arXiv:2304.00871v1 [eess.AS] 3 Apr 2023

# 自监督学习 习驱动的会议数据声源分离 A

*Yuang Li*、*Xianrui Zheng**、*Philip C. Woodland*

剑桥大学工程系，特朗普顿街，剑桥，CB2 1PZ，英国{yl807、xz396、pcw}@eng.cam.ac.uk

**摘要**

在多方会议场景中，声源分离可以从重叠语音中提取单说话人信号，从而提升自动语音识别（ASR）效果。尽管自监督学习模型在单通道声源分离方面取得了成功，大多数研究仍聚焦于模拟设置。本文对七种自监督学习模型在模拟语料库和真实语料库上的表现进行了比较。随后，我们提出一种新颖的迭代声源选择方法，将表现最佳的 WavLM 模型集成到自动转录系统中。为提升真实场景下的性能，我们将时域无监督混合不变训练改编至时频域。实验表明，在转录系统中，如果将声源分离置于经分离语音微调的自动语音识别模型之前，则在 AMI 开发集和测试集上，未知说话人数情况下的拼接最小排列词错误率（cpWER-us）分别绝对降低了 1.9% 和 1.5%。

索引词— 声源分离、自监督学习、自动语音识别

## 1. 引言

日常对话中，语音重叠十分常见，会降低语音可懂度，并给自动语音识别系统带来重大挑战。为解决这一问题，可以采用声源分离来估计单说话人信号。早期研究（如深度聚类 [1] 和话语级置换不变训练 [2] ）在时频（T-F）域处理单通道源分离，并忽略相位信息，因此任务是估计单说话人的幅度谱。近年来，时域方法因能够联合建模幅度和相位而日益流行。从 Tas-Net [3] 直接通过编码器-解码器结构预测波形开始，人们提出了各种时域架构 [4, 5]。迄今为止，基于Transformer的模型 [6, 7]展现出最有前景的性能。

有监督训练需要以单说话人信号作为目标，但从真实的重叠信号中获取此类参考信号颇具挑战，因此一种常见做法是组合多个单说话人片段 [1]。为提高真实感，人们加入了噪声和混响 [8, 9], ，但真实数据与合成数据之间的差异仍然是个问题。为了利用真实的重叠语音，研究人员提出了无监督训练方法，这些方法使用训练良好的分离模型生成伪标签 [10] ，或利用分离声源与混合信号之间的一致性 [11]。Sivaraman 等人[12] 表明，无监督混合不变训练（MixIT）在 AMI 上效果良好，但评估中未纳入基于自动语音识别的评估，且他们关注的是时域模型，而非基于时频域掩蔽的模型。

自监督学习（SSL）通过从大量未标注信号中学习，使各种语音相关下游任务受益 [13]。以 Conformer 网络 [14]作为下游模型，WavLM [15] 在 LibriCSS [16]上取得了最先进的分离性能。Huang 等人[17] 进一步在 LibriMix [18] 上比较了 13 种自监督学习模型，以展示它们的有效性。然而，这些比较仅限于模拟数据集。此外，自监督学习模型处于冻结状态，仅用于生成声学表示。

本文旨在全面分析自监督学习模型在真实场景源分离任务中的表现。针对传统的逐话语分离任务，我们使用基于信号和基于自动语音识别的评估标准，在混响版本的 LibriMix [18]（R-LibriMix）、LibriCSS [16], 和 AMI [19] 上比较了七种自监督学习模型。通过两阶段微调方案充分发挥自监督学习模型的潜力：先训练轻量级下游模型，再微调自监督学习模型。随后，我们将实验扩展到一项真实场景应用：会议转录系统。为缓解领域不匹配，我们引入了采用相位敏感掩蔽（PSM）的时频域无监督 MixIT，使用真实重叠数据对模型进行微调。分离模型会产生多个输出声源，需要根据说话人信息自动选出目标声源。为此，所提出的声源选择方法根据说话人分离提供的说话人标签，迭代计算并优化平均说话人嵌入。

