奏/振奋人心的歌曲，将抒情歌曲、节奏布鲁斯/灵魂乐、慢节奏摇滚和民谣/布鲁斯归类为慢节奏/柔和歌曲。振奋人心/快节奏歌曲的结果显示，其反应的精确率和召回率均高于柔和/慢节奏歌曲，因此得分相对较高$F_1$ 。具体而言，头部运动和无反应的 $F_1$ 得分分别为0.75和0.85，

表 3。与基线的比较

| | 受控 | MusicReactionSet | | | | |
|---|---|---|---|---|---|---|
| | | 休息室 | 办公室 | Car | Cafe | Avg. |
| **头部运动** | | | | | | |
| 随机森林 | 0.96 | 0.61 | 0.54 | 0.67 | 0.59 | 0.60 |
| CNN | 0.93 | 0.74 | 0.68 | 0.75 | 0.66 | 0.71 |
| 卷积长短期记忆网络 | 0.97 | 0.66 | 0.57 | 0.73 | 0.64 | 0.65 |
| GrooveMeter | 0.94 | 0.75 | 0.72 | 0.75 | 0.72 | 0.74 |
| **无反应** | | | | | | |
| 随机森林 | 0.92 | 0.81 | 0.90 | 0.80 | 0.86 | 0.84 |
| CNN | 0.85 | 0.84 | 0.92 | 0.86 | 0.89 | 0.89 |
| 卷积长短期记忆网络 | 0.92 | 0.81 | 0.89 | 0.86 | 0.89 | 0.86 |
| GrooveMeter | 0.86 | 0.84 | 0.91 | 0.85 | 0.90 | 0.88 |

对于激昂/快节奏歌曲，相应数值分别为 0.70 和 0.89。聆听快节奏歌曲时，人们往往会大幅度点头。因此，动作反应呈现出更明显的信号和清晰的周期性，从而取得更好的表现。相比之下，聆听轻柔/慢节奏歌曲时，动作反应往往较弱，动作轨迹也较小。因此，与快节奏歌曲的情况相比，更多反应数据可能会被误认为是非反应动作。

*6.2.2* 与基线的深入比较。为了更好地了解与基线之间的性能差异，我们还在受控实验室环境中收集了受控数据集。我们从大学校园招募了10名参与者（男性：6名，女性：4名），$P_B1$ 至 $P_B10$；他们的年龄为20至26岁（平均：23.1岁）。他们均自愿参与，并获得了价值9美元的礼品卡作为补偿。我们邀请每位参与者来到实验室，并要求其遵循一个包含3个环节的指定场景：2个环节用于收集音乐聆听反应，1个环节用于其他活动。在前2个环节中，我们提供了音乐流媒体服务中的音乐排行榜前100名，并让他们在每个环节中自由选择一首歌曲进行聆听。随后，他们被要求自然地做出给定的反应，但要持续60到90秒。最后一个环节的任务是在实验室内自由走动，并在聆听音乐的同时做出与音乐无关的动作，这代表无反应。

表 3 分别展示了受控数据集和 MusicReactionSet 数据集中的头部运动和无反应 $F_1$ 分数。我们首先考察头部运动的检测性能。有趣的是，尽管包括 GrooveMeter 在内的所有方法在受控数据集中的表现相近，但在 MusicReactionSet 数据中（该数据是在自然环境中采集的），性能差异十分明显。例如，在受控数据集中，头部运动的 $F_1$ 分数均超过 0.93。在 MusicReactionSet 中，由于存在各种动作噪声和多样的动作反应模式，$F_1$ 分数通常会下降，但与基线相比，GrooveMeter 的差距较小。

我们进一步考察了不同场景下的头部运动检测。一个有趣的观察是，尽管 GrooveMeter 在各个场景中的表现相近，但如表 2 所述，在观察到更多动作噪声的办公室和咖啡馆场景中，所有基线的性能均明显下降。例如，CNN 对头部运动的平均 $F_1$ 分数（0.71）与 GrooveMeter（0.74）相近，但与休息室和车内场景相比，其在办公室和咖啡馆场景中的 $F_1$ 分数下降幅度为 0.06 至 0.09。相比之下，GrooveMeter 仅下降 0.03，这表明 GrooveMeter 对日常动作噪声具有更强的鲁棒性。

我们考察了无反应类别的检测性能。在受控数据中，CNN 和 GrooveMeter 的 $F_1$ 分数明显低于随机森林和卷积长短期记忆网络，但它们在 theMusicReactionSet 上的表现更好。在 InMusicReactionSet 中，对于无反应类别，$F_1$ 分数

