# 自监督表征学习在说话人日志与语音分离中的应用

S´everin Baroudi[1], Herv´e Bredin[2],Joseph Razik[1],Ricard Marxer[1]

[1]土伦大学、艾克斯-马赛大学、法国国家科学研究中心、LIS，土伦，法国[2]IRIT / pyannoteAI，图卢兹，法国

{severin.baroudi,joseph.razik,ricard.marxer}@lis-lab.fr, herve@pyannote.ai

arXiv:2512.15224v1 [eess.AS] 17 Dec 2025

摘要——**近年来，诸如 wav2vec 2.0 和 WavLM 等自监督语音模型已被证明能够显著提升许多下游语音任务的性能，尤其是在资源匮乏的情况下。尽管如此，目前针对说话人日志和语音分离等任务的评估仍然十分有限。本文研究了近期自监督语音表征在这两项与说话人身份相关的任务上的质量，重点指出当前文献中的不足之处，这些不足源于现有基准测试的局限性——尤其是评估数据集缺乏多样性，以及与说话人日志和语音分离相关的下游系统缺乏多样性。**

索引词—**说话人日志、说话人分离、自监督表征学习。**

## 一、引言

近年来，语音表征的自监督学习（SSL）范式展现出巨大潜力。通过提升自动语音识别（ASR）或说话人情感识别 [1]，等多种下游任务的表现，这些方法在转录或语音分类任务中均展现出强大的泛化能力。尽管当前许多研究探讨了在自动语音识别领域中使用自监督学习表征，但说话人日志（SD）和源分离（SS）任务受到的关注相对较少。说话人日志的任务是根据活跃说话人的身份，将音频流划分为多个片段。这项任务旨在回答音频文件中“谁在何时说话？”这一问题。与之相反，语音分离的任务是从混合语音中估计每位说话人的目标语音，同时去除背景噪声和其他干扰。这项任务的难度很大程度上取决于需要分离的说话人数，以及录音的声学条件。

近期与 SD 或 SS 相关的研究，例如 [2]–[4]，往往只采用一种主要的语音自监督学习基础模型（WavLM [5]），旨在提升其中某项任务的性能。尽管这种方法有助于获得性能最佳的系统，却忽视了在更广泛的自监督学习模型中对这些内部表征进行深入评估，例如 HuBERT[6] 或 wav2vec 2.0.0 [7]。这类研究的另一个重要缺陷，是没有评估语音基础模型的关键因素，例如预训练数据集的类型以及上游前置任务的性质（预测式或对比式）。最后，尽管已有研究在多种语音基础模型上，使用语音处理通用性能基准 [1]（SUPERB）对 SD 和 SS 开展了大规模评估，例如 [8] 和 [5]，但仍存在一些局限。其中一个局限涉及 SUPERB 用于评估这两项任务的数据集（LibriMix [9]），该数据集源自与某些自监督学习模型（如 wav2vec 2.0、HuBERT 和 WavLM）预训练所用的相同语料库（LibriSpeech-960h [10]），因而可能导致领域不匹配和结果偏差。此外，SUPERB 所用模型与较新方法在 SD 和 SS 下游架构方面的差异，可能会影响自监督学习模型内部表征的激活程度，从而引发对这些基准测试跨模型逐层分析有效性的担忧。

为解决这些局限性，我们作出了以下几项重要贡献：

- 我们对一系列常用的自监督学习架构的原始性能进行了基准测试。这些架构在预训练数据集（LibriSpeech 或自定义对话数据集 [3]）、大小和前置任务方面各不相同。我们的目标不是试图击败当前最先进水平来优化每个系统的性能，而是仅基于给定语音基础模型提供的原始表征，明确比较哪种架构最适合每项任务。为此，我们的研究重点是冻结的自监督学习模型，而非微调版本。

- 我们分别在说话人日志（DIHARD3 [11]）和语音分离（WSJ0 [12]）领域广泛使用的基准数据集上进行评估，以此缓解我们此前引入的领域不匹配问题；此前的基准测试仅使用了 LibriMix。
- 为了提供近期说话人分离和语音分离模型的基准测试，我们采用*Pyannote* [13] 的神经端到端分割方法进行说话人日志分析，并采用仅在时域运行的架构 [14] 进行分离，而非使用 SUPERB 提供的下游模型。对于后者，我们评估了三种不同的掩码架构（ConvTasNet [15], DPRNN [16] 和 SepFormer [17]），以评估结果的一致性。
- 我们针对每项任务，对每个自监督学习模型进行完整的逐层分析，从而能够比较预训练数据集和前置任务对下游评估的影响。此外，通过利用 MetaAI [1] 最新的语音基础模型 w2v-BERT 2.0 [18]——该模型融合了对比式和预测式学习方法——我们能够直接评估这两种训练范式对所评估的自监督学习模型的影响。