本文其余部分安排如下：第 2 节回顾 SSL 模型。第 3 节介绍源分离框架、监督式和无监督式训练方法，以及本文提出的迭代式声源选择方法。第 4 节和第 5 节介绍实验设置和结果。第 6 节总结全文。音频样本可公开获取 [1]。

## 2. 语音信号的 SSL 模型

自监督学习模型可为下游任务提供信息丰富的表征，并为进一步微调提供良好的初始化。这些模型可分为两类：生成式模型会重建原始特征，判别式模型则会估计通过聚类或量化生成的离散化语音表示。**TERA** [20] 是一种自监督 Transformer 编码器，通过生成式任务进行训练；训练过程中，输入频谱图会被随机遮蔽。该模型通过 L1 损失优化，以重建原始频谱图。**Wav2Vec2** [21] 使用

*Zheng 获得了亚马逊奖学金。本研究使用剑桥大学研究计算服务运营的剑桥 Tier-2 系统资源（www.hpc.cam.ac.uk）完成，该系统由 EPSRC Tier-2资本拨款EP/T022159/1资助。

[1]https://sites.google.com/view/ssl-ss/ho me



© 电气电子工程师学会 2023

原始波形作为输入，并使用潜在空间中的遮蔽特征。它引入了一个量化模块，将潜在特征离散化为训练目标。模型通过对比学习任务进行训练，需要从一组干扰项中辨别出被遮蔽时间步上的真实量化特征。**HuBert** [22] 采用离线聚类方法生成离散标签。模型估计这些伪标签的分布，然后在遮蔽区域计算交叉熵损失。**UniSpeech-SAT**和**WavLM** [23, 15] 是 HuBert 的变体，旨在促进说话人身份的提取。它们共有的一项关键特征是采用话语混合增强，即将主要话语与随机选取的另一位说话人的片段混合。

## 3. 方法

### 3.1. 用于声源分离的 SSL 表征

基于 SSL 的声源分离框架如图 1 所示。SSL 模型处理时域波形并生成 SSL 表征。随后，为了利用多层级信息，使用加权和 [17] 合并来自每个 SSL Transformer 层的特征。合并特征后，通过最近邻插值阶段提高 SSL 表征的时间分辨率，使其与频谱图具有相同的时间分辨率。最后，单个 Conformer 模块估计多个时频掩码，用于恢复分离后的信号。Conformer 模块的参数量显著少于先前研究中的下游模型 [15, 17]，因此 SSL 模型在声源分离中发挥着关键作用。请注意，由于 SSL 特征与频谱图具有相似的步幅，而时域模型往往采用小得多的步幅，因此本文未使用时域模型。

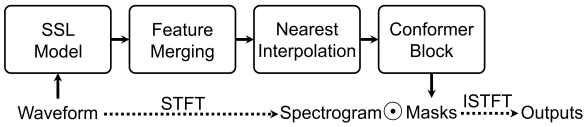


**图 1.** 使用 SSL 表征进行时频域声源分离的框架。

为优化分离模型，标准方法采用有监督的置换不变训练（PIT） [2]，通过在所有可能的置换中选取损失最小值进行训练。本文中，当模型在合成数据集上训练时，按照公式（1）计算损失，其中使用了相位敏感目标 [24] 频谱图。

$$\mathcal{L}_{\text{PIT}} = \min_{\phi \in P} \sum_{(i,j)\in\phi} ||\boldsymbol{M}_i \odot |\boldsymbol{Y}| - |\boldsymbol{X}_j| cos(\theta_Y - \theta_{X_j})||_F^2 \quad (1)$$

其中，$\phi$ 表示一种置换；$P$ 是所有可能置换的集合；$\boldsymbol{Y}$ 是时频域混合信号；$\boldsymbol{M}_i$ 是一个输出掩码；$\boldsymbol{X}_j$ 是目标频谱图；$\theta_Y$ 和 $\theta_{X_j}$ 分别是 $\boldsymbol{Y}$ 和 $\boldsymbol{X}_j$ 的相位；$||\cdot||_F$ 是弗罗贝尼乌斯范数。