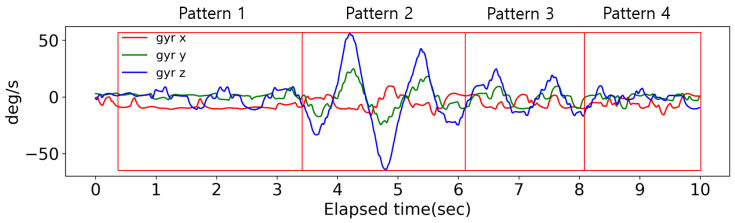



图22.陀螺仪数据与头部运动，持续10秒。模式 1:非常缓慢地摇摆头部。模式 2-4:用力点头，然后逐渐变弱

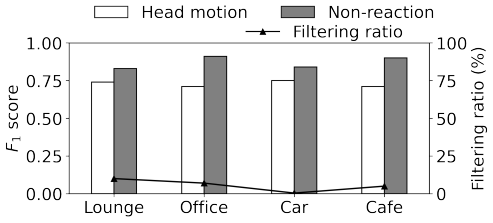



图23. 不同场所中活动的影响

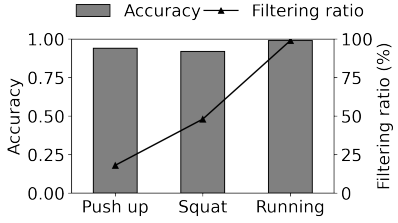



图24。对其他活动的鲁棒性

CNN 和 GrooveMeter 的分数分别为 0.89 和 0.88，而随机森林和卷积长短期记忆网络的分数分别为 0.84 和 0.86。我们推测，这是因为在受控数据中，头部运动与无反应片段具有明显可区分的模式。然而，在 MotionReactionSet 中，由于自然头部运动和日常动作噪声呈现多种模式，存在更多容易混淆的情况，从而增加了随机森林和卷积长短期记忆网络的假阳性错误。

我们展示了一个具有代表性的示例，说明 GrooveMeter 为何比基线在头部运动检测方面更准确、更稳健。图 22 展示了持续进行头部运动时长达 10 秒的陀螺仪信号。如图所示，即使在一次头部运动过程中，运动模式（即相应的信号模式）也并非规律或均匀重复；头部运动的方向和力度会自然地在较短时间间隔内发生变化。现有的运动感知流程（例如随机森林）可能无法检测此类情况，因为它们通常假设统计特征在窗口大小（例如 7 秒）的时间范围内相似。另一方面，GrooveMeter 利用长短期记忆网络分类器，通过运动单元（100 毫秒）的时间序列进行头部运动检测，因此能够处理各种头部运动模式。

*6.2.3* 活动的影响。我们根据活动考察动作反应检测的性能。如第 §5 节所述，参与者在不同地点进行不同的活动，即在休息室休息、在办公室工作、乘车以及在咖啡馆放松，这些活动会引发各种与反应无关的运动。

图 23 展示了四种情境下宏平均的 $F_1$ 分数和过滤率。无论进行何种活动，GrooveMeter 的表现都相近。办公室和咖啡馆情境下头部运动的 $F_1$ 分数略低于其他情境。相反，无反应的 $F_1$ 分数则呈现相反趋势。各情境的过滤率各不相同。休息室的过滤率为 10%，高于其他情境。汽车情境下的过滤率仅为 0.4%，因为汽车行驶产生的动作使大多数无反应数据都处于过滤范围内。

表 4. 系统成本；时间（毫秒），功耗（毫瓦）。

| | 操作 | GALAXY S21 | | GALAXY S8+ | |
|---|---|---|---|---|---|
| | | Time | 功率 | Time | 功率 |
| 发声系统 | 过滤 | 1.2 | 247.0 | 3.4 | 248.8 |
| | 分类 | 14.1 | 417.2 | 39.4 | 630.9 |
| | 校正 | 1.4 | 261.9 | 10.9 | 275.8 |
| | 平滑 | 0.1 | 27.1 | 0.3 | 152.0 |
| 动作 | 过滤 | 0.5 | 27.9 | 1.2 | 28.4 |
| | 分类 | 12.4 | 92.62 | 34.6 | 82.3 |
| 能量 | 全部（不含过滤） | 7.3 毫焦/秒 | | 70.4 毫焦/秒 | |
| | 全部（含过滤） | 3.7 毫焦耳/秒 | | 28.5 毫焦耳/秒 | |