## II. 语音的自监督学习

近期的语音自监督方法受到自然语言处理模型（如 BERT [19]）的启发，并采用共同的主干网络来学习和生成语音表征。首先，它们使用具有不同卷积核和步幅的卷积神经网络（CNN），将语音音频编码为潜在表示。每个编码帧实质上表示 16 kHz 音频中的 20 毫秒，因此帧率为 50 Hz。随后，按照给定的掩码比例对 CNN 输出的随机时间片段进行掩码处理。掩码后的特征被送入一系列 Transformer 层（对于 w2v-BERT，则为 Conformer [20] 层），通过基于自注意力的机制对被掩码部分进行上下文预测。预测得到的上下文向量会与目标伪标签进行比较。

目标生成方式因架构而异：HuBERT 使用通过对训练集进行 k-means 聚类预测出的离散化隐藏单元 [6]，，而 wav2vec 2.0 则利用量化码本单元 [21]。前者采用掩码语言建模（MLM）损失，后者则使用对比损失。HuBERT 仅使用源自有声读物数据集 LibriSpeech 的单说话人语音内容进行训练，而自监督学习模型 WavLM [5]则尝试通过对音频输入应用话语混合策略，使基于隐藏单元的自监督学习专门适用于多说话人内容。另一方面，对话式 WavLM[3] 依赖不同的预训练数据集，并利用一个组装的多领域语料库，其中包含 663 小时的真实场景对话数据集，例如 DIHARD-3 [11],AiSHELL-4 [22],或 AMI [23]。

最后，w2v-BERT [18] 同时受益于对比式学习和预测式学习，从而提供有意义的语音表示。该架构包含 24 个 Conformer 层，其中前一半支持模型的对比式部分，后一半增强其预测式组件。与 wav2vec 2.0 类似，对比式部分的目标上下文向量通过量化模块生成，该模块使用由码本导出的码字。该量化器还用于生成待预测的相关隐藏单元，这些隐藏单元服务于模型的预测式部分。模型最后的 12 个 Conformer 层采用与 HuBERT 相同的学习策略，利用 MLM 损失，并将最后一层的输出作为上下文向量、将量化器单元作为目标。模型通过联合优化对比式损失和 MLM 损失来结合这两种学习方式，并使用来自 143 种以上不同语言的 450 万小时音频数据进行预训练。

## III. 评估说话人日志

我们使用最新的开源*pyannote 3.1* 工具包 [24]，作为说话人日志任务的评估基线，该工具包在多个说话人日志数据集上展现了最先进的性能。它采用局部监督式说话人分割模型，并结合全局无监督聚类 [25]。分割模型包含一个逐帧特征提取器（SincNet [26]），用于处理时长为 10 秒的音频片段。音频片段随后作为多个双向长短期记忆网络和线性层的输入，以在幂集 空间中执行多类分类 [27]。幂集包含一个非语音类别、每位单独说话人各自对应的类别，以及每对独特重叠说话人对应的类别。训练完成后，说话人分割模型会输出 10 秒音频片段中活跃说话人的类别。要处理完整的音频文件，必须拼接各个局部窗口。为此，系统会从每个窗口中仅有一位活跃说话人的片段提取说话人嵌入，并执行凝聚层次聚类（AHC），以确定每个片段中各说话人的全局身份。说话人日志错误率（DER）是用于评估各系统的指标。该指标通过对一个音频片段中的三种特定错误率求和得出：说话人混淆率、虚警率和漏检率。

为了评估每个自监督模型，我们用一个冻结的 SSL 模型替换了 SincNet 特征提取器。随后我们

[1]https://pytorch.org/audio/main/pipelines[2] https://huggingface.co/facebook/w2v-bert-2.0

表 I 不同自监督语音表征模型在各种基于 TasNet 的语音分离架构（使用 WSJ0-2 混合测试集）以及说话人分离流水线（使用 DIHARD 3 测试集）上的总结与评估。对于分离任务，报告的指标为 SDRi（dB）。对于说话人日志，FA、MD 和 SC 分别指虚警率、漏检率和说话人混淆率。去噪指 WavLM 论文 [5]中介绍的语音去噪前置任务。