### 3.2. 时频域混合不变训练（MixIT）

PIT 仅适用于合成数据，而 MixIT 可以将真实混合音频作为标签，前提是输出声源能够重新混合成这些标签。MixIT 首先通过组合真实的重叠信号来创建混合物的混合（MoMs）。然后，模型将 MoMs 作为输入，并估计多个音频流。为计算损失，MixIT 会穷举搜索误差最小的重新混合方式。

最初的 MixIT 流程用于时域源分离，在这种情况下，可以通过将波形样本相加来重混合输出；但在时频域中，由于存在相位差，直接将幅度相加并不恰当。因此，MixIT 结合 PSM 进行了适配，以用于时频域，因为时域重混合等同于相位敏感目标的重混合，所以损失函数变为公式（2），其形式与 PIT 类似。区别在于，预测结果是一个混合信号（$\sum_{i\in\mathbf{I}} \boldsymbol{M}_i \odot |\boldsymbol{Y}|$），而不是单个频谱（$\boldsymbol{M}_i \odot |\boldsymbol{Y}|$）。

$$\mathcal{L}_{\text{MixIT}} = \min_{\phi \in \mathcal{M}} \sum_{(\mathbf{I},j)\in\phi} ||(\sum_{i\in\mathbf{I}} \boldsymbol{M}_i \odot |\boldsymbol{Y}|) - |\boldsymbol{X}_j| cos(\theta_Y - \theta_{X_j})||_F^2 \quad (2)$$

其中 $\phi$ 表示一次重混音；$\mathcal{M}$ 表示所有可能的重混音；$\mathbf{I}$ 表示一组应进行组合的输出流；$\boldsymbol{Y}$ 表示时频域中的混合物的混合；$\boldsymbol{M}_i$ 表示一个输出掩码；$\boldsymbol{X}_j$ 表示一个目标频谱图；$\theta_Y$ 和 $\theta_{X_j}$ 分别表示 $\boldsymbol{Y}$ 和 $\boldsymbol{X}_j$ 的相位。

MixIT 并未明确引导模型将单个说话人的信号分离到一个输出流中，这可能导致过度分离，因此，结合 PIT 与 MixIT 的半监督学习更加有效。

### 3.3. 迭代式声源选择

与不含目标说话人的声源分离相比，说话人提取是一项更贴近实际的任务，在该任务中，模型需要从混合信号中估计目标说话人的语音。为了识别目标说话人，通常会将辅助分支从注册音频中提取的说话人嵌入注入主网络 [25]。在本文中，说话人提取被解耦为声源分离和声源选择。对于声源选择，假设话语来自能够提供说话人标签的说话人分离系统。因此，可以根据同一说话人的话语得到平均说话人嵌入，随后利用该嵌入从输出声源中进行选择。然而，这些话语受到重叠数据的污染，因此需要对说话人嵌入进行迭代式优化，并移除部分离群值。具体步骤如下：

- **步骤 1**：计算每个话语的说话人嵌入。
- **步骤 2**：计算每位说话人的平均嵌入。
- **步骤 3**：对于每位说话人，剔除部分与平均嵌入的欧氏距离较大的离群值。
- **步骤 4**：计算剔除离群值后的平均说话人嵌入。
- **步骤 5**：选择嵌入与平均说话人嵌入的余弦相似度最高的输出声源。
- **步骤 6**：计算所选输出声源的嵌入。
- **步骤 7**：返回步骤 2。

借助声源选择，无需修改架构或重新训练，即可直接应用分离模型进行说话人提取。此外，我们的方法无需注册音频，且可以轻松集成到转录系统中。

## 4. 实验设置

### 4.1. 数据集与配置

在 PIT 实验中，本文通过 imagemethod [26]模拟混响，将 LibriMix [18]扩展为 R-LibriMix。

并添加各向同性噪声 [27]。混响信号被用作目标，因此模型专注于分离。在 MixIT 实验中，使用了 AMI [19]远端麦克风录制的波束形成音频。通过将信噪比（SNR）设为 -5dB 至 5dB，生成了 MoMs。为便于同时进行域内基于信号和真实场景中基于自动语音识别的比较，评估采用了合成的 R-LibriMix、模拟的 LibriCSS [16] 以及真实的 AMI。

