利用语音打断分析提升会议包容性

Szu-Wei Fu
Szu
WeiFu@microsoft.com 微软
美国华盛顿州雷德蒙德

Yaran Fan
Yaran.Fan
@microsoft.com 微软 美国华盛顿州雷德蒙德

Yasaman Hosseinkashi
yahossei@microsoft.com 微软 美国华盛顿州雷德蒙德

Jayant Gupchup
Jayant.Gupchup@microsoft.com 微软 美国华盛顿州雷德蒙德

Ross Cutler
R
oss.Cutler@microsoft.com
微软 美国华盛顿州雷德蒙德

arXiv:2304.00658v2 [eess.AS] 5 Apr 2023

摘要

会议是各类公司和组织中普遍采用的沟通方式。自新冠肺炎疫情以来，使用远程协作系统召开会议的情况大幅增加。然而，并非所有会议都具有包容性，尤其是在与会者的参与率方面。微软最近开展的一项大规模调查显示，会议参与者提出的改善包容性的首要建议，是提升远程参与者在会议中打断他人并获得发言权的能力。我们发现，使用虚拟举手（VRH）功能可以提高微软会议包容性的预测值。一项挑战是，VRH 在全部会议中的使用率不到 1%。为了推动这项功能的使用，从而改善包容性（和参与度），我们提出了一种基于机器学习的系统，用于预测会议参与者尝试获得发言权却未能成功打断他人的情况（称为“失败的打断”）。该预测可用于提示用户在会议中举起虚拟手。我们认为，这是首个语音打断失败检测器；在一个贴近真实情况的测试集上，其曲线下面积（AUC）为 0.95，在假正率（FPR）为 $< 1\%$ 时，真正率（TPR）为 50%。据我们所知，这也是首个针对远程会议的打断类别数据集（包括失败的打断类别）。最后，我们认为这是首个旨在通过语音打断分析和主动干预来改善会议包容性的系统。

CCS 概念

• **计算方法** → **自然语言处理**；**语音识别**；• **以人为中心的计算** → **协作式交互**；• **信息系统** → **网络会议**。

允许免费复制本作品的全部或部分内容（无论是数字版还是纸质版），用于个人或课堂用途，前提是不得为盈利或商业利益而复制或传播，且每份副本首页均须附上本声明和完整引文。对于本作品中由作者以外人员拥有版权的部分，必须尊重其版权。注明出处后，可以摘录。若要以其他方式复制、再版、发布至服务器或分发至列表，须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。
*MM '22，2022*年*10*月*10*日至*14*日，葡萄牙里斯本 © 2022年版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548379

关键词

语音打断分析，会议包容性，远程协作，机器学习

ACM 参考格式：
Szu-Wei Fu、Yaran Fan、Yasaman Hosseinkashi、Jayant Gupchup 和 Ross Cutler。2022年。运用语音打断分析提升会议包容性。载于《第*30*届*ACM* 国际多媒体会议（*MM '22*）论文集》，*2022*年*10*月*10*日至*14*日，葡萄牙里斯本。ACM，美国纽约州纽约市，9页。https://doi.org/10.1145/3503161.3548379

1 引言

计算机媒介通信（CMC）系统，尤其是视频会议系统，改变了公司和组织召开会议的方式。自 COVID-19 疫情开始以来的两年间，随着许多人被迫居家办公，此类系统的使用量大幅增加。这些系统的一个常见目标是尽可能促成高效的会议 [10]。近来，另一个目标是尽可能促成更具包容性的会议，让每个人都感到自己融入其中。近期一项研究发现，用户提出的提升会议包容性的首要建议，是让远程会议参与者更方便地打断发言并参与会议讨论 [10]。该研究还利用会议遥测数据表明，用户参与时，会议更有可能具备包容性；而远程用户参与会议的可能性则低得多。

为提升会议包容性，我们着手解决一个首要问题：增强远程参与者在会议中打断他人并发言的能力。尽管大多数远程协作系统都提供虚拟举手功能来帮助解决这一问题（见图 1），但我们发现，在 Microsoft Teams 中，使用该功能的会议不到 1%。我们表明，虚拟举手功能一经使用，便能有效提升会议包容性。我们介绍了一种基于机器学习的语音分析模型，可检测打断失败，即有人尝试发言却未能获得发言权。随后，可利用该分类器提示用户举起虚拟手并发言，或通知会议主持人，以帮助确保参与者有机会在会议中发言。构建此类分类器的一大挑战，是缺乏有标注且具有代表性的数据集。我们使用了来自公开来源的 250 小时会议数据，并创建了 90 小时新数据，从而构建了具有代表性的训练集和测试集，其中包含 40,000 个片段。这些会议被切分为包含语音重叠的音频片段。为了

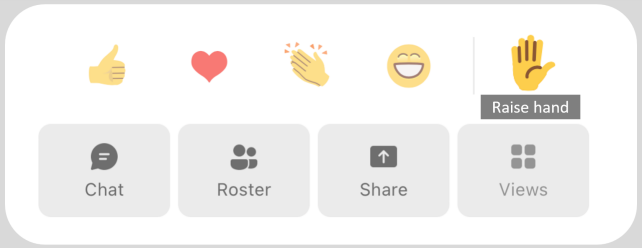


图 1：Microsoft Teams 中提供的虚拟举手（VRH）选项。

准确标注这些片段，我们开发了一套众包标注系统。我们对标注准确率进行了全面验证，这是确保模型达到预期准确率的关键步骤。据我们所知，这是首个对会议中不同类型打断进行分类的标注数据集。

拥有数百万用户的生产级计算机媒介沟通系统，对准确率有着非常严格的要求，以确保流畅的用户体验。根据我们的生产数据，对于有 4 位参与者、时长通常为半小时的会议，我们观察到 40 次打断。我们确定检测器所需的假阳性率为 $\leq 1\%$；这意味着普通用户平均每 10 场半小时会议会遇到一次误报。此外，我们发现打断失败占语音重叠的 10%，因此将真正率设为 $> 40\%$ 可确保平均每场会议至少捕捉到两次打断失败。要达到如此严格的标准，我们需要评估多种架构选择。打断失败通常表现为简短的语音话语。核心挑战在于区分这类事件与其他简短语音话语，例如赞同（如“是的”“有道理”）和回应语（如“嗯嗯”“嗯……嗯”）。鉴于标注片段数量很少（4 万个），从头开始训练模型是不可能的。因此，我们借助了自监督学习（SSL）表征方面的最新进展；这些表征是在数万小时的未标注数据上学习得到的（[3], WavLM [8]）。