| 模型概述 | | | | | | | 评估结果 | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 学习任务 | | | | | 分离 | | | 说话人日志 | | | |
| 自监督模型 | 预训练数据集 | Predictive | Contrastive | Denoising | 层 Size | 模型 Size | ConvTasNet（SDRi ↑） | DPRNN（SDRi ↑） | SepFormer（SDRi ↑） | FA (%) | MD (%) | SC (%) | 说话人日志错误率 ↓ (%) |
| wav2vec2.0 [7] [1] | LibriSpeech（960 小时） | | ✓ | | 12 Transformer | **95M** | 17.0 | 19.2 | 20.2 | 4.8 | 7.7 | 7.9 | 20.4 |
| HuBERT [6] [1] | | ✓ | | | | | 17.0 | 18.9 | 20.3 | 4.6 | 8.1 | 7.6 | 20.3 |
| WavLM [5] [1] | | ✓ | | ✓ | | | 17.5 | 19.2 | 20.9 | 5.4 | 7.2 | 8.2 | 20.9 |
| 会话式 WavLM [3] | 会话式语音（663 小时） | ✓ | | ✓ | 12 Transformer | 95M | 17.8 | 19.7 | 21.2 | 5.1 | 6.4 | 5.7 | 17.3 |
| w2v-BERT [2] [18] | 组装数据集（450 万小时） | ✓ | ✓ | | 24 配置 | 581M | **17.9** | **19.9** | **21.5** | 4.9 | 6.3 | 5.1 | **16.4** |
| 基线 / 复现 | - | - | | | - | - | 15.6 [15] / 16.4 | 19.0 [16] / 18.0 | 20.5 [17] / 20.4 | 6.2 | 8.1 | 7.6 | 22.0 |

直接从原始的 10 秒音频片段中提取表示，提取范围涵盖 Transformer（或 Conformer）的每一层，并对这些表示应用可学习加权平均（例如 $\sum \alpha_i = 1$，其中 $\alpha$ 表示与第 $i^{th}$ 个“former”层相关的权重）。这些权重（训练后见图1）有助于我们了解模型的哪些部分对当前任务贡献最大。为进行完整评估，我们在多样化的多领域 DIHARD 3 数据集[11]上训练分割模型。由于该数据集不提供训练 集，我们使用 DIHARD 3开发 集的 77% 作为训练数据集，其余 23% 作为真正的开发集。测试 集保持不变。为了获得全局说话人日志结果，我们使用 Wespeaker 工具包 [28]中的预训练 ResNet-34 嵌入模型提取说话人嵌入。分割模型（220 万个参数）在单块 V100-32GB 上训练约 100 个周期。

如表 I 所示，与基于 SincNet 的特征提取器相比，改用现成的自监督表示，平均而言性能相对提升了 8%，提升主要体现在虚警率上。这表明，自监督学习表示有助于分割模型判断说话人何时处于活动状态。比较在 LibriSpeech 上预训练的三个模型可以发现，HuBERT 和 wav2vec 2.0 的表现几乎不相上下。令人意外的是，在所有自监督学习模型中，WavLM 的表现最差，尽管训练期间观察到的指标显示，它在训练集和开发集上的表现都优于 HuBERT 和 wav2vec 2.0。相反，对话式 WavLM 带来了显著的性能提升（相比在 LibriSpeech 上预训练的模型，相对提升约 20%；相比 SincNet，提升 26%），并且改善了说话人日志错误率的全部三个组成部分。这一结果表明，与单人语音内容相比，通过专门在对话数据集上训练来自监督模型，使其适应特定领域，对提升性能起着重要作用。最后，w2v-BERT 凭借更大的架构（24 个 Conformer 层）以及用于预训练的更多样、数量更大的未标注音频内容（450 万小时），相对性能比对话式 WavLM 提升了 5%，比现成模型提升了 25%。

观察层激活权重（见图 1）可以发现，在 LibriSpeech 上预训练的自监督学习模型，其前几层基于 Transformer 的层激活程度最高（主要是第 1 、 2 和 3 层），这表明模型前段对说话人日志任务的贡献最大。wav2vec 2.0 和 HuBERT 的激活分数相近，而 WavLM 的激活更早、程度也更高，主要集中在第 1 和第 2 层。

与此相反，对话式 WavLM 主要激活架构的中上层（第 7、8、9 和 10 层）。此外，与其他模型相比，其激活似乎更为均匀。对说话人日志贡献最大的层位于模型顶部，这似乎对性能有显著影响，因为对话式 WavLM 的表现优于此前仅激活较早层的模型。这也表明，该模型更充分地利用了来自整个网络的表征。w2v-BERT 上观察到的激活分数进一步验证了这一假设。对于后者，可以观察到两个激活峰值，分别对应于对比式（第 10、11 和 12 层）以及