下游模型由单个 Conformer 模块 [28]组成，注意力维度为 256，注意力头为 4 个，隐藏状态维度为 1024。监督训练（PIT）的输出掩码数量为 2，半监督训练（PIT&MixIT）的输出掩码数量为 4。半监督训练中使用 Softmax 替代 Sigmoid 激活函数，因为它在掩码之间引入了连接，这与 MixIT 的假设一致，有助于更快、更稳定地收敛。前端短时傅里叶变换（STFT）使用长度为 400 点的汉宁窗，步长为 160 点。微调计划分为两个阶段。第一阶段仅更新 Conformer 模块，训练 100k 步。第二阶段解冻自监督学习模型，并再微调 80k 步。半监督训练中，以 80% 的概率使用 MixIT，并从 AMI 中采样批次；否则，使用 PIT，并从 R-LibriMix 中采样批次。

### 4.2. 转录系统

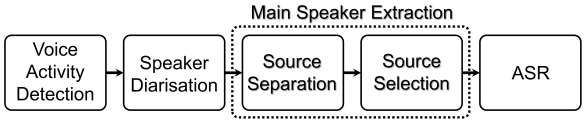


**图 2**。自动转录系统流程图。

图 2 展示了转录系统的流程图。首先，通过语音活动检测（VAD）和说话人分离，将连续录音裁剪为话语并标注说话人标签。随后，通过声源分离和选择移除重叠语音，并保留主要说话人的语音信号。最后，使用自动语音识别模型生成转录文本。具体如下：**1）语音活动检测（VAD）**：在 Wav2Vec2 上添加线性层，并接入 Softmax 激活函数，以输出语音成分存在或不存在的概率；随后在 AMI 上对模型进行微调。语音区域之间的时间间隔被限制为大于 0.04 秒，与 [29] 中的流程相同。**2）说话人分离**：ECAPA-TDNN [30]以 3 秒的窗口长度和 1.5 秒的步长生成说话人嵌入。随后采用谱聚类，最多识别 10 位说话人。**3）声源分离**：选择在 R-LibirMix 和 AMI 上通过半监督训练进行微调的 WavLM 模型。**4）声源选择**：使用基于 ECAPA-TDNN 生成的说话人嵌入进行迭代式声源选择。在 AMI 开发集上发现，词错误率在经过两次迭代后显著下降，此后基本保持不变。因此，为提高效率，将迭代次数设为两次。此外，还移除了 60% 的离群值，以稳定收敛。**5）自动语音识别**：比较了三个以 Wav2Vec2-Robust [31] 为基础模型的自动语音识别模型：W2V2-SWB 在 Switchboard [32] 上进行微调；W2V2-AMI 在 AMI 上进行微调；W2V2-AMI-Sep 则是在 W2V2-AMI 的基础上进一步微调，并使用来自 AMI 的分离后音频。

### 4.3. 评估

在对自监督学习模型进行初步比较时，采用了尺度不变信噪比（SI-SNR） [3] 和词错误率（WER）。R-LibriMix 提供了参考信号，因此计算了所有可能排列中的最高 SI-SNR。对于 LibriCSS [2] 和 AMI [3]，由于只有话语的转录文本可用，因此找出了使用分离声源时可能达到的最低词错误率。

通过与理想选择进行 WER 比较，评估声源选择 [4]。此外，还计算了选择准确率，其定义为估计选择与理想选择相匹配的时长百分比。

对于转录系统，由于语音活动检测所提供的片段与参考转录中话语的起止时间可能相差很大，因此难以计算传统 WER。因此，采用了会话级别的 cpWER-us [29]。首先，按时间顺序分别连接每个会话中各说话人的假设转录和参考转录。然后，计算所有可能的说话人排列对应的 WER，并移除冗余说话人。最后，选取其中最低的 WER 作为 cpWER-us。