本文的贡献如下：

- 我们将检测打断失败的问题形式化，并创建了首个已知数据集，用于对远程会议中的打断进行分类，标注准确率达到 95%。
- 据我们所知，我们提供了首个失败语音打断预测器。在真实的测试集中，该失败语音打断预测器的曲线下面积达到 0.95，真正率为 $> 50\%$，假阳性率为 $< 1\%$，足以用于生产级 CMC 系统。
- 我们认为，这是首个旨在通过语音打断分析提升会议包容性的功能性系统。

第 2 节回顾该领域的相关研究。第 3 节展示了使用 Microsoft Teams 中的虚拟举手功能时，包容性在统计上会提高。第 4 节介绍我们的模型所使用的训练数据集和测试集，第 5 节介绍模型架构和训练过程。第 6 节和第 7 节分别给出模型的结果及其一些可能的应用。第 8 节总结全文并讨论未来的工作。

2 相关工作

对会话过程进行刻画是一个经过充分研究的课题。Schegloff 详细阐述了“轮流发言”和“打断”之间的差异 [39]。该研究定义了打断，并列举了不应被视为打断的明确示例（例如，示意当前说话者继续发言，以及笑声、烦恼等手势）。文献 [37] 提出了一种有组织的轮流发言的形式化结构。该研究将“错误起始”定义为需要修复的轮流发言违规。这些修复规则旨在处理获取发言权的意图。然而，该研究假设修复机制会得到遵循（但实践中并未观察到这一点）。Margariti 等人对计算机媒介通信系统中的轮流发言进行了定量分析 [29]。他们发现，如果发生一次语音重叠，抢占发言的可能性与原发言者保持发言轮次的可能性大致相同。然而，当同一打断者多次尝试时，发生抢占发言的可能性会高得多。该研究没有区分失败的打断和插话（也称为“反馈语”）；二者都会导致抢占发言失败。

Sellen 研究了与同室对话相比，使用计算机媒介通信系统（仅音频和视频）对轮流发言和打断的影响 [40]。Sellen 的研究结果表明，同室条件下同时开始发言（导致打断失败）的发生率，与技术媒介解决方案中的发生率相同。该研究还指出，如果说话者在超过 1.5 秒内未被打断，即可认为其已获得发言轮次（或发言权）；这是一个有助于我们进行标注的定量定义。然而，Sellen 的研究假设网络和设备条件良好。Hammer 等人研究了网络延迟对网络电话系统中“对话交互性”的影响 [15]。该研究提出了主动打断和被动打断的概念。主动打断是有意为之的，而被动打断则是无意的，因延迟听到远端说话者的声音而发生。作者报告称，延迟对被动打断的影响远大于对主动打断的影响。延迟每增加 100 毫秒，被动打断率就相对上升 15 %。双讲场景产生的回声抑制伪影会导致衰减，使得在 C MC 系统中进行打断变得困难 [31]。

国际计算机科学研究所（ICSI）和 AMI 多方会议数据集为分析会议中的打断提供了良好的起点（ICSI [20], AMI [7]）。ICSI 数据集包含 75 场真实会议和 72 小时的语音。所有会议都有一个混合音频文件，以及每位说话者各自的音频文件（耳机、开放式麦克风等）。该数据集提供详细的词级转录，以及对打断和反馈语的标注。AMI 是一个时长 100 小时的会议语料库，记录了非母语说话者在三个具有不同声学特性的房间中的会议。ICSI 和 AMI 中的说话者在同一房间内参会时都佩戴耳机。虽然每位说话者都对应一个独立通道，但说话者之间的串音并未消除，因为与一位参与者关联的麦克风也会拾取另一位参与者的语音。仅远程会议不会有这个问题。为解决这种偏倚，我们使用自己的会议系统收集了仅远程参与者的数据。因此，我们主要使用这些数据来源来扩充训练数据；由于典型的仅远程会议场景

中不存在串音，因此我们不使用这些数据来构建测试集。这些数据集没有失败的打断标签。据我们所知，目前不存在能够捕捉失败打断类别的带标签音频数据集。在这项工作中，我们为众包工作者制定了详细说明，以便对不同的打断类型（包括打断失败）进行分类。

在 VoIP 应用中检测语音重叠，主要通过语音活动检测器（VAD）实现 [5, 6, 24, 43]。Baron 使用基于韵律的方法来预测打断和"插话点"。在预测插话点方面，基频和停顿特征的表现显著优于语言模型 [4]。Makhervaks等人构建了一个模型，用于检测会议中的"热点"（即兴趣高度集中的区域）[28]。该模型包含从语音活动、语言模型嵌入和韵律中提取的特征。诸如"嗯""是的"这样的回应（或赞同）在文献中通常称为"反馈语" [30, 36]。在 [36], Ruede等人使用词嵌入对反馈语进行分类。在 [23] Kennedy等人构建了一个基于支持向量的模型，利用梅尔频率倒谱系数（MFCC）对笑声进行分类。关键词检出模型是一种特殊的声学事件检测方法，旨在仅检测少量特定词语 [38, 42]。

据我们所知，目前尚无研究旨在对远程会议中的不同类型语音重叠进行分类。区分打断失败与反馈语是这一挑战的核心，因为这两类事件的声学特征相似，持续时间也非常接近。Yang等人和Fitzgerald等人构建了模型，用于检测他们称为"非流畅现象"的事件 [14, 26]。这些事件包括话语重复、修正和起始错误。在这两项研究中，研究者均假设这些非流畅现象得到了修正，因此不会导致打断失败。构建用于检测非流畅现象（即打断失败）的模型是一项复杂的工作。这项任务可以借鉴应用于其他类似下游任务（如关键词检测）的自监督方法 [8]。目前，已有越来越多基于原始音频数据训练的自监督学习表征（嵌入）（Wav2Vec [3], HuBERT [16], WavLM [8]）。在自动语音识别（ASR）这一类下游任务中，这些自监督方法均在前一代最先进水平的基础上取得了提升，展现了这些嵌入在提取语言上下文方面的有效性。在本研究中，我们将从原始音频生成的这些嵌入用作打断分类器的输入。

据我们所知，只有一项研究测量了会议包容性 [31]。作者在一家大型科技公司对 1.6 万名员工开展了调查（收到 3,300 份有效答复），并建立了一个多变量模型，以提取会议成效、包容性及其影响因素之间的关系。研究显示，该公司 80% 的会议具有包容性，表明仍有很大的改进空间。模型指出，参与是影响感知到的会议包容性的首要因素，其重要性比次重要因素高出 47%。此外，调查结果显示，参与者提出的改善会议包容性的首要功能需求是"让远程参与者更容易打断发言"。