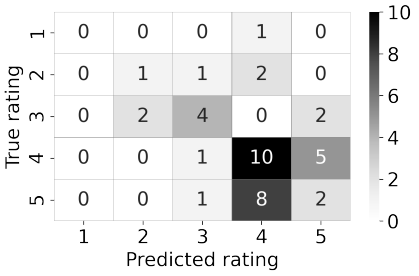



图25。评分预测的混淆矩阵。

*6.2.4*对其他活动的稳健性。我们进一步研究了现实生活场景中的稳健性，在这些场景中，人们经常会在进行锻炼等其他活动时听音乐。为此，我们按照§5中所述的相同方式，额外收集了5名参与者的IMU数据。他们在聆听自己喜欢的三首歌曲时，进行了三项活动，即跑步、做俯卧撑和做深蹲。他们在锻炼过程中没有做出任何反应。为检验检测性能，我们仅使用§5中包含30名参与者的数据集训练分类模型，并将其他活动数据用于测试。由于这里只包含无反应数据，因此我们使用准确率。图24显示，GrooveMeter能够稳健地应对三项活动中涉及的动作。对于俯卧撑、深蹲和跑步，准确率分别为0.94、0.92和0.99。尽管这些活动表现出周期性动作，GrooveMeter仍能准确区分无反应。此外，除俯卧撑情况外，过滤率都很高，分别为48%和99%，显著减少了处理开销。即便如此，俯卧撑情况下的过滤率也达到了18%，能够减少相当可观的处理成本。

### 6.3 系统成本

我们从执行时间和能耗开销两个方面考察 GrooveMeter 的系统成本。表4展示了其在两部 Android 手机 Galaxy S21 和 Galaxy S8+ 上的结果。在 Galaxy S21 上，语音和动作反应检测组件的端到端执行时间分别约为 17 毫秒和 13 毫秒，适合支持交互式应用。即使在低端手机 Galaxy S8+ 上，总耗时仍低于 90 毫秒。请注意，平均延迟短于总执行时间，这是因为由于过滤操作，分类和校正操作并不总是会执行。

我们使用 Monsoon 功率监测仪测量能耗。手机通过蓝牙在 Apple AirPods 上播放音乐，并测量执行操作所带来的净功率增量。尽管执行期间的瞬时功率较高，但包含过滤操作在内的总能耗仍然很小，即在 Galaxy S21 和 Galaxy S8+ 上分别为 3.7 毫焦耳/秒和 28.5 毫焦耳/秒。这是由于：a) 能耗可以搭载音乐播放器对中央处理器和通信功能的使用；以及 b) 过滤操作。此外，考虑到 GrooveMeter 仅在用户听音乐时运行，其能耗开销几乎不会影响电池续航。

此外，我们还考察了 eSense 耳塞式耳机的能耗。由于 eSense 未提供读取设备电池电量的应用程序接口，我们使用 Android 提供的电池电量信息。该信息的精度较粗，分辨率为 10%。我们的测量结果显示，使用 eSense 听音乐并进行惯性测量单元传感和音频录制，持续一小时所消耗的电池电量比

仅使用 eSense 听音乐多 20%。请注意，优化耳塞式耳机的能耗是我们未来的研究方向。

### 6.4 应用案例研究

我们通过开展应用案例研究，展示 GrooveMeter 的潜力。需要说明的是，我们既不讨论这些技术的总体性能，也不主张其具有新颖性。我们考察在有趣的应用场景下，音乐聆听反应信息的可行性与潜力。

*6.4.1* 体验采集。在这项研究中，我们进一步招募了十名参与者（$P_C1$ 至 $P_C10$），并请他们聆听八首歌曲；其中四首从排行榜前 50 名中随机选取（他们此前从未听过），另外四首则从参与者的播放列表中选取。

**自动音乐评分：**我们研究基于发声和动作反应的自动音乐评分的可行性。在这项研究中，我们请参与者为他们聆听的八首歌曲打分，评分采用 5 分制；5 分表示“我非常喜欢这首歌。”。为了构建评分预测模型，我们从发声和动作反应检测结果中提取统计特征，即对每种反应标签的持续时间和次数进行归一化，并使用决策树作为分类器。我们采用留一受试者交叉验证来评估评分预测性能。训练和测试时，我们仅使用排行榜前 50 名中参与者不熟悉的歌曲数据，原因有二。首先，熟悉歌曲的评分偏向 4 分和 5 分，因为参与者从自己的播放列表中挑选了熟悉的歌曲，而这些歌曲可能是他们最喜欢的歌曲。其次，评分预测对不熟悉的歌曲更有用，因为熟悉歌曲的评分很可能已经存在。图 25 展示了我们的评分预测混淆矩阵。对角线附近的数值较高，表明基于反应预测的评分与实际评分较为接近（MAE：0.22）。

**熟悉度检测：** 我们研究了是否可以利用音乐聆听反应来检测对一首歌曲的熟悉程度，即检测用户之前是否听过某首歌曲。这一功能将帮助音乐流媒体服务在不明确询问用户之前喜欢过哪些歌曲的情况下，加快建立新订阅用户的音乐偏好。我们使用人声和动作反应事件的统计特征训练了一个决策树模型。与自动音乐评分类似，我们以LOSO方式验证其性能，但使用的是完整数据集。对于已知歌曲和未知歌曲的检测，$F_1$得分分别为0.78（精确率：0.85，召回率：0.72）和0.81（精确率：0.76，召回率：0.88）。