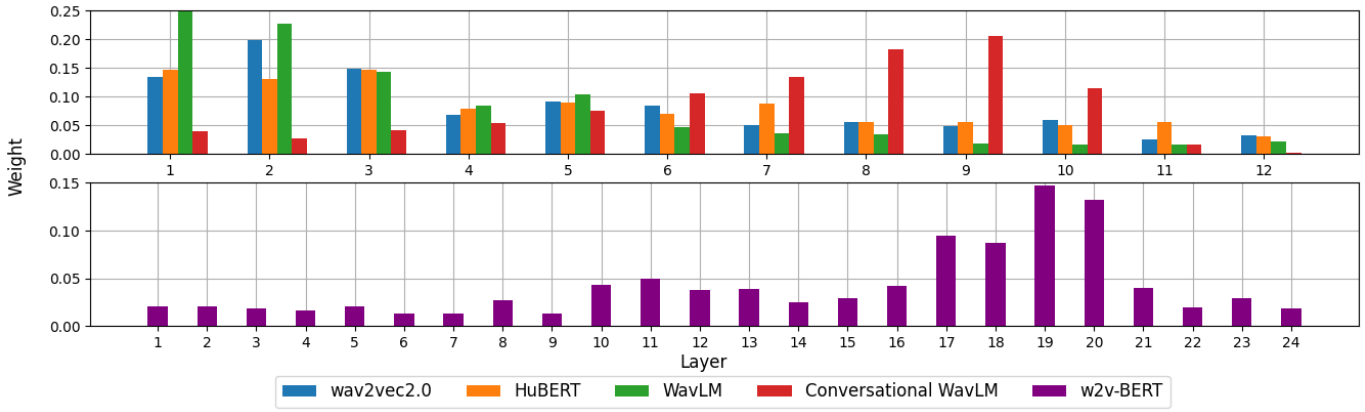


图1. 用于说话人日志任务的各自监督模型的逐层分析。

预测式（第 17、18、19 和 20 层）模型顶部区域，这表明 w2v-BERT 试图利用这两类预训练学习任务来解决说话人日志问题。根据这一自监督学习模型的结果，预测式学习任务似乎比对比式学习任务贡献更大。最后，自监督学习预训练数据集的差异对观察到的层激活有显著影响：与对话式 WavLM 或 w2v-BERT 相比，LibriSpeech 模型在说话人日志任务上的逐层激活模式截然不同。

## IV. 评估说话人分离

用于评估说话人分离的基线是时域音频分离网络 [14] (TasNet)。以往的框架依赖短时傅里叶变换 (STFT) 进行分离，但在处理混合语音 [29] 的幅度和相位时遇到了一些局限。因此，采用基于编码器—解码器架构的纯时域神经网络更适用于语音分离。TasNet架构包含一个直接作用于混合语音上的编码器（以一维卷积模块的形式实现）。随后，生成的潜在空间表示经由掩码网络处理，再通过转置一维卷积进行解码，如图3所示。

本研究评估了三种掩蔽网络，以考察不同的掩蔽策略能否在每种自监督学习架构中取得一致的结果。第一种是 ConvTasNet [15], , 其特点是包含多层深度卷积（沿帧方向应用）和逐点卷积（沿特征方向应用），并采用不同大小的卷积核和不同的步幅（用于捕获输入混合语音中的长程依赖）。第二种是 DPRNN [16], , 它首先将编码器的潜在表示分割成固定大小的块（块与块之间有 50% 的重叠），然后将这些块拼接起来，形成一个三维张量。随后，系统沿块轴和帧轴多次处理该张量，使用长短期记忆层处理编码器提供的特征和帧信息，兼顾局部与全局信息。为生成所需数量的掩码，系统使用一个 $Conv2D$ 层，将三维张量（沿特征维度）的大小扩展至待预测的声源数量。最后，在解码之前，系统使用 $overlap-add$ 将其重塑，为每位说话人重建一个输出。最后，SepFormer [17] 通过引入基于Transformer的模块取代长短期记忆层，改进了DPRNN，并利用注意力机制扩展三维张量在帧维度和块维度上的上下文窗口。

为了评估说话人分离任务的性能，我们采用源失真比提升 (SDRi) 作为指标，量化重建波形相对于原始信号的保真度。该指标通过从估计源和混合语音各自的SDR中相减得出。SDR的计算方法是取干净信号与估计失真之比的对数。用于计算SDR的失真，是根据真实信号与对应估计信号之间的残差误差推断得出的。

### A. 实现自监督表示

由于 TasNet 网络采用基于编码器-解码器的架构，因此自监督学习模型不能直接替代编码器，因为我们没有有效的方法将自监督学习表示解码为时域波形以进行重建。相反，在我们的设置中，自监督学习表示被拼接到编码器的潜在表示上，作为额外信息，如图3所示。

通常，TasNet 模型在最高编码帧率下表现更好 [16]（即较小的卷积核和步幅）。相反，自监督学习特征提取器会以固定帧率 $f_{SSL}$ 50 Hz 生成表示。为匹配两个模块的帧率并使其正确对齐，我们沿帧方向复制