性别一直被作为对话和参与中的一个影响因素进行研究。Eecke 等人发现，女性在对话中比男性更常被打断，而且男性打断女性的频率高于打断男性的频率 [12]。Leaper 等人 [25] 开展了一项荟萃分析，结果表明，在对话中，男性比女性说得更多。James 等人回顾了关于对话中性别偏见这一主题的大量研究 [19]。

3 虚拟举手与包容性分析

虚拟举手功能旨在促进远程参与者使用一款热门视频会议应用程序时的参与。本节旨在量化使用虚拟举手功能对会议包容性的影响。此处，使用虚拟举手功能是指会议期间任何参与者与该功能进行的任何交互（自行举手或放下手，或代他人举手或放下手）。我们通过比较使用和未使用虚拟举手功能的可比会议，来估算其影响。

在我们的系统中，所有会议的 VRH 使用信息均可通过通话技术遥测数据获取。会议包容性通过一个机器学习（ML）模型进行测量，该模型根据其余通话遥测数据（不包括 VRH 使用情况）预测包容性评分。与人工设置的研究小组不同，这种方法基于大量真实会议样本进行估算，因此具有更强的泛化能力。接下来，我们将介绍包容性评分预测模型的开发，以及该模型在估算 VRH 影响方面的应用。

3.1 包容性预测模型

来自 [31] 的初步调查结果促使我们开发了一款应用内通话结束问卷，用于衡量用户感知的会议效果和包容性，并在随机抽取的一组会议中进行调查。问卷包含两个 5 分制问题（3 分为中立）：1. 本次会议在实现目标方面的成效如何？2. 您觉得自己在会议中有多大程度的参与感?

应用内调查收集了超过 4 万条评分。将这些评分与会议遥测数据结合后，我们据此拟合了一个用于预测会议包容性和成效的模型。该模型使用 28 项工程化特征，这些特征源自通话质量、可靠性、会议规模、音频参与（通话期间发言）、视频或屏幕共享使用、会议时长和会议时间。通过 [31]中报告的图模型可以看出，所有这些特征对于解释评分差异都很重要。与 [31],类似，应用内数据也表明，参与度是预测和描述包容性的主要特征之一。

本研究用于预测的模型是一个轻量级 GBM [21] 二元分类器，用于生成用户对包容性问题给出 4 星或 5 星评分的概率。测试集上的表现通过 50 次随机训练集-测试集划分进行交叉验证来衡量，结果显示测试集的曲线下面积为 0.75+/-0.02。

为了预测没有用户评分的通话的用户评分，我们通过应用一个阈值，将预测概率转换为二元分数，以确保假阳性率不超过 5%。当实际包容性用户评分不是 4星或5星时，错误分类率即定义为假阳性率。在假阳性率为 5% 时，模型的真正率为 32%，利用现有遥测数据只能标记出 32% 的 4星或5星通话。因此，模型预测的包容性评分基线远低于实际用户评分（如果有的话）。

表1：按来源和标签划分的片段数量

数据来源	原始音频时长（小时）	标注者	反馈语	失败的打断	打断	笑声	其他	总计
AMI	100	众包	8,760	2,270	5,630	1,650	150	18,450
ICSI	72	众包	4,650	1,060	3,570	50	100	9,420
原始数据	77	众包	3,840	1,140	2,570	1,430	200	9,180
原始数据	15	专家	1,310	320	860	340	150	2,970
All	264	All	18,560	4,790	12,630	3,470	600	40,020

不过，该模型仍足够敏感，可以估算使用虚拟举手功能后预测分数的提升。下一节将对此进行讨论。

3.2 举手对预测值会议包容性的影响

我们比较了有和没有 VRH参与的会议的预测包容性分数。预测值是由上一节所述模型生成的二元值。由于 VRH 是所有用户都可使用的功能，且未进行任何受控实验 [22]，，因此无法开展因果分析。取而代之的是，我们采用倾向得分分层 [2] 作为一种伪实验方法。

在缺乏受控实验的情况下，倾向得分（PS）方法是估计因果关系的有效技术 [9, 17, 35]。 [9] 指出，如果恰当设定倾向得分，就能消除混杂因素造成的 90% 以上的偏倚。混杂因素是指数据中任何可能在处理组和对照组样本中分布不同的属性；此外，这些属性对结果变量的影响可能与结果变量的效应混杂在一起。在本研究中，结果变量是会议包容性，处理组为使用 VRH 的通话集合（称为 VRH 通话），对照组则为未参与 VRH 的通话集合。

在本研究中，主要的混杂因素包括会议规模、会议时长以及媒体选择（例如视频或屏幕共享）。例如，VRH 主要用于有 4 名或更多参与者、时长超过 30 分钟的会议。因此，VRH 通话中的会议规模和会议时长分布与平均通话显著不同。为减少这种偏倚，我们使用逻辑回归模型生成 PS 值，该模型根据混杂因素预测 VRH 活动。PS 值介于 0 和 1 之间，并用于生成五个大小相等的分箱。在每个分箱内，混杂因素的分布在统计上相似，从而提供可比的数据组来估计 VRH 的影响。根据 [33]，增加分箱数量可以提高推断的准确率，但超过 10 个分箱后，提升幅度会变小。在我们的分析中，使用 5 至 10 个分箱得出的结果大体一致，预测增量的方差改善有限，即 95% 置信区间（CI）略有缩小。

采用倾向评分分层后，VRH 对预测包容性分数的影响为绝对提升 3.4%，95% 置信区间为（2.9%，3.9%）。这意味着，对于参与者超过 2 人的会议，根据包容性机器学习模型，在 95% 的置信水平下，使用 VRH 平均可使预测包容性分数绝对提升 3.4%。

4 数据集

本节详细介绍我们为开发打断检测器而构建准确训练集和测试集所做的工作。我们在第 4.1 节定义语音重叠类别，并在第 4.2 节说明数据集要求。接着，我们在第 4.3 节汇总所用的数据来源。在第 4.4 节，我们概述标注流程，说明此类数据集标注中的细微之处。最后，第 4.5 节介绍训练集和测试集。