## 5. 实验结果

### 5.1. 自监督学习模型之间的比较

如表 1 所示，仅训练下游模型时，所有自监督学习模型在 R-LibriMix 和 LibriCSS 上的表现都优于频谱图。然而，在 AMI 开发集和测试集上，频谱图的词错误率分别为 38.1% 和 40.5%，低于 Wav2Vec2 和 Hu- Bert；后两者在 AMI 开发集上的词错误率分别为 38.7% 和 38.3%，在 AMI 测试集上则均为 41.0%。TERA 在 R-LibriMix 上的表现略逊于 Wav2Vec2 和 HuBert，但在 AMI 开发集和测试集上的表现更好，词错误率分别为 37.8% 和 40.3%；这可能是因为用于预训练 TERA 的频谱变化提升了其泛化能力。这些比较表明，尽管 R-LibriMix 和 LibriCSS 是为模拟会议录音而创建的，但由于噪声类型、混响和语音内容不同，其结果仍可能与真实语料库不一致。如果自监督学习模型仅使用朗读文本的干净录音进行训练，那么它在真实的重叠对话场景中可能泛化不佳。即使没有扩大预训练数据集，UniSpeech-SAT 和 WavLM 在所有数据集上的表现也都显著优于其他模型。在 AMI 语料库上，UniSpeech-SAT 在开发集和测试集上的词错误率分别为 36.6% 和 39.6%，而 WavLM 的词错误率略高，分别为 36.8% 和 40.1%。可以得出结论，话语混合使模型能够同时接触多个说话人，这是提取用于声源分离的稳健表征的关键。

对 SSL 模型进行微调后，除 TERA 外，所有 SSL 模型的可训练参数均从 1.8M 大幅增加至 92M，并观察到一致的性能提升。TERA 的参数量最少，仅为 23M，因此提升幅度最小。通过微调实验可以得出结论：解冻 SSL 模型是有益的，因为这使模型能够学习任务特定表示；同时，由于 AMI 上的跨数据集性能也得到了提升，因此不会导致过拟合。尽管所有 SSL 模型都针对源分离进行了直接优化，

[2]LibriCSS 的自动语音识别模型：https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self[3]AMI 的自动语音识别模型：https://huggingface.co/facebook/wav2vec2-large-robust-ft-swbd-300h[4]理想选择是指选择转录内容与真实标注最匹配的声源

| 模型 | R-LibriMix 尺度不变信噪比↑ | LibriCSS WER↓ | AMI WER↓ |
|---|---|---|---|
| 无分离 | -0.43 / -0.44 | 14.7 | 41.2 / 43.8 |
| 频谱图 | 3.15 / 2.96 | 12.4 | 38.1 / 40.5 |
| 阶段 1：训练下游模型。 | | | |
| TERA [20] | 4.80 / 4.60 | 11.2 | 37.8 / 40.3 |
| Wav2Vec2 [21] | 5.31 / 5.18 | 11.4 | 38.7 / 41.0 |
| HuBert [22] | 5.12 / 4.93 | 11.4 | 38.3 / 41.0 |
| UniSpeech-SAT [23] | 6.89 / 6.83 | 8.0 | 36.6 / 39.6 |
| UniSpeech-SAT+ [23] | 7.01 / 6.87 | 7.9 | 36.4 / 39.3 |
| WavLM [15] | 6.51 / 6.36 | 7.7 | 36.8 / 40.1 |
| WavLM+ [15] | **7.40 /7.27** | **7.4** | **36.2 /38.9** |
| 第二阶段：微调基础模型。 | | | |
| TERA [20] | 5.86 / 5.67 | 10.5 | 37.6 / 40.3 |
| Wav2Vec2 [21] | 6.74 / 6.59 | 9.8 | 37.3 / 39.7 |
| HuBert [22] | 6.67 / 6.50 | 9.8 | 37.2 / 39.9 |
| UniSpeech-SAT [23] | 7.73 / 7.65 | 7.1 | 35.5 / 38.6 |
| UniSpeech-SAT+ [23] | 7.83 / 7.76 | **6.7** | 35.5 / 38.3 |
| WavLM [15] | 7.80 / 7.69 | 6.8 | 35.6 / 38.6 |
| WavLM+ [15] | **8.05 /7.95** | **6.7** | **34.7 /38.0** |