*6.4.2* 小规模部署体验。 我们开展了额外的实验，以观察GrooveMeter在非受控、自然的音乐聆听情境下的检测性能和推荐可行性。为开展这项研究，我们开发了一个具有两项功能的应用程序：一项用于调查反应检测的性能，另一项用于研究基于反应的音乐推荐，分别如图1b和1c所示。我们招募了三名参与者（$P_D1$至 $P_D3$），并在他们的Android手机上安装了GrooveMeter和这些应用程序。他们自由使用GrooveMeter听音乐一天。他们平均聆听了21.6首歌曲。随后，我们开展了半结构化的一对一访谈。

**基于反应的音乐推荐：** 播放一首歌曲后，应用会计算反应模式，并从候选曲目池中推荐一首反应模式最相似的新歌曲；我们使用 MusicReactionSet 中的 240 次音乐聆听会话作为候选曲目池。我们将标签映射为反应索引（无反应：0，唱歌：1，吹口哨：2，动作：3），并将反应模式定义为由 1 秒长的索引组成的序列，再使用动态时间规整计算相似度。总体而言，参与者喜欢这些推荐歌曲，尽管他们从未听过这些歌曲。有趣的是，他们提到，听推荐歌曲时，他们做出的反应模式与听上一首歌曲时相似。 $P_D1$ 说，"它推荐的都是我喜欢的风格的歌曲。..... 我

我想我也做出了类似的反应。。 $P_D3$ 表示：“我实际听了之后，喜欢这个推荐。。共同点是，他们都认为基于反应的音乐推荐很有用。

**真实场景中的反应检测：**为考察真实场景中的反应检测，应用会在歌曲播放后显示如图 1b 所示的反应时间序列输出。在访谈中，我们询问了他们对反应输出的印象和看法。请注意，由于缺少用于获取真实数据的摄像头，这是一项主观、粗粒度的评估，但旨在评估用户在完全不受控环境中自然产生的反应。所有参与者都表示，GrooveMeter 根据他们的回忆给出了合理的反应输出。$P_D2$ 提到：“我觉得相当准确。”。我们请他们以 5 分制评价感知到的检测准确度；5 分表示“总结看起来非常准确”，平均得分为 4 分。

**建议：**参与者根据自身体验提出了 GrooveMeter 的扩展功能建议。$P_D2$和 $P_D3$提到，GrooveMeter 可以通过指出听众喜欢并对此作出反应的歌曲片段，帮助音乐作曲者。$P_D2$ 补充道：“如果它能显示观众何时情绪高涨，即使在音乐会上也会很有用。”

虽然当前研究展示了实际部署应用场景的潜力，但参与者数量和使用时长有限。未来，我们将开展更全面、周期更长的用户研究。

## 7 讨论

**对音乐的其他类型反应：**根据现有的音乐心理学文献，[25],人们在听音乐时通常会产生多种反应。除了身体反应外，这些反应还包括生理反应（例如心率、皮肤电导和生化反应）以及情绪反应（例如主观感受和情绪表达）。在这项工作中，我们重点关注身体反应，并以我们在真实环境数据集中常见的发声和动作反应为目标。通过为耳塞式耳机增添其他传感模态（例如光电容积描记法（PPG）[20],皮肤电导（EDA）[50],）并采用相关识别模型，GrooveMeter 可扩展至检测生理和情绪反应。此外，还可以区分其他类型的身体反应，例如跺脚和打响指。这将作为我们未来的工作。

**基于阈值的操作：**GrooveMeter 包含多个基于阈值开展工作的组件，例如过滤的阈值范围和用于校正的相似度阈值。出于以下原因，我们选择了这类简单方法，而不是基于复杂机器学习（ML）的操作。首先，根据我们的初步分析，基于阈值的操作与基于机器学习的操作表现相当，同时无需承担模型训练和执行的成本。其次，阈值以保守的方式设定，例如，过滤操作的阈值范围被设定为尽可能纳入更多反应片段，因此错误可能会导致系统成本增加，但几乎不会影响检测准确率。不过，我们承认，基于阈值的静态操作对于多样化的现实场景可能不够稳健。作为未来工作，我们将进一步研究这两种方法，并探索在线适应机制，使其即使在未见过的环境中也能保持稳健，即通过反映运行时数据，动态调整阈值或机器学习模型。