4.1 语音重叠类别

语音重叠分类假设发言权由一位说话者掌握。打断者是另一位参与者，通过发言制造重叠语音；他们可能有意获取发言权，也可能无意获取。

- **反馈语**：说话者在背景中表达关注、理解或赞同的一小段对话。其目的不是获取发言权。例如：“是的”“嗯”“嗯哼”。
- **打断失败**：打断者试图在当前说话者发言的同时开口，以获取发言权，但未能获得发言权。
- **打断（或成功打断）**：在第一位说话者尚未说完整个句子之前，打断者便与其语音重叠。打断者成功获得发言权，并吸引了所有其他参与者的注意。
- **笑声**：打断者在第一位说话者发言时发笑。打断者没有获得发言权，也无意获得发言权。
- **其他**：这表示没有明确语音内容的音频重叠情形。这可能是没有可辨认词语的重叠（例如含混语音、清嗓），也可能是背景噪声（例如鼠标点击声）。

4.2 数据集要求

我们研究的 CMC 系统（Microsoft Teams）能够分离每位远程参与者的音频。因此，我们只考虑每位参与者的音频都在独立通道中采集的数据集。这样的设计选择简化了解决方案，因为对于混合声道数据集，需要先分离说话者（即进行说话人分离）。我们的场景代表多方商务会议，因此我们关注的是有三（3）位或更多参与者的远程会议中产生的对话动态。我们将研究范围限定在英语，但采集了不同口音的数据。我们优先考虑以下地区的口音：美国（US）、英国（GB）、印度（IN）和德国（DE）。这些地区是根据产品使用量最高的国家选定的

表2：原始数据来源的人口统计信息

人口统计信息	子类别	数量	百分比
性别	女性	292	53%
	Male	251	46%
	非二元性别	7	1%
口音	US	256	47%
	GB	111	20%
	IN	111	20%
	DE	63	11%
	其他	9	2%
年龄组	18-24	98	18%
	25-34	184	34%
	35-44	107	19%
	45+	161	29%
总计	总计	550	100%

。数据集中男性和女性说话者的比例应相同。由于这些对话代表商务会议，数据集中的说话者年龄须超过 18 岁。在当前版本的数据集中，我们将范围限定为佩戴耳机且网络状况良好的低噪声条件下的说话者。未来，我们计划采集网络（延迟）、设备（例如外放扬声器）和声学（例如噪声）条件各异的数据。

为确保能够验证检测器预测结果的质量，我们要求测试集中标签的准确率达到 95%，该准确率以 Fleiss' Kappa 指标 [34] 衡量。

4.3 数据来源

在开发过程中，我们使用经过筛选的公开数据集和自有原始数据集作为输入数据源。

4.3.1 公开数据来源。 我们使用 AMI 会议语料库 [7] 和 ICSI 会议语料库 [20] 作为公开数据来源。这两个语料库都符合我们的场景需求，但也存在一些局限性。两个语料库的采集方式都是让参与者佩戴耳机并坐在同一个房间内。这会带来两个问题：首先，所有会议都是面对面进行的，这与在线交流可能有所不同；其次，由于人们坐在同一个房间里，当人们的麦克风拾取到其他人的声音时，就会产生串音。我们使用这些数据集来扩充训练数据。

4.3.2 原始数据来源。为缓解公开数据集带来的挑战，并为远程会议场景创建准确的测试集，我们使用自家的会议产品采集了数据。采集这些数据的另一个目标是提升数据集多样性。我们在148场会议中收集了90小时的会议录音。每场会议有3或4位参与者。参与者通过我们的产品远程加入会议，讨论一个或多个存在自然争议的商业话题（例如优先级排序、产品设计等）。在数据收集阶段，我们还确保说话者在性别、口音和年龄方面具有足够的多样性。表2展示了我们原始数据在会议和参与者层面的详细人口统计信息。

4.4 标注流程

创建标注数据主要分为两个步骤：1. 使用准确的语音活动检测（VAD）检测语音重叠，以生成候选片段；2. 按照准确的标注流程标注候选片段。这些片段用于创建训练集和测试集。训练数据集通过众包服务 [1] 进行标注。我们请专家标注了3,000个片段，以创建准确率为 $\geq 95\%$ 的测试集，并通过黄金标准集帮助验证众包标注质量。

4.4.1 候选片段。 对于每场会议，我们使用准确的语音活动检测 [6] 扫描各个通道，以定位语音重叠发生的时间戳。考虑到我们不希望过于频繁地触发该功能，尤其是在讨论不活跃时，我们只保留这样的重叠片段：打断者插话前至少保持沉默 3 秒，且打断者的话语至少持续 0.3 秒。随后，为了进行标注和训练，我们为检测到的每处语音重叠导出一个 10 秒的立体声音频片段。每个片段的第 5 秒是语音重叠的起始点。为了提高标注准确率，片段采用立体声格式，其中打断者的音频位于右声道，其他所有音频则合并到左声道。

4.4.2 众包标注。 我们为众包工作者提供了详细的标注任务说明。专家也使用相同的说明，以验证这些说明是否有效。说明中包括设置指南（例如耳机的使用）、培训模块、资格测试，以及用于验证专家标注质量的黄金标准片段 [32]。说明还列举了示例，帮助厘清容易混淆的情况。其中一个值得注意的挑战是，如何为包含多个类别的片段提供标签（例如，先出现反馈语，随后发生打断失败）。我们沿用 ImageNet 论文中的做法，建立类别层级，并规定每项任务只标注一个类别 [11]。该层级将语音重叠分为两个级别。第 1 级区分成功打断与未发生打断。对于未发生打断这一类别，如果片段中出现多个类别，则按以下优先顺序标注：失败的打断 >反馈语 > 笑声 > 其他。此优先顺序是根据误分类的代价确定的（打断失败较为少见，因此优先级更高）。我们通过调整标注员人数来评估标签质量，最终确定采用标注员的 7 张独立投票。

4.4.3 专家标注。我们从原始数据来源中随机选取了 3,000 个片段，并由内部专家进行标注。具体而言，其中 450 个片段由五位专家标注。我们进行了多轮讨论，以确保充分达成一致并完善说明。总体而言，这些片段中有 92% 达成了五位专家意见完全一致，Fleiss κ系数为 95%。其余 2,550 个片段最多由两位专家标注。我们通过随机抽取 150 个片段并让所有专家对其进行标注，验证这 2,550 个片段的质量，以确保达到准确率门槛。专家标注对于创建准确的测试集以及评估众包标注系统至关重要。我们通过多轮迭代这一流程，确保所有存在语音重叠的片段均得到标注。在这些候选片段中，我们达到了 95% 的准确率。