**表 1**。自监督学习模型之间的比较。对于 R-LibriMix 和 AMI，给出了开发集/测试集的尺度不变信噪比或词错误率。对于 Lib- riCSS，给出了所有数据子集的平均词错误率。标记为"+"的模型是在包含 94k 小时语音数据集上预训练的。其他模型则是在包含 960 小时 LibriSpeech 数据集上预训练的。

UniSpeech-SAT 和 WavLM 仍明显优于 Wav2Vec2 和 HuBert，这表明，即使对于源分离这类可以自动生成大型合成数据集的任务而言，从 SSL 任务进行初始化也至关重要。此外，如果将 WavLM+ 与 UniSpeech-SAT+ 进行比较，可以看出 WavLM+ 从大型预训练数据集中获益更多，因为在大多数情况下，WavLM 与 WavLM+ 之间的性能差距更大。例如，在 AMI 开发集和测试集上，WavLM 到 WavLM+ 的绝对词错误率提升分别为 0.9% 和 0.6%。然而，对于 UniSpeech-SAT，在 AMI 测试集上观察到的绝对提升仅为 0.3%。总体而言，WavLM+ 是性能最佳的模型，因此本文其余部分均使用该模型。

### 5.2. MixIT 和声源选择的结果

表 2 对比了 PIT 与半监督 PIT&MixIT。使用理想声源时，PIT&MixIT 在 AMI 开发集和测试集上的词错误率均有所改善，绝对降幅约为 1%。这是因为 MixIT 利用了真实的重叠语音，丰富了训练数据。此外，MixIT 并不局限于语音成分。使用四个输出声源时，咳嗽声和笑声等非语音成分有时会与语音信号分离开来。然而，采用迭代式声源选择时，MixIT 的效果有所下降，因为输出声源数量越多，声源选择就越困难。对于 PIT& MixIT，选择准确率约为 85%；而对于仅有两个声源的 PIT，选择准确率则高于 90%。因此，在真实场景中，不能使用传统的置换不变准则来评估输出声源数量不同的模型。分离性能与选择准确率之间的权衡更为重要。就此而言，半监督训练仍略胜一筹，其词错误率更低。

| 训练 | 理想声源 WER↓ | 迭代式选择 WER↓ | 迭代式选择 选择准确率↑ |
|---|---|---|---|
| PIT | 34.7 / 38.0 | 36.3 / 39.4 | 93.3 / 92.5 |
| PIT&MixIT | 33.6 / 37.0 | 35.5 / 39.1 | 87.0 / 85.4 |

**表 2**。AMI 上 PIT 与半监督PIT&MixIT 的比较。提供了开发集/测试集的词错误率和选择准确率。

| ASR | 分离 | cpWER-us↓ |
|---|---|---|
| W2V2-SWB | × | 43.9 / 46.0 |
| W2V2-SWB | ✓ | 40.8 / 43.8 |
| W2V2-AMI | × | 34.2 / 35.1 |
| W2V2-AMI | ✓ | 36.4 / 36.8 |
| W2V2-AMI-Sep | × | 33.5 / 34.4 |
| W2V2-AMI-Sep | ✓ | **32.3 /33.6** |

**表 3**。AMI 数据集上采用或未采用分离的转录系统对比。给出了开发集/测试集的 cpWER-us。

### 5.3. 转录系统的结果

表 3 展示了声源分离对转录系统的影响。对于 W2V2-SWB——一种使用非重叠语音训练的自动语音识别模型——在未进行声源分离时，其 cpWER-us 在 AMI 开发集和测试集上分别为 43.9% 和 46.0%。引入声源分离后，cpWER-us 降至 40.8% 和 43.8%。然而，如果自动语音识别模型在 AMI 上进行微调（W2V2-AMI），声源分离反而会损害性能，并导致删除错误大幅增加。一个可能的原因是，W2V2-AMI 接触过重叠数据，因此学会了忽略背景中的说话人。分离之后，更多语音成分可能会被错误地识别为背景。此外，分离模型还引入了域外数据，例如包含长时间静音和系统噪声的话语。为了解决这些问题，研究者使用分离后的数据对 W2V2-AMI 模型进行微调，得到 W2V2-AMI-Sep。与使用 W2V2-AMI 且未进行声源分离的系统相比，该模型在 AMI 开发集和测试集上的 cpWER-us 绝对降幅分别为 1.9% 和 1.5%，达到最低值。使用未分离的数据时，W2V2-AMI-Sep 的表现也更好；与 W2V2-AMI 相比，其在 AMI 开发集和测试集上的 cpWER-us 绝对降幅均为 0.7%，这表明重叠语音可能会影响自动语音识别模型的训练。