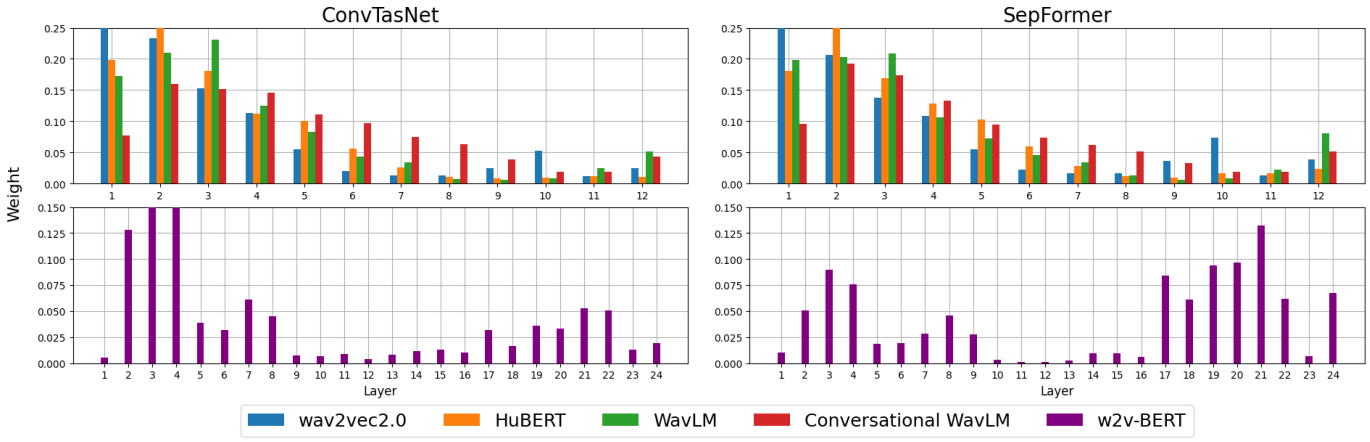


图 2. 使用两种掩码网络（ConvTasNet 和 SepFormer）对各自监督学习模型在说话人分离任务中的逐层表现进行分析。（DPRNN 呈现出与 ConvTas Net 类似的趋势，因此我们决定不予展示。）

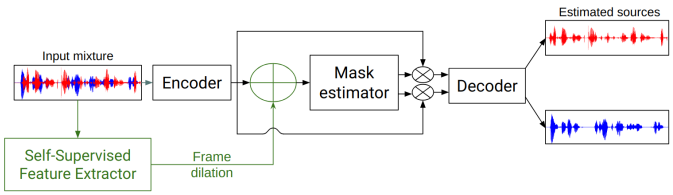


图 3. 使用自监督学习表示的 TasNet 架构

自监督学习特征，使其帧数与编码器相同。针对这项任务，我们在 WSJ0-2mix [12] 8kHz 数据集上评估每个模型。由于测试的自监督模型是在 16 kHz音频内容上预训练的，我们事先使用凯泽窗化 sinc 插值对混合音频进行上采样（从 8 kHz 提高到 16 kHz）。估计出声源后，我们再将其下采样回 8 kHz，以进行损失计算。

为生成自监督学习表示，我们采用与说话人日志相同的提取策略（对各层使用可学习加权平均）。我们使用 Asteroid [30] 和 Speechbrain[31] 库，根据提供的配方分别训练 ConvTasNet（5.1M 个参数）、DPRNN（14.8M 个参数）和 SepFormer（25.9M 个参数）200、50 和 100 个周期。训练 30 个周期后，我们采用耐心值为 3 个周期的学习率减半策略。每项实验均在单台 A40-48GB 上进行。

*B.* 基准测试结果

最终结果见分离部分的表 I。对于基线部分（不使用自监督学习表示），左侧列出的分数是各论文作者报告的 SDRi，右侧列出的分数则是我们按照各个库提供的基础配置方案自行训练后得到的结果。

结果表明，将自监督学习表示集成到掩码网络中，能够持续提升模型的分离能力。对于 ConvTasNet，使用 wav2vec 2.0 或 HuBERT 时，性能至少提升 0.6 dB（相对提升 3.7%）。与基础模型相比，WavLM 将 SDRi 提高了更多，达到 0.9 dB（相对提升 5.5%）。最后，使用对话式 WavLM 和 w2v-BERT 时取得了最佳结果，两种模型分别将 SDRi 提高了 1.4 dB 和 1.5 dB（相对提升 9%）。对于 DPRNN，增益更为显著：使用现成模型的自监督学习特征，SDRi 相较基线相对提升了 7%。值得注意的是，WavLM 和 wav2vec 2.0 的性能相近（19.2 dB），而 HuBERT 的表现略逊一筹。与 ConvTasNet 的情况类似，对话式 WavLM 和 w2v-BERT 进一步扩大了其与现成模型之间的差距（与基线相比，性能分别相对提升 9% 和 11%）。然而，对于 SepFormer架构，集成 wav2vec 2.0 和 HuBERT 的特征时，明显出现了瓶颈。具体而言，加入这些特征后，性能反而低于基础 SepFormer。此外，使用更适配的表示（例如 WavLM 或对话式 WavLM）所带来的增益，也明显小于其他分离网络中的增益。例如，使用 WavLM 时，SDRi 相较基线仅提升 0.5 dB，而其他模型的增益均超过 1 dB。这表明，对于更复杂的基于掩蔽的分离网络，这些表示中编码的信息可能存在冗余或多余。