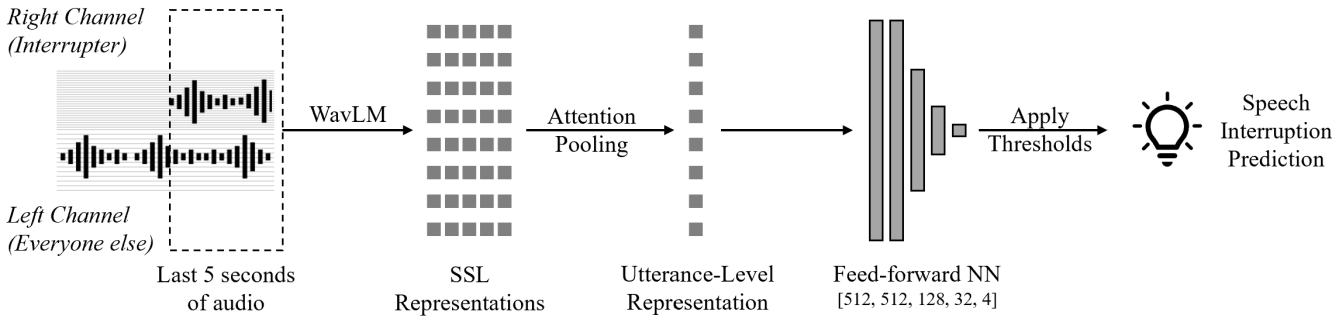


图 2：所　提出的语音打断　　检测模型结构。

4.5 训练集和测试集

众包工作者和专家标注的片段总数如表 1 所示。

测试集：对于测试集，我们使用高质量的专家标注音频片段。我们从每个类别中随机选取 250 个片段（不考虑"其他"类别），构建一个包含 1,000 个片段的固定测试集。

训练集：训练集由众包工作者标注的音频片段组成。我们发现，与多数投票相比，70% 的一致率（即 7 人中有 5 人给出相同标签）能带来更好的模型性能。该阈值在较高的标签质量和训练数据量之间实现了良好的平衡。我们使用专家标签作为真实值，对 1,100+ 个片段进行了比较，从而得出这一结论。因此，我们只使用能够达到 70% 共识水平的片段。

5 模型

整体模型结构如图 2 所示。在模型训练中，我们只使用语音重叠开始后的音频。因此，模型输入为第 4.4.2 节所述双通道音频片段的最后 5 秒。右声道仅包含打断者的语音，左声道则包含所有其他参与者的混合语音。随后，使用特征提取器从原始波形中提取有用特征。鉴于语音的自监督学习表征在多个下游任务中已达到最先进水平，本研究也通过预训练自监督学习模型提取高层嵌入。

将嵌入输入分类器之前，沿时间轴进行池化操作，以降低输入维度。这里我们使用注意力池化（AP）[44]，从一系列帧级表征$\mathbf{U} \in R^{d\times 1}$中获取话语级表征$\mathbf{H}$，其中$\mathbf{H} \in R^{d\times M}$是从SSL模型中提取的嵌入，$d$是功能维度，$M$是帧数。首先使用以下方法计算每一帧的注意力权重$\mathbf{Q} \in R^{1\times M}$：

$$\mathbf{Q} = \text{softmax}(\mathbf{WH}) \tag{1}$$

其中，$\mathbf{W} \in R^{1\times d}$可视为用于决定哪个帧更重要的模板。最后，通过加权求和可以得到话语级表征 $\mathbf{U}$：

$$\mathbf{U} = \mathbf{HQ}^T \tag{2}$$

随后将话语级表征 $\mathbf{U}$ 输入一个 5 层前馈神经网络（也称为深度神经网络），每层的节点数为 [512, 512, 128, 32, 4]，并以 LeakyReLU [27]作为激活函数。采用交叉熵作为损失函数，以随机梯度下降作为优化器，学习率为 0.0015。

6 结果

6.1 不同输入特征和分类器模型的性能比较

本节首先展示不同特征提取方法的结果。除了自监督学习嵌入外，我们还展示了使用MFCC、幅度频谱图和openSMILE声学特征 [13] 的结果，作为基线（对于openSMILE特征，我们未使用注意力池化，因为它本身就是话语级表征）。由于本研究更关注检测失败的打断类别的性能，表3中的曲线下面积和假阳性率为1%时的真正率均以失败的打断类别为准。从表中可以看出，MFCC和幅度频谱图等传统特征的性能远未达到我们的目标。原因可能是，适用于此任务的特征应当**说话人/噪声无关** [18]，并包含**语义信息**。在训练数据有限的情况下，可能难以从传统特征中提取有用信息并舍弃无用信息。比较三个基础模型规模的基于自监督学习的嵌入（即，$Wav2vec2_{Base}$、$HuBERT_{Base}$和$WavLM_{Base+}$），WavLM表现最佳，这与SUPERB基准 [45] 中其他语音任务的结果一致。对于这三种嵌入，我们使用可学习权重对不同Transformer层的表征进行加权求和，并将其作为注意力池化模块的输入。我们发现，与仅使用最后一个Transformer层的嵌入相比，这种方法可以显著提升性能。为进一步提升性能，我们使用 $WavLM_{Large}$进行特征提取，在假阳性率为=1%时，真正率达到 > 50%。

接下来，我们固定 $WavLM_{Large}$作为特征提取器，考察输入声道数量和分类器模型不同时的性能。从图3和表 4的前两行可以看出，如果没有左声道（所有其他参与者的混合语音）提供的信息，性能会大幅下降

表3：使用不同输入特征时，失败的打断 类别的分类器结果（每个数值为10次不同运行结果的平均值）。同时列出 H 的维度以及端到端模型（SSL模型 + 分类器）的参数数量（请注意，对于维度 M，SSL嵌入的帧移为20毫秒；梅尔频率倒谱系数和频谱图采用torchaudio库的默认设置）。

			失败的打断：	
输入特征	H的维度（$d \times M$）	# 参数	AUC	真正率@ 1% 假阳性率
MFCC	(2×40) × 401	0.4M	0.667	3.48%
频谱图	(2×257) × 313	0.4M	0.736	6.24%
openSMILE	(2×88) × 1	0.4M	0.816	11.56%
$Wav2vec2_{Base}$	(2×768) × 249	95M	0.925	32.28%
$HuBERT_{Base}$	(2×768) × 249	95M	0.934	31.52%
$WavLM_{Base+}$	(2×768) × 249	95M	0.943	37.80%
$WavLM_{Large}$	(2×1024) × 249	316M	0.949	50.93%