## 6. 结论

本文研究了七种用于声源分离的自监督学习模型。我们强调，评估时使用真实会议数据至关重要，因为模型在真实数据上的表现可能与在模拟数据集上的表现不一致。实验表明，自监督学习模型不仅可用作特征提取器，也可作为可训练的基础模型。然而，在干净的单说话人信号上训练的自监督学习模型（如 Wav2Vec2 和 HuBert）在真实数据上的泛化能力有限。对于转录系统，可在说话人日志之后、自动语音识别之前，直接通过所提出的迭代式声源选择方法引入声源分离。为最大限度地提升性能，我们采用 MixIT，利用真实混合语音进行训练，并进一步在分离后的音频上微调自动语音识别模型。因此，AMI 上的 cpWER-us 显著降低。

## 7. 参考文献


[1] J. R. Hershey、Z. Chen、J. Le Roux 和 S. Watanabe，“深度聚类：用于分割与分离的判别式嵌入”，*ICASSP*会议论文集，2016年。[2] M. Kolbæk、D. Yu、Z. Tan 和 J. Jensen，“采用深度循环神经网络话语级排列不变训练的多说话人语音分离”，《电气电子工程师学会音频、语音与语言处理汇刊》，第25卷，第10期，第1901–1913页，2017年。[3] Y. Luo 和 N. Mesgarani，“TasNet：用于实时单通道语音分离的时域音频分离网络”，*ICASSP*会议论文集，2018年。[4] Y. Luo 和 N. Mesgarani，“Conv-TasNet：超越语音分离中的理想时频幅度掩蔽”，《电气电子工程师学会音频、语音与语言处理汇刊》，第27卷，第8期，第1256–1266页，2019年。[5] Y. Luo、Z. Chen 和 T. Yoshioka，“双路径RNN：用于时域单通道语音分离的高效长序列建模”，*ICASSP*会议论文集，2020年。[6] C. Subakan、M. Ravanelli、S. Cornell、M. Bronzi 和 J. Zhong，“注意力就是语音分离所需的一切”，*ICASSP*会议论文集，2021年。[7] J. Chen、Q. Mao 和 D. Liu，“双路径Transformer网络：用于端到端单耳语音分离的直接上下文感知建模”，*Interspeech*会议论文集，2020年。[8] G. Wichern、J. Antognini、M. Flynn、L. R. Zhu、E. McQuinn、D. Crow、E. Manilow 和 J. L. Roux，“WHAM!：将语音分离拓展至噪声环境”，*Interspeech*会议论文集，2019年。[9] M. Maciejewski、G. Wichern、E. McQuinn 和 J. Le Roux，“WHAMR!：噪声和混响环境下的单通道语音分离”，*ICASSP*会议论文集，2020年。[10] J. Han 和 Y. Long，“用于无监督语音分离适配的异质分离一致性训练”，*arXiv*预印本 *arXiv:2204.11032*，2022年。[11] S. Wisdom、E. Tzinis、H. Erdogan、R. Weiss、K. Wilson 和 J. Hershey，“使用混合不变训练的无监督声音分离”，*NeurIPS*会议论文集，2020年。[12] A. Sivaraman、S. Wisdom、H. Erdogan 和 J. R. Hershey，“使用混合不变训练使语音分离适用于真实会议”，*ICASSP*会议论文集，2022年。[13] S. Yang、P. Chi、Y. Chuang、C. J. Lai、K. Lakhotia、Y. Y Lin、A. T Liu、J. Shi、X. Chang、G. Lin 等，“SUPERB：语音