arXiv:2305.05443v1 [eess.AS] 9 May 2023

探索无监督跨任务语音表示在“真实环境”边缘应用中的性能

Heitor Guimar˜ães[1], A rthur Pimentel[1], Anderson Avila[2], Mehdi Rezagholizadeh[2], Tiago H. Falk [1]

[1] INRS-EMT，魁北克大学′，蒙特利尔′，加拿大[2] 华为
诺亚方舟实验室*，蒙特利尔，加拿大

摘要

无监督语音模型正在语音和机器学习领域中变得无处不在。上游模型负责从原始音频中学习有意义的表示。随后，这些表示作为输入交由下游模型，用于解决多种任务，例如关键词检测或情感识别。随着边缘语音应用开始涌现，评估这些跨任务表示在资源有限、噪声水平各异的边缘设备上的鲁棒性至关重要。为此，本研究评估了 HuBERT 的四种不同版本的鲁棒性，即基础版、大型版和超大型版，以及近期推出的 Robust-HuBERT。在不同的加性噪声和卷积噪声条件下，针对三项下游任务进行了测试：关键词检测、意图分类和情感识别。结果表明，较大的模型虽然能够在一定程度上增强对环境因素的鲁棒性，但可能不适用于边缘应用。另一方面，较小的模型在噪声条件下准确率显著下降，尤其是在存在房间混响时。这些发现表明，跨任务语音表示尚未做好用于边缘应用的准备，仍需进一步创新。

引言

从高维数据中提取有意义的特征，是机器学习面临的主要挑战之一。对于语音应用，如何提取可用于多种任务的有意义特征，一直是近期研究的重点（例如，Chen、Wang 等人（2022年））。事实上，这些学习得到的表示对于边缘语音应用的发展至关重要，因为资源限制会使得在边缘设备上存储并运行不同模型变得困难，甚至不可能。

如今，HuBERT，即基于Transformer的隐藏单元双向编码器表示，是领先的自监督语音表征之一（Hsu 等人，2021年）。它已被广泛用于语音处理通用性能基准（SUPERB），该基准基于成熟的公共数据集构建，涵盖十项语音处理任务（Yang、Chi 等人，2021年）。HuBERT 已向社区提供多种规模的版本，从基础模型的9500万个参数到超大型版的近 10 亿个参数不等。近期研究表明，HuBERT 的性能会在噪声环境下下降；Huang 等人（2022年）近期提出了一个更稳健的版本，本文称之为 Robust-HuBERT。

本研究旨在进一步了解加性噪声和卷积噪声（房间混响）可能对基于 HuBERT 表示的跨任务性能产生的影响。此外，鉴于我们的最终目标是在边缘设备上应用这些任务，因此有必要评估模型大小对整体准确率的影响。本文仅关注三项任务：（i）内容任务（关键词检测）；（ii）语义任务（意图分类）；以及（iii）副语言任务（情绪识别）。

方法与材料

自监督语音表征

如前所述，HuBERT 是一种自监督学习方法，能够基于仅包含语音的数据获得表示。该模型依赖于学习所谓的隐藏声学单元（即音素单元），这些单元可用作帧级目标（Hsu 等人 2021年）。因此，给定一个话语 $X = [x_1, x_2, ..., x_N]$，其中包含 N声学帧，以及相应的隐藏单元目标序列 $Z = [z_1, z_2, ..., z_N]$，聚类模型必须学会将 z_t分配给一组类别 $C = [c_1, c_2, ..., c_K]$，其中 C是一个 K-分类变量。因此，采用离线聚类方法，为类似 BERT 的预测损失提供对齐的目标标签。该损失应用于掩码区域，迫使模型在连续输入上学习结合声学和语言模型，即：

$$L_m(f; X, M, Z) = \sum_{t \in M} log p_f(z_t | \tilde{X}, t), \qquad (1)$$

其中，$M \in [N]$ 是长度为 N 的序列 X 中需要进行掩码处理的索引集合，f 是一个掩码预测模型。因此，X 的扰动版本表示为$\tilde{X} = r(X, M)$，其中 r是一个掩码函数。

请注意，HuBERT 模型主要依赖无监督聚类步骤，而非..... 的质量。

*本稿中表达的任何观点、研究结果和结论均为作者所有，不一定反映华为的观点、官方政策或立场。版权所有 © 2022年，人工智能促进协会（www.aaai.org）。保留所有权利。

上游	#参数 (M)	关键词检测 (c)	(n)	(n + r)	意图分类 (c)	(n)	(n + r)	情感识别 (c)	(n)	(n + r)
HuBERT 基础版	95	0.9630	0.8377	0.5294	0.9831	0.7962	0.5663	0.6487	0.5272	0.3173
HuBERT 大型版	317	0.9529	0.7478	0.4450	0.9873	0.8408	**0.7171**	**0.6745**	0.5696	0.4114
HuBERT 超大型版	964	**0.9705**	0.8747	**0.7485**	**0.9921**	0.8165	0.6599	0.6650	0.5731	**0.4880**
鲁棒 HuBERT	95	0.9662	**0.9140**	0.6702	0.9808	**0.9262**	0.7142	0.6470	**0.5872**	0.3494

表 1：关键词检测、意图分类和情感识别在干净 (c)、有噪声 (n) 以及噪声加混响 (n+r) 测试条件下的实验结果。

聚类分配标签。以包含 100 个簇的简单 k 均值教师模型为起点，HuBERT 模型的表现与 wav2vec 2.0 模型相当或更优。有关该模型的更多详情，请参阅 Hsu 等人（2021年）。