表 4：使用 $WavLM_{Large}$ 作为特征提取器时，不同输入通道数和分类器模型的比较

		失败的打断：	
输入通道	分类器模型	AUC	真正率@ 1% 假阳性率
双声道	注意力池化+深度神经网络	0.949	50.93%
右声道	注意力池化+深度神经网络	0.918	37.93%
双声道	DNN	0.912	31.05%

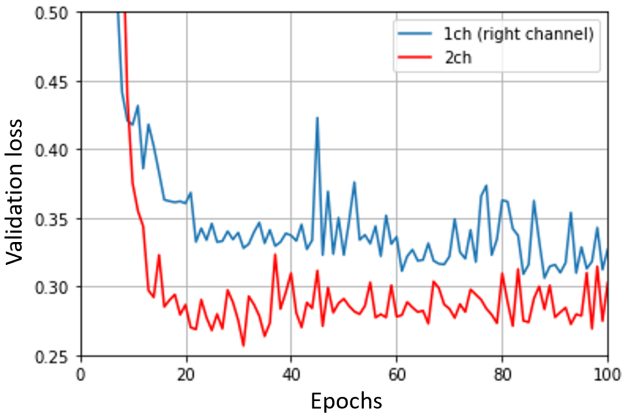


图3：不同输入通道数量下的学习曲线

，这可能是由于模型混淆了失败的打断和成功打断所致（我们将在下一节验证这一点）。接下来，我们想考察注意力池化对模型性能的影响。从下游任务中移除注意力池化，意味着将展平后的帧级表征 H 直接输入深度神经网络分类器。这不仅会显著增加分类器的模型规模，而且从表 4的第三行可以看出，真正率也降低了 20%。原因可能是相较于训练数据的数量，输入维度过大。

6.2 混淆矩阵分析

为进一步分析 $WavLM_{Large}$的预测结果，我们分别在表5和表6中展示了双通道输入和单通道输入的混淆矩阵。我们增加了一个名为“低于1%假阳性率阈值”的列，用于表示以下情况：根据 $argmax$（.）函数，失败的打断本应是预测类别，但其置信度分数低于实现1%假阳性率所需的阈值。从表5可以看出，模型最容易混淆的是反馈语和失败的打断。我们认为，这是因为两者的语音活动模式非常相似（即打断者短暂发言，而其他参与者继续说话）。换言之，模型不能仅凭音频能量分布区分二者，还必须通过自监督学习嵌入中的语义信息推断说话意图。比较表5和表6可知，当真实值为失败的打断时，单通道输入的模型更容易将其误分类为成功打断。这验证了我们在上一节中的假设：左声道（所有其他参与者的混合语音）有助于模型区分成功打断和失败的打断。

6.3 与其他语音任务的关系

语音打断分类是一个相对较新的主题。因此，我们想了解它与传统语音任务之间的关系。如第 6.1 节所述，基于自监督学习的嵌入由不同 Transformer 层的加权求和构成，并使用可学习权重。因此，学习得到的权重可以提供一些信息，帮助我们了解哪些层对特定任务更为重要。在图 4 中，我们以 $WavLM_{Base+}$ 的学习层权重为例，将我们的权重与 SUPERB基准中其他语音任务学习得到的权重进行比较。从图中可以看出，我们学习得到的权重模式与关键词检测（KS）任务学习得到的模式最为相似。这表明这两个任务使用的输入特征彼此相似，因此它们之间的关联最为密切。我们推测，这是因为模型在区分反馈语和失败的打断时，依赖于某种类似于 KS 的机制（例如，在反馈语的情况下，打断者通常会说“yeah”“ Mm-hmm”等）。

表5： $WavLM_{Large}$**双通道输入的预测结果混淆矩阵。**

	预测值				
真实值	反馈语	失败打断	打断	笑声	低于1% 假阳性率阈值
反馈语	219	6	8	11	6
失败的打断	47	131	19	7	46
打断	9	1	225	7	8
笑声	37	0	1	212	0

表6： $WavLM_{Large}$**一个通道输入的预测结果混淆矩阵** **t.**

	预测值				
真实值	反馈语	失败打断	打断	笑声	低于1% 假阳性率阈值
反馈语	201	4	3	20	22
失败的打断	42	85	35	12	76
打断	8	4	207	11	20
笑声	29	0	0	221	0

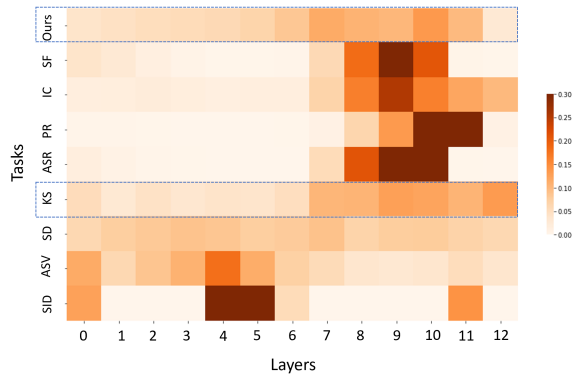


图 4：针对不同任务，从$WavLM_{Base+}$ **Transformer模型不同层学习到的权重。**

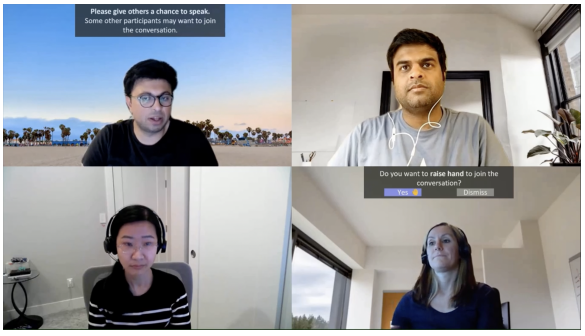


图 5：失败语音打断检测器的应用示例

7 应用

作为失败的语音打断检测的一种应用示例，我们可以提醒失败的打断者使用虚拟举手功能，如图 5 所示。此外，我们还可以提醒没有让出发言机会的说话者：当每个人都有机会发言时，会议会更具包容性；尤其是当他们在该会议中反复不让出发言机会时。最后，可以将失败的语音打断检测结果记录在 CMC 系统的通话遥测数据中并加以分析，以进一步加深对会议成效和包容性的理解；开展 A/B 测试，评估 CMC 系统中用于减少失败语音打断的新功能；并改进会议成效和包容性的预测模型。