**数据集：**我们的数据集包含 30 名参与者在四种人们经常享受音乐的现实场景中进行的 240 次音乐聆听会话。它有助于系统地了解音乐聆听反应的信号特征，但在周围环境和参与者群体方面仍存在局限。现实场景中的音乐聆听情境可能多种多样，影响检测性能的因素也可能有所不同，例如聆听者的反应模式、移动模式和噪声条件。值得收集涵盖多种条件的更大规模数据集，并探索如何提高模型的鲁棒性。

**情境依赖分析：**当人们在进行其他主要活动（例如学习、阅读、锻炼）时听音乐，可能不如将听音乐作为主要活动时反应强烈。此外，即使听的是同一首歌曲，对歌曲的反应也可能随时间变化，或因具体情境而异。提供有关用户情境的更多信息，例如活动、时间和地点，将会有所帮助。此外，情境依赖分析未来将成为 GrooveMeter 的一项高级功能。

**超越音乐聆听反应：**耳塞式耳机主要用于聆听音乐，但也广泛应用于许多日常生活场景，例如线上会议、线上讲座，甚至面对面交谈中的实时翻译。然而，耳塞式耳机目前仍局限于音频流媒体，无法理解用户在将耳塞式耳机用于其他用途时对对话对象做出的反应。我们设想，GrooveMeter可以进一步扩展，以检测不同情境下的反应，并丰富这些应用的情境感知能力。例如，GrooveMeter可以检测学生在线上讲座中表现出的非语言和行为线索，帮助教师即时掌握学生的学习状态。同样，翻译应用可以根据说话者或对话对象的反应，即时调整翻译结果。

## 8 结论

我们提出了GrooveMeter，这是一种通过耳戴式感测自动检测音乐听众发声和动作反应的新型系统。我们设计了复杂的处理流程，使反应检测更加准确、稳健且高效。我们通过包含30位参与者在日常听音乐情境下产生的总时长926 分钟的惯性测量单元和音频数据的数据集，开展了广泛的实验来评估 GrooveMeter。我们还通过一项案例研究展示了其用处。

**REFERENCES**

[1] [n.d.]. The Future Of Live Events: AR, VR, And Advertising. https://www.forbes.com/sites/legalentertainment/2021/ 02/09/the-future-of-live-events-ar-vr-and-advertising/. 访问日期：2022年11月13日。[2] [n.d.]. How Does Spotify Know You So Well? https://medium.com/s/story/spotifys-discover-weekly-how-machine-learning-finds-your-new-music-19a41ab76efe. 访问日期：2022年11月13日。[3] [n.d.]. Music Listening 2021. https://www.ifpi.org/wp-content/uploads/2021/10/IFPI-Engaging-with-Music-report.pdf. 访问日期：2022年11月13日。[4] [n.d.]. Tech Tuesdays: The Secret A.I. Recommender Systems Behind Spotify's Success. https://bennettdatascience. com/tech-tuesdays-the-secret-ai-recommender-systems-behind-spotifys-success/. 访问日期：2022年11月13日。[5] [n.d.]. TensorFlow Lite. https://www.tensorflow.org/lite. 访问日期：2022年11月13日。[6] Yudhik Agrawal、Samyak Jain、Emily Carlson、Petri Toiviainen 和 Vinoo Alluri。2020。迈向多模态音乐信息检索：根据音乐诱发的动作预测个体差异。*arXiv* 预印本 *arXiv:2007.10695* (2020)。[7] Xuan Bao、Songchun Fan、Alexander Varshavsky、Kevin Li 和 Romit Roy Choudhury。2013。你的反应表明你喜欢这部电影：通过反应感测自动进行内容评级。载于*2013*年美国计算机协会普适与泛在计算国际联合会议论文集。197–206。[8] Oswald Barral、Ilkka Kosunen 和 Giulio Jacucci。2017。不必开怀大笑：在真实环境中基于生理信号预测对连环漫画的幽默评价。《*ACM*人机交互汇刊》*(TOCHI)* 24, 6 (2017), 1–29。[9] Abdelkareem Bedri、Richard Li、Malcolm Haynes、Raj Prateek Kosaraju、Ishaan Grover、Temiloluwa Prioleau、Min Yan Beh、Mayank Goel、Thad Starner 和 Gregory Abowd。2017。EarBit：使用可穿戴传感器在不受限制的环境中检测进食事件。《*ACM*交互式、移动、可穿戴与泛在技术论文集》1, 3 2017 1 20( ), – 。[10] Anne J Blood 和 Robert J Zatorre。2001。对音乐产生的强烈愉悦反应与奖赏和情绪相关脑区的活动有关。《美国国家科学院院刊》98, 20 (2001), 11818–11823。[11] Lars Ole Bonde 和 Tony Wigram。2002。音乐治疗综合指南：理论、临床实践、研究与培训。Jessica Kingsley出版社。[12] Nam Bui、Nhat Pham、Jessica Jacqueline Barnitz、Zhanan Zou、Phuc Nguyen、Hoang Truong、Taeho Kim、Nicholas Farrow、Anh Nguyen、Jianliang Xiao 等。2019。eBP：一种从用户耳部进行频繁且舒适血压监测的可穿戴系统。载于第*25*届年度移动计算与