本文探讨了 HuBERT 的三种变体。基础版的 Transformer 网络有 12 层。大型版的层数翻倍，超大型版则是四倍。HuBERT 基础版使用 960 小时的 Librispeech 数据集进行训练（Panayotov 等人，2015），而大型版和超大型版则使用 Librilight 数据集中的 6 万小时数据进行训练（Kahn、Rivi`ere 等人，2020）。Robust-HuBERT 则是一项近期创新，它采用域对抗训练，其中判别器是一个域网络，负责对施加于话语的失真来源进行分类，以提高噪声条件下的准确率（Huang 等人，2022年）。

下游任务与评价指标

我们的实验采用了三项下游任务。关键词检出是一项 12 分类任务，需要识别 12 条语音命令。由于这对边缘应用至关重要，因此在真实环境中的鲁棒性和模型大小是关键因素。其次，意图分类旨在将样本归入三个可能的类别之一：动作、物体和位置。最后，顾名思义，情感识别是一项多分类任务，需检测说话人的情绪属于中性、开心、悲伤或愤怒。所有任务均以识别准确率作为评估指标，因此数值越高越好。有关这些任务及其所依赖的公共数据集的更多详情，请参阅 Yang、Chi 等人（2021年）。

实验设置

为模拟边缘应用中可能遇到的条件，我们在 SUPERB 测试集上添加了加性噪声和卷积噪声（即房间混响）。为此，我们使用了 Deep Noise Suppression Challenge 4 噪声数据集（Dubey、Gopal 等人，2022年）。该数据集包含 180 小时的噪声，共有 62,000 条语句，涵盖 150 种不同的噪声类型。为进一步模拟房间混响，我们使用了 openSLR28 数据集中的 248 个真实房间脉冲响应，以及来自 openSLR26 的 60,000 个合成房间脉冲响应（Ko 等人，2017年），覆盖小型、中型和大型房间。所有文件均重采样至 16 kHz。

在我们的实验中，将采用三种条件：(1)“干净 (c)”，即使用 SUPERB 中的测试条件；(2)“含噪 (n)”，即向 SUPERB 测试文件中加入噪声，信噪比 (SNR) 从区间 $[0, 20]$ dB 内均匀采样；(3)“噪声加混响 (n+r)”，即将房间脉冲响应与 SUPERB 测试文件进行卷积，然后按照条件“(n)”加入加性噪声。为确保三项任务中的退化条件一致，采样过程中采用自定义种子。

结果

表 1 展示了我们针对全部三项任务和三种条件进行分析所得的结果。如图所示，对于关键词检测，随着探索更大版本的 HuBERT，准确率有所提升。在干净条件下，Robust-HuBERT 取得了与 HuBERT 超大型模型相当的结果，同时所需参数量减少了一个数量级。事实上，在噪声条件下，Robust-HuBERT 的表现优于 HuBERT 超大型模型，从而进一步印证了对抗训练的优势。总体而言，噪声和混响对关键词检测的影响最为不利，即使是 Robust-HuBERT，准确率也大幅下降。

对于意图分类，在干净条件下可以观察到类似的表现：随着模型规模增大，准确率也随之提升。在噪声条件和噪声加混响条件下，Robust-HuBERT 取得了具有竞争力的性能，甚至优于超大型模型。然而，最终准确率仍显著低于干净条件下的结果，这表明边缘应用仍有很大的改进空间。最后，对于情感识别，Robust-HuBERT 在噪声条件下取得了最高准确率，但仅略高于超大型 HuBERT 模型。噪声加混响条件再次表现出最大的挑战性，准确率下降到几乎只有 Robust-HuBERT 在干净条件下所能达到水平的一半。总体而言，尽管超大型模型对噪声条件表现出一定的韧性，Robust-HuBERT 在这些条件下仍取得了最高准确率。然而，当噪声和混响同时存在时，所有模型的性能都大幅下降。

结论

在这项工作中，我们评估了四种 HuBERT 跨任务语音表征变体在三项 SU- PERB 任务中的表现，涵盖噪声以及噪声加混响条件。虽然较大的模型往往能在干净条件下取得更高的准确率，但在嘈杂环境中的表现会下降，这表明仍需进一步创新，尤其是对于边缘语音应用而言，因为这类应用对模型大小的要求是

一个重要的限制性参数。

参考文献

Chen, S.; Wang, C.; 等. 2022年. WavLM：面向全栈语音处理的大规模自监督预训练. IEEE J. Selected Topics Signal Proc. Dubey, H.; Gopal, V.; 等. 2022年. ICASSP 2022 深度噪声抑制挑战赛. 见 ICASSP 论文集. Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; 和 Mohamed, A. 2021年. HuBERT：通过隐藏单元的掩码预测进行自监督语音表征学习. IEEE/ACM 音频、语音与语言处理汇刊, 29: 3451–3460. Huang, K. P.; Fu, Y.-K.; Zhang, Y.; 和 Lee, H.-y. 2022年. 通过领域自适应提升自监督语音处理任务的失真鲁棒性 . arXiv:2203.16104. Kahn, J.; Rivi`ere, M.; 等. 2020年. Libri-light：一个监督有限或无监督的自动语音识别基准. 见 IEEE ICASSP 论文集 , 7669–7673. Ko, T.; Peddinti, V.; Povey, D.; Seltzer, M. L.; 和 Khudanpur, S. 2017年. 面向鲁棒语音识别的混响语音数据增强研究. 见 IEEE ICASSP 论文集, 5220–5224. Panayotov, V.; Chen, G.; Povey, D.; 和 Khudanpur, S. 2015年. Librispeech：基于公共领域有声读物的自动语音识别语料库. 见 2015年 IEEE 声学、语音与信号处理国际会议（ICASSP）, 5206–5210. IEEE. Yang, S.-w.; Chi, P.-H.; 等. 2021年. SUPERB：语音处理通用性能基准. arXiv 预印本 arXiv:2105.01051.