8 结论

本文介绍了一种通过语音重叠分析提升会议包容性的方法。根据我们的发现，我们提出了远程会议中"打断失败"这一挑战。我们创建了首个准确的标注数据集，以应对这一挑战，适用于计算机媒介通信系统。借助语音自监督学习表征领域的最新进展，我们构建了一个检测器，其真正率超过 50%，假阳性率低于 1%。该数据集需要扩展，以支持多种语言和不同网络及设备条件等场景。我们计划利用该数据集发起一项挑战赛，并发布基线模型，以推动这一主题的积极研究。我们还在将该模型集成到 Microsoft Teams 中，并开展 A/B 测试，以衡量包容性和整体会议成效的提升。最后，我们正在开发一种全双工机器学习声学回声消除器（例如，参见 [41]），它也有助于远程参会者在会议中打断他人并更好地参与会议。

回应

我们感谢 Teams 团队的 Juhee Cho、Alex Chzhen 和 Chinmaya Madan 对模型和数据基础设施所作的贡献。Azure 认知团队为模型开发工作提供了有力指导。特别感谢 Yu Wu、Zhou Chen 和 Dimitrios Dimitriadis 进行了富有成效的讨论。

参考文献

[1] Scale AI。2022年。AI 数据平台：为 AI 应用提供高质量训练和验证数据。https://scale.com/about [2] Peter C. Austin。2011年。倾向得分方法简介：减少观察性研究中的混杂影响。多变量行为研究 46, 3 2011 399 424 10 1080 00273171 2011 568786 (5月)，–。[3] https://doi.org/。/。。Alexei Baevski、Yuhao Zhou、Abdelrahman 2020 Mohamed 和 Michael Auli。。wav2vec 2.0：一种用于语音表征自监督学习的框架。神经信息处理系统进展 33 (2020)，12449–12460。[4] Don Baron。2002年。基于韵律的 ICSI 会议录音语料库中标点符号和打断事件自动检测。研究项目，加州大学伯克利分校，美国 (2002)。[5] Kofi Boakye、Beatriz Trueba-Hornero、Oriol Vinyals 和 Gerald Friedland。2008年。通过重叠语音检测改进多方会议中的说话人分离。载于《2008 IEEE 声学、语音与信号处理国际会议》。IEEE，4353–4356。[6] Sebastian Braun 和 Ivan Tashev。2021年。关于抗噪语音活动检测的训练目标。载于《2021年第29届欧洲信号处理会议 (EUSIPCO)》。421–425。https://doi.org/10.23919/EUSIPCO54536.2021.9616082 ISSN：2076-1465。[7] Jean Carletta、Simone Ashby、Sebastien Bourban、Mike Flynn、Mael Guillemot、Thomas Hain、Jaroslav Kadlec、Vasilis Karaiskos、Wessel Kraaij、Melissa Kronen- thal、Guillaume Lathoud、Mike Lincoln、Agnes Lisowska、Iain McCowan、Wilfried Post、Dennis Reidsma 和 Pierre Wellner。2006年。AMI 会议语料库：预告。载于《多模态交互机器学习》（《计算机科学讲义》），Steve Renals 和 Samy Bengio（编）。施普林格，柏林，海德堡，28–39。https://doi.org/10.1007/11677482_3[8] Sanyuan Chen、Chengyi Wang、Zhengyang Chen、Yu Wu、Shujie Liu、Zhuo Chen、Jinyu Li、Naoyuki Kanda、Takuya Yoshioka 和 Xiong Xiao。2021年。WavLM：面向全栈语音处理的大规模自监督预训练。arXiv 预印本 arXiv:2110.13900 (2021)。[9] W. G. Cochran。1968年。通过亚分类调整消除观察性研究中偏倚的有效性。《生物统计学》24卷，2期 (1968年6月)，295。https://doi.org/10.2307/2528036[10] Ross Cutler、Yasaman Hosseinkashi、Jamie Pool、Senja Filipi、Robert Aichner、Yuan Tu 和 Johannes Gehrke。2021年。远程协作中的会议成效与包容性。《ACM 人机交互汇刊》5卷，CSCW1 (2021)，1–29。[11] Jia Deng、Wei Dong、Richard Socher、Li-Jia Li、Kai Li 和 Li Fei-Fei。2009年。ImageNet：大型分层图像数据库。载于《2009 IEEE 计算机视觉与模式识别会议》。IEEE，迈阿密，佛罗里达州，248–255。https://doi.org/ 10 1109 2009 5206848 。/CVPR.。。Paul Van Eecke 和 [12] 2016 Raquel Fernández。。性别对多方对话中打断行为的影响。2070–2074。https://doi.org/10.21437/Interspeech. 2016-951[13] Florian Eyben、Martin Wöllmer 和 Björn Schuller。2010年。openSMILE：慕尼黑通用快速开源音频特征提取器。载于《第18届 ACM 多媒体国际会议论文集》（MM '10）。美国计算机协会，美国纽约州纽约市，1459–1462。https://doi.org/10.1145/1873951. 1874246[14] Erin Fitzgerald、Keith B Hall 和 Frederick Jelinek。2009年。重建自发语音文本中的错误起始。(2009)。[15] Florian Hammer、Peter Reichl 和 Alexander Raake。2005年。《调谐良好的对话：交互性、延迟与感知 VoIP 质量》。载于 IEEE 国际通信会议论文集 (ICC)，韩国首尔。[16] Wei-Ning Hsu、Benjamin Bolte、Yao-Hung Hubert Tsai、Kushal Lakhotia、Rus- lan Salakhutdinov 和 Abdelrahman Mohamed。2021年。HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习。《IEEE/ACM音频、语音与语言处理汇刊》29 (2021)，3451–3460。出版方：IEEE。[17] K. H. Hullsiek。2002年。用于观察性数据因果分析的倾向得分建模策略。《生物统计学》3卷，2期 (2002年6月)，179–193。https://doi.org/10. 1093/biostatistics/3.2.179[18] Kuo-Hsuan Hung、Szu-Wei Fu、Huan-Hsin Tseng、Hsin-Tien Chiang、Yu Tsao 和 Chii-Wann Lin。2022年。增强用于语音增强的自监督嵌入。arXiv:2204.*03339 [*eess*]* (2022年4月)。http://arxiv.org/abs/2204. 03339 arXiv：2204.03339。[19] Deborah James 和 Janice Drakich。1993年。理解谈话量中的性别差异：研究批判性综述。载于《性别与会话互动》。牛津大学出版社，美国纽约州纽约市，281–312。[20] A. Janin、D. Baron、J. Edwards、D. Ellis、D. Gelbart、N. Morgan、B. Peskin、T. Pfau、E. Shriberg、A. Stolcke 和 C. Wooters。