网络通信国际会议论文集。1–17。[13] Birgitta Burger、Marc R Thompson、Geoff Luck、Suvi Saarikallio 和 Petri Toiviainen。2013。节奏和音色相关音乐特征对音乐诱发动作特征的影响。《心理学前沿》4 (2013), 183。[14] Birgitta Burger、Marc R Thompson、Geoff Luck、Suvi H Saarikallio 和 Petri Toiviainen。2014。追寻身体中的节拍：音乐诱发动作中的周期与相位锁定。《人类神经科学前沿》8 (2014), 903。[15] Gaoshuai Cao、Kuang Yuan、Jie Xiong、Panlong Yang、Yubo Yan、Hao Zhou 和 Xiang-Yang Li。2020。EarphoneTrack：将耳机纳入声学动作追踪生态系统。载于第18届嵌入式网络传感器系统会议论文集。95–108。[16] Yetong Cao、Huijie Chen、Fan Li 和 Yu Wang。2021。CanalScan：通过耳道形变感测识别舌颌动作。载于IEEE INFOCOM 2021——IEEE通信会议。IEEE，1–10。[17] Seokmin Choi、Yang Gao、Yincheng Jin、Se jun Kim、Jiyang Li、Wenyao Xu 和 Zhanpeng Jin。2022。PPGface：喜欢你正在看的内容吗？耳机也能“感知”你的面部表情。《ACM交互、移动、可穿戴与泛在技术》6, 2 (2022), 1–32。[18] Andrea Ferlini、Dong Ma、Robert Harle 和 Cecilia Mascolo。2021。EarGate：利用入耳式麦克风进行基于步态的用户识别。载于第27届年度移动计算与网络国际会议论文集。337–349。[19] Andrea Ferlini、Alessandro Montanari、Cecilia Mascolo 和 Robert Harle。2019。通过入耳式可穿戴设备追踪头部运动。载于第1届耳戴计算国际研讨会论文集。8–13。[20] Andrea Ferlini、Alessandro Montanari、Chulhong Min、Hongwei Li、Ugo Sassi 和 Fahim Kawsar。2022。用于生命体征监测的入耳式PPG。《IEEE普适计算》21, 1 (2022), 65–74。[21] Davide Figo、Pedro C Diniz、Diogo R Ferreira 和 Joao MP Cardoso。2010。利用加速度计数据进行情境识别的预处理技术。《个人与泛在计算》14, 7 (2010), 645–662。[22] G David Forney。1973。维特比算法。《IEEE学报》61, 3 (1973), 268–278。[23] Rolf Inge Godøy 和 Marc Leman。2010。音乐手势：声音、动作与意义。Routledge。[24] Victor E Gonzalez-Sanchez、Agata Zelechowska 和 Alexander Refsum Jensenius。2018。静止状态下音乐与非自主人体微动作之间的对应关系。《心理学前沿》9 (2018), 1382。[25] Susan Hallam、Ian Cross 和 Michael Thaut。2016。《牛津音乐心理学手册》（第2版）。牛津大学出版社。[26] Yan He、George Chernyshov、Jiawen Han、Dingding Zheng、Ragnar Thomsen、Danny Hynds、Muyu Liu、Yuehui Yang、Yulan Ju、Yun Suen Pai 等。2022。Frisson Waves：探索音乐表演中审美颤栗的自动检测、触发与分享。《ACM交互、移动、可穿戴与泛在技术》6, 3 (2022), 1–23。[27] Romain Hennequin、Anis Khlif、Felix Voituret 和 Manuel Moussallam。2020。Spleeter：一种基于预训练模型的快速高效音乐源分离工具。《开放源代码软件期刊》5, 50 (2020), 2154。https://doi.org/10. 21105/joss.02154 Deezer Research。[28] Shawn Hershey、Sourish Chaudhuri、Daniel P. W. Ellis、Jort F. Gemmeke、Aren J ansen、Channing Moore、Manoj Plakal、Devin Platt、Rif A. Saurous、Bryan Seybold、Malcolm Slaney、Ron Weiss 和 Kevin Wilson。2017。面向大规模音频分类的卷积神经网络架构。载于国际声学、语音与信号处理会议（ICASSP）。https://arxiv.org/abs/ 1609 09430 。[29] Nan Jiang、Terence Sim 和 Jun Han。2022。EarWalk：探索利用耳戴设备识别行走姿势。载于第23届年度移动计算系统与应用国际研讨会论文集。35–40。[30] Yincheng Jin、Yang Gao、Xiaotao Guo、Jun Wen、Zhengxiong Li 和 Zhanpeng Jin。2022。EarHealth：一种基于耳机的声学耳镜，可在日常生活中检测多种耳部疾病。载于第20届年度移动系统、应用与服务国际会议论文集。397–408。[31] Pragma Kar、Samiran Chattopadhyay 和 Sandip Chakraborty。2020。Gestatten：通过眼动注视与注视手势追踪，估算用户在移动慕课中的注意力。《ACM人机交互论文集》4, EICS (2020), 1–32。[32] Fahim Kawsar、Chulhong Min、Akhil Mathur 和 Allesandro Montanari。2018。用于个人级行为分析的耳戴设备。《IEEE普适计算》17, 3 (2018), 83–89。[33] Prerna Khanna、Tanmay Srivastava、Shijia Pan、Shubham Jain 和 Phuc Nguyen。2021。JawSense：使用低成本耳戴系统识别无声发音。载于第22届移动计算系统与应用国际研讨会论文集。44–49。[34] Rushil Khurana、Karan Ahuja、Zac Yu、Jennifer Mankoff、Chris Harrison 和 Mayank Goel。2018。Gy mCam：在不受限制的场景中检测、识别和追踪同时进行的锻炼。《ACM交互、移动、可穿戴与泛在技术》2, 4 (2018), 1–17。[35] Jong Wook Kim、Justin Salamon、Peter Li 和 Juan Pablo Bello。2018。CREPE：一种用于音高估计的卷积表征。arXiv:1802.06182 [eess.AS]