最后，通过查看语音分离任务的逐层分析（图 2），我们观察到，与说话人日志任务类似，

仅考察现成模型时：激活程度最高的层位于模型靠前的部分。然而，对于对话式 WavLM 和 w2v-BERT，无法进行相同的比较，因为我们观察到的激活情况与使用 LibriSpeech 预训练的自监督学习模型所观察到的情况类似。对话式 WavLM 在说话人日志任务中主要激活第 7、8、9 和 10 层；而在语音分离任务中，使用最多的是第 2、3、4 和 5 层。

w2v-BERT也呈现出类似的行为。由于说话人日志和语音分离都是与说话人身份相关的任务，人们可能会认为这两个任务激活的层相似，但对于在对话数据集上预训练的模型而言，情况似乎并非如此。一种假设是，后者对说话人轮次（以及预测被掩蔽对话片段）更加敏感，而这些内容在说话人日志中发挥着重要作用。另一方面，说话人分离任务仅处理混合语音，其所依赖的更多是语音的低级声学特征，而非对话中的长期模式。

尽管逐层分析揭示了本研究考察的各种掩蔽网络中存在一致的趋势，但我们观察到，对于 w2v-BERT，在比较 SepFormer 与 ConvTasNet 时，预测层的激活程度高于对比式层。这表明，通过预测式自监督学习任务学得的表示可能比通过对比学习获得的表示与 SepFormer 更契合。

## V. 结论

本文研究了将自监督语音表征用于说话人日志和语音分离任务。我们首先回顾了与语音相关的先进自监督学习架构（wav2vec 2.0、HuBERT、WavLM、w2v-BERT 和对话式 WavLM），以及各任务对应的下游模型（用于说话人日志的 EEND 分割模型和用于语音分离的 TasNet），随后对若干最具代表性的自监督模型进行了性能研究。结合逐层分析，我们发现，预训练数据集、学习任务以及下游模型的类型都会显著影响性能。尽管对话式 WavLM 和 w2v-BERT 在预训练所用语音时长（663 小时对 450 万小时）和参数量（9500 万对 5.81 亿）方面差异巨大，但二者在这两项任务上的表现相近，这凸显了所用数据类型比数据量更为重要。分析还发现，某些自监督学习模型在采用更有效的掩蔽网络进行语音分离时，性能会有所下降。最后，通过使用 DIHARD3 进行说话人日志评估、使用 WSJ0 进行语音分离评估，我们能够有效地评估模型性能，避免数据集偏差，并深入了解观察到的性能提升能否迁移至更新的架构 [17], [32]。

为进一步探究对话式自监督学习模型在说话人日志任务中观察到的层贡献与语音分离任务中的差异，值得识别各层能够生成的信息类型（语音学或语义信息），并将其与现成模型进行比较。一种假设是，从现成模型（如 WavLM）和对话式模型（如对话式 WavLM 或 w2v-BERT）中提取的表征具有互补性。另一个值得关注的问题是分离任务中采用的特征融合技术。虽然在本研究对该特定任务的评估中，我们主要使用与 TasNet 编码器进行特征拼接的技术，但未来工作应考虑更复杂且性能更优的方法，例如交叉注意力或特征级线性调制（FiLM） [33] ，因为这些方法可能会使自监督学习模型各层产生不同的结果和表现。最后，本文提出的逐层分析为将这些发现应用于整合说话人日志和语音分离的联合模型（如 [34]）开辟了有趣的可能性。

## VI. 致谢

本研究获得了 GENCI 分配的 AD011014044R2 和 2024- AD011015163 配额下 IDRIS 高性能计算资源的使用权限，并得到国防创新局 2022 650079 资助项目的支持。此外，本研究还获得了法国国家研究署通过 ANR-20-CE23-0012-01 (MIM) 项目提供的支持。