2003年。ICSI 会议语料库。载于《2003 IEEE 声学、语音与信号处理国际会议论文集》(ICASSP '03)，第1卷。I–I。https://doi.org/10.1109/ICASSP.2003. 1198793 ISSN：1520-6149。[21] Guolin Ke、Qi Meng、Thomas Finley、Taifeng Wang、Wei Chen、Weidong Ma、Qiwei Ye 和 Tie-Yan Liu。2017年。LightGBM：高效的梯度提升决策树。神经信息处理系统进展 30 (2017)。[22] Oscar Kempthorne。1952年。实验设计与分析。（1952年）。出版商：Wiley。[23] Lyndon S Kennedy 和 Daniel PW Ellis。2004年。会议中的笑声检测。（2004年）。[24] Juntae Kim 和 Minsoo Hahn。2018年。使用自适应上下文注意力模型进行语音活动检测。《IEEE 信号处理快报》25，8（2018），1181–1185。出版商：IEEE。[25] Campbell Leaper 和 Melanie M. Ayres。2007年。成人语言使用中性别差异的元分析综述：健谈程度、亲和性言语与坚定性言语。《人格与社会心理学评论》11，4（2007年11月），328 363 10 1177 1088868307302221 –。https://doi.org/ 。/[26] Yang Liu、Elizabeth Shriberg 和 Andreas Stolcke。2003年。利用多个知识源自动识别会话语音中的非流利现象。载于《第八届欧洲语音通信与技术会议》。Citeseer。[27] Andrew L. Maas、Awni Y. Hannun 和 Andrew Y. Ng。2013年。整流器非线性改善神经网络声学模型。载于 ICML 音频、语音与语言处理深度学习研讨会。[28] Dave Makhervaks、William Hinthorn、Dimitrios Dimitriadis 和 Andreas Stolcke。2020年。结合声学、内容和交互特征，发现会议中的热点。载于《ICASSP 2020-2020 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，8054–8058。[29] Eleni Margariti、Sean Rintel、Brendan Murphy 和 Abigail Sellen。2022年。自动绘制视频会议中的竞争性与协作性重叠对话。（2022年），8。[30] Louis-Philippe Morency、Iwan de Kok 和 Jonathan Gratch。2008年。预测听者反馈语：一种概率多模态方法。载于《智能虚拟代理国际研讨会》。施普林格，176–190。[31] 不适用。不适用。匿名评审隐藏。（不适用）。[32] Babak Naderi 和 Ross Cutler。2020年。ITU-T 建议书 P.808 的开源实现与验证。《Interspeech》（2020年10月），2862–2866。https://doi.org/ 10 21437 2020 2665 。/Interspeech。- Markus [33] Neuhäuser、Matthias Thielmann 和 Graeme D Ruxton。2018年。二元结局倾向得分分层中的层数。《医学科学档案》：AMS 14，3（2018），695。出版商：Termedia Publishing。[34] Thomas R Nichols、Paola M Wisner、Gary Cripe 和 Lakshmi Gulabchand。2010年。κ统计量的应用。《质量保证期刊》13，3-4（2010），57–61。出版商：Wiley Online Library。[35] Paul R. Rosenbaum 和 Donald B. Rubin。1984年。使用倾向得分亚分类减少观察性研究中的偏倚。《美国统计学会杂志》。79, 387 1984 516 524 10 1080 01621459 1984 10478078 （9[36] 2017 月），–。https://doi.org/ 。/。。Robin Ruede、Markus Müller、Sebastian Stüker 和 Alex Waibel。。利用词嵌入增强反馈语预测。。载于 Interspeech。879–883。[37] Harvey Sacks、Emanuel A Schegloff 和 Gail Jefferson。1978年。会话话轮交替组织的最简系统。载于《会话互动组织研究》。Elsevier，7–55。[38] Tara Sainath 和 Carolina Parada。2015年。用于轻量级关键词检测的卷积神经网络。（2015年）。[39] Emanuel A Schegloff。2000年。重叠对话与会话话轮交替的组织。《语言与社会》29，1（2000），1–63。出版商：剑桥大学出版社。[40] Abigail Sellen。1995年。远程对话：媒介化对话与技术的影响。《人机交互》10，4（1995年12月），401–444。https: //doi.org/10.1207/s15327051hci1004_2 [41] Kusha Sridhar、Ross Cutler、Ando Saabas、Tanel Parnamaa、Markus Loide、Hannes Gamper、Sebastian Braun、Robert Aichner 和 Sriram Srinivasan。2021年。ICASSP 2021 声学回声消除挑战赛：数据集、测试框架与结果。载于《ICASSP 2021-2021 IEEE 声学、语音与信号处理国际会议（ICASSP）》。IEEE，151–155。[42] Ming Sun、Anirudh Raju、George Tucker、Sankaran Panchapagesan、Gengshen Fu、Arindam Mandal、Spyros Matsoukas、Nikko Strom 和 Shiv Vitaladevuni。2016年。用于轻量级关键词检测的长短期记忆网络最大池化损失训练。载于《2016 年 IEEE 口语技术研讨会（SLT）》。IEEE，474–480。[43] Zheng-Hua Tan、Najim Dehak 等。2020年。rVAD：一种基于无监督分段的鲁棒语音活动检测方法。《计算机语音与语言》59（2020），1–21。出版商：Elsevier。[44] Wei-Cheng Tseng、Chien-yu Huang、Wei-Tsung Kao、Yist Y Lin 和 Hung-yi Lee。2021年。利用自监督表示进行平均意见分预测。arXiv 预印本 arXiv:2104.03017（2021）。[45] Shu-wen Yang、Po-Han Chi、Yung-Sung Chuang、Cheng-I. Jeff Lai、Kushal Lakho-tia、Yist Y. Lin、Andy T. Liu、Jiatong Shi、Xuankai Chang、Guan-Ting Lin、Tzu-Hsien Huang、Wei-Cheng Tseng、Ko-tik Lee、Da-Rong Liu、Zili Huang、Shuyan Dong、Shang-Wen Li、Shinji Watanabe、Abdelrahman Mohamed 和 Hung-yi Lee。2021年。SUPERB：语音处理通用性能基准测试。arXiv:2105.*01051 [*cs，eess*]*（2021年10月）。http://arxiv.org/abs/2105.01051 arXiv：2105.01051。