[36] Stefan Koelsch、Thomas Fritz、D Yves v. Cramon、Karsten Müller 和 Angela D Friederici。2006。借助音乐研究情绪：一项功能性磁共振成像研究。《人脑图谱》27卷3期（2006年），239–250页。[37] Seungchul Lee、Chulhong Min、Alessandro Montanari、Akhil Mathur、Youngjae Chang、Junehwa Song 和 Fahim Kawsar。2019。利用动觉耳戴设备自动识别微笑与皱眉。载于《2019年第10届增强人类国际会议论文集》。1–4页。[38] Daniel J Levitin、Jessica A Grahn 和 Justin London。2018。音乐心理学：节奏与动作。《心理学年度评论》69卷（2018年），51–75页。[39] Ke Li、Ruidong Zhang、Bo Liang、François Guimbretière 和 Cheng Zhang。2022。Eario：一种低功耗声学感知耳戴设备，可持续追踪细致的面部动作。《ACM交互式、移动、可穿戴与泛在技术论文集》6卷2期（2022年），1–24页。[40] Geoff Luck、Suvi Saarikallio、Birgitta Burger、Marc R Thompson 和 Petri Toiviainen。2010。大五人格和音乐流派对音乐诱发动作的影响。《人格研究杂志》44卷6期（2010年），714–720页。[41] Emily Lynar、Erin Cvejic、Emery Schubert 和 Ute Vollmer-Conna。2017。动人心弦音乐的喜悦：对情绪和生理反应的考察。《国际心理生理学杂志》120卷（2017年），118–125页。[42] Dong Ma、Andrea Ferlini 和 Cecilia Mascolo。2021。OESense：利用堵耳效应进行入耳式人体感知。载于《第19届年度移动系统、应用与服务国际会议论文集》。175–187页。[43] Claudio Martella、Ekin Gedik、Laura Cabrera-Quiros、Gwenn Englebienne 和 Hayley Hung。2015。演出效果如何？利用智能手机感知测量观众对现场演出的隐性反应。载于《第23届美国计算机协会国际多媒体会议论文集》。201–210页。[44] Chulhong Min、Akhil Mathur 和 Fahim Kawsar。2018。探索耳戴设备上的音频与动觉感知。载于《第4届美国计算机协会可穿戴系统与应用研讨会论文集》（德国慕尼黑）（WearSys '18）。计算机协会，纽约，美国，