## 参考文献

[1] S. wen Yang, P.-H. Chi, Y.-S. Chuang, C.-I. J. Lai, K. Lakhotia, Y. Y. Lin, A. T. Liu, J. Shi, X. Chang, G.-T. Lin, T.-H. Huang, W.-C. Tseng, K. tik Lee, D.-R. Liu, Z. Huang, S. Dong, S.-W. Li, S. Watanabe, A. Mohamed, and H. yi Lee, 《SUPERB：语音处理通用性能基准》，载于*Interspeech 2021*论文集，2021年，第1194–1198页。[2] J. Han, F. Landini, J. Rohdin, A. Silnova, M. Diez, and L. Burget, 《利用自监督学习进行说话人日志》，载于*ICASSP*论文集，2025年。[3] S. Baroudi, T. Pellegrini, and H. Bredin, 《面向说话人分割的自监督语音表征专门化》，载于国际语音通信协会第*25*届年会*INTERSPEECH 2024*，2024年9月，第3769–3773页。[4] Z. Chen, N. Kanda, J. Wu, Y. Wu, X. Wang, T. Yoshioka, J. Li, S. Sivasankaran, and S. E. Eskimez, 《利用大规模自监督学习进行语音分离》，载于*ICASSP 2023——2023*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2023年，第1–5页。