5 10 10 1145 3211960 3211970 – 。

[45] 2014 https://doi.org/ 。/ 。Dan Morris、T Scott Saponas、Andrew Guillory 和 Ilya Kelner。 。RecoFit：使用可穿戴传感器查找、识别并计数重复性练习。载于《SIGCHI计算机系统中的人为因素会议论文集》。3225–3234页。[46] Shahriar Nirjon、Robert F Dickerson、Qiang Li、Philip Asare、John A Stankovic、Dezhi Hong、Ben Zhang、Xiaofan Jiang、Guobin Shen 和 Feng Zhao。2012。Musicalheart：聆听音乐的贴心方式。载于《第10届美国计算机协会嵌入式网络传感器系统会议论文集》。43–56页。[47] Adrian C North、David J Hargreaves 和 Jon J Hargreaves。2004。音乐在日常生活中的用途。《音乐感知》22卷1期（2004年），41–77页。[48] Francisco Javier Ordóñez 和 Daniel Roggen。2016。用于多模态可穿戴活动识别的深度卷积神经网络和长短期记忆循环神经网络。《传感器》16卷1期（2016年），115页。[49] Alisun Pawley 和 Daniel Müllensiefen。2012。跟唱的科学：一项关于英格兰北部跟唱行为的定量实地研究。《音乐感知》30卷2期（2012年），129–146页。[50] Nhat Pham、Tuan Dinh、Taeho Kim、Zohreh Raghebi、Nam Bui、Hoang Truong、Tuan Nguyen、Farnoush Banaei-Kashani、Ann Halbower、Thang N Dinh 等。2021。利用耳后可穿戴系统检测微睡眠事件。《IEEE移动计算汇刊》（2021年）。[51] Nhat Pham、Tuan Dinh、Zohreh Raghebi、Taeho Kim、Nam Bui、Phuc Nguyen、Hoang Truong、Farnoush Banaei-Kashani、Ann Halbower、Thang Dinh 等。2020。WAKE：一种用于微睡眠检测的耳后可穿戴系统。载于《第18届移动系统、应用与服务国际会议论文集》。404–418页。[52] Meera Radhakrishnan、Darshana Rathnayake、Ong Koon Han、Inseok Hwang 和 Archan Misra。2020。ERICA：为自由重量训练实现实时错误检测与纠正性反馈。载于《第18届嵌入式网络传感器系统会议论文集》。558–571页。[53] Tobias Röddiger、Daniel Wolffram、David Laubenstein、Matthias Budde 和 Michael Beigl。2019。探索使用入耳式耳机惯性测量单元监测呼吸频率。载于《第1届耳戴计算国际研讨会论文集》。48–53页。[54] Charissa Ann Ronao 和 Sung-Bae Cho。2016。利用深度学习神经网络和智能手机传感器进行人体活动识别。《专家系统与应用》59卷（2016年），235–244页。[55] Tobias Scheffer、Christian Decomain 和 Stefan Wrobel。2001。用于信息抽取的主动隐马尔可夫模型。载于《智能数据分析进展》，Frank Hoffmann、David J. Hand、Niall Adams、Douglas Fisher 和 Gabriela Guimaraes 编。Springer Berlin Heidelberg，柏林、海德堡，309–318页。[56] Claude Elwood Shannon。1948。《通信的数学理论》。《贝尔系统技术杂志》27卷3期（1948年），379–423页。https://doi.org/10.1002/j.1538-7305.1948.tb01338.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x [57] Xingjian Shi、Zhourong Chen、Hao Wang、Dit-Yan Yeung、Wai-Kin Wong 和 Wang-chun Woo。2015。卷积长短期记忆网络：一种用于降水临近预报的机器学习方法。《神经信息处理系统进展》28卷（2015年）。

[58] Dhruv Verma、Sejal Bhalla、Dhruv Sahnan、Jainendra Shukla 和 Aman Parnami。2021。ExpressEar：利用耳戴设备感知细粒度面部表情。《*ACM*交互式、移动、可穿戴与泛在技术论文集》5卷3期（2021年），1–28页。[59] Sergey Volokhin 和 Eugene Agichtein。2018。了解日常活动中的音乐聆听意图及其对情境音乐推荐的启示。载于《*2018*年人类信息交互与检索会议论文集》。313–316页。[60] Chen Wang 和 Pablo Cesar。2017。这场演出很精彩：但你如何判断？载于《*2017*年美国计算机协会*SIGCHI*创造力与认知会议论文集》。336–347页。[61] Zi Wang、Sheng Tan、Linghan Zhang、Yili Ren、Zhi Wang 和 Jie Yang。2021。EarDynamic：基于耳道形变、利用入耳式可穿戴设备进行持续用户认证。《*ACM*交互式、移动、可穿戴与泛在技术论文集》5卷1期（2021年），1–27页。[62] Yadong Xie、Fan Li、Yue Wu、Huijie Chen、Zhiyuan Zhao 和 Yu Wang。2022。TeethPass：通过入耳式声学感知进行基于牙齿咬合的用户认证。载于《*IEEE INFOCOM 2022—IEEE*通信计算机会议》。IEEE。[63] Yi-Hsuan Yang 和 Yuan-Ching Teng。2015。智能手机情境下音乐聆听行为的定量研究。《美国计算机协会交互式智能系统汇刊（*TiiS*）5卷3期（2015年），1–30页。