[5] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, J. Wu, L. Zhou, S. Ren, Y. Qian, Y. Qian, M. Zeng, X. Yu, and F. Wei, 《Wavlm：面向全栈语音处理的大规模自监督预训练》，*IEEE*信号处理精选主题期刊，第16卷，第1–14页，2022年10月。[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, 《Hubert：通过隐藏单元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，第29卷，第3451–3460页，2021年。[7] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, 《wav2vec 2.0：语音表征自监督学习框架》，载于第*34*届神经信息处理系统国际会议论文集，NIPS'20系列。美国纽约州Red Hook：Curran Associates Inc.，2020年。[8] S.-w. Yang, H.-J. Chang, Z. Huang, A. T. Liu, C.-I. Lai, H. Wu, J. Shi, X. Chang, H.-S. Tsai, W.-C. Huang, T.-h. Feng, P.-H. Chi, Y. Y. Lin, Y.-S. Chuang, T.-H. Huang, W.-C. Tseng, K. Lakhotia, S.-W. Li, A. Mohamed, S. Watanabe, and H.-y. Lee, 《语音基础模型的大规模评估》，*IEEE/ACM*音频、语音与语言处理汇刊，第32卷，第2884–2899页，2024年。[9] J. Cosentino, M. Pariente, S. Cornell, A. Deleforge, and E. Vincent, 《LibriMix：一个用于泛化语音分离的开源数据集》，2020年5月，工作论文或预印本。 [在线]。可访问：https://inria.hal.science/hal-03354695[10] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, 《LibriSpeech：基于公共领域有声读物的自动语音识别语料库》，载于*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，第5206–5210页。 [在线]。可访问：http://ieeexplore.ieee.org/document/7178964/[11] N. Ryant, P. Singh, V. Krishnamohan, R. Varma, K. Church, C. Cieri, J. Du, S. Ganapathy, and M. Liberman, 《第三届Dihard说话人日志挑战赛》，载于*Interspeech 2021*，2021年，第3570–3574页。[12] J. R. Hershey, Z. Chen, J. Le Roux, and S. Watanabe, 《深度聚类：用于分割与分离的判别式嵌入》，载于*2016*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2016年，第31–35页。[13] H. Bredin, 《pyannote.audio 2.1说话人分离流水线：原理、基准测试与实践指南》，载于*INTERSPEECH 2023*论文集，2023年，第1983–1987页。[14] Y. Luo and N. Mesgarani, 《Tasnet：用于实时单通道语音分离的时域音频分离网络》，载于*2018*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2018年，第696–700页。[15] ——, 《Tasnet：在语音分离方面超越理想时频掩蔽》，*CoRR*，第abs/1809.07454卷，2018年。 [在线]。可访问：http://arxiv.org/abs/1809.07454[16] Y. Luo, Z. Chen, and T. Yoshioka, 《双路径循环神经网络：用于时域单通道语音分离的高效长序列建模》，*ICASSP 2020——2020*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），第46–50页，2019年。 [在线]。可访问：https://api.semanticscholar.org/CorpusID:204575939[17] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi, and J. Zhong, 《语音分离只需注意力机制》，载于*ICASSP 2021——2021*年*IEEE*声学、语音与信号处理国际会议（*ICASSP*），2021年，第21–25页。[18] Y. Chung, Y. Zhang, W. Han, C. Chiu, J. Qin, R. Pang, and Y. Wu, 《W2v-bert：结合对比学习与掩码语言建模进行自监督语音预训练》，*CoRR*，第abs/2108 06209卷，。2021 [在线]。可访问：https://arxiv.org/abs/2108 06209。[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, 《BERT：面向语言理解的深度双向变换器预训练》，载于*2019*年计算语言学协会北美分会会议：人类语言技术论文集，第*1*卷（长短论文），J. Burstein、C. Doran和T. Solorio编。明尼阿波利斯，明尼苏达州：计算语言学协会，2019年6月，第4171–4186页。 [在线]。可访问：https://aclanthology.org/N19-1423/[20] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, 《Conformer：用于语音识别的卷积增强Transformer》，载于 Interspeech 2020，2020年，第5036–5040页。[21] A. Baevski, S. Schneider, and M. Auli, 《vq-wav2vec：离散语音表征的自监督学习》，载于第8届国际学习表征会议（ICLR）论文集，2020年。[在线]。网址：https://openreview.net/forum?id=rylwJxrYDS[22] Y. Fu, L. Cheng, S. Lv, Y. Jv, Y. Kong, Z. Chen, Y. Hu, L. Xie, J. Wu, H. Bu, X. Xu, J. Du, and J. Chen, 《AiSHELL-4：面向会议场景下语音增强、分离、识别和说话人日志的开源数据集》，载于 Interspeech 2021，2021年，第3665–3669页。[23] J. Carletta, 《AMI会议语料库发布公告》，《ELRA通讯》，第11卷第1期，第3–5页，2006年。 [在线]。网址：http://groups.inf.ed.ac.uk/ami/corpus/overview.shtml[24] H. Bredin, 《pyannote.audio 2.1说话人分离流水线：原理、基准测试与实践指南》，载于 INTERSPEECH 2023论文集，2023年，第1983–1987页。[25] K. Kinoshita, M. Delcroix, and N. Tawara, 《面向真实对话语音的端到端神经网络与基于聚类的说话人分离集成技术进展》，载于 Interspeech 2021，2021年，第3565–3569页。[26] M. Ravanelli and Y. Bengio, 《利用原始波形和SincNet进行说话人识别》，载于2018年IEEE口语技术研讨会（SLT），2018年，第1021–1028页。[27] A. Plaquet and H. Bredin, 《用于神经网络说话人分离的幂集多类交叉熵损失》，载于 INTERSPEECH 2023论文集，2023年，第3222–3226页。[28] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xiang, Y. Deng, and Y. Qian, 《WeSpeaker：面向研究与生产的说话人嵌入学习工具包》，载于ICASSP 2023——2023年IEEE声学、语音与信号处理国际会议（ICASSP），2023年，第1–5页。[29] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, 《利用深度循环神经网络进行相位敏感且识别增强的语音分离》，载于2015年IEEE国际声学、语音与信号处理会议，2015年，第708–712页。[30] M. Pariente, S. Cornell, J. Cosentino, S. Sivasankaran, E. Tzinis, J. Heitkaemper, M. Olvera, F.-R. St¨oter, M. Hu, J. M. Mart´ın-Do˜nas, D. Ditter, A. Frank, A. Deleforge, and E. Vincent, 《Asteroid：面向研究人员的基于PyTorch的音频源分离工具包》，载于 Interspeech 2020，2020年，第2637–2641页。[31] M. Ravanelli, T. Parcollet, A. Moumen, S. de Langen, C. Subakan, P. Plantinga, Y. Wang, P. Mousavi, L. D. Libera, A. Ploujnikov, F. Paissan, D. Borra, S. Zaiem, Z. Zhao, S. Zhang, G. Karakasidis, S.-L. Yeh, P. Champion, A. Rouhe, R. Braun, F. Mai, J. Zuluaga-Gomez, S. M. Mousavi, A. Nautsch, H. Nguyen, X. Liu, S. Sagar, J. Duret, S. Mdhaffar, G. Laperri`ere, M. Rouvier, R. D. Mori, and Y. Est`eve, 《SpeechBrain 1.0：开源对话式人工智能》，《机器学习研究期刊》，第25卷第333期，第1–11页，2024年。 [在线]。网址：http://jmlr.org/papers/v25/24-0991.html[32] S. Zhao and B. Ma, 《MossFormer：利用门控单头Transformer和卷积增强联合自注意力，突破单声道语音分离的性能上限》，载于ICASSP 2023——2023年IEEE声学、语音与信号处理国际会议（ICASSP），2023年。[33] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, 《FiLM：采用通用条件层的视觉推理》，载于第32届AAAI人工智能会议（AAAI-18）论文集。AAAI出版社，2018年，第3942–3951页。 [在线]。网址：https://ojs.aaai.org/index.php/AAAI/article/view/[34] J. Kalda, C. Pages, R. Marxer, T. Alumae, and H. Bredin, 《PixIT：基于真实场景多说话人录音联合训练说话人分离与语音分离》，载于说话人与语言识别研讨会（Odyssey 2024），2024年，第115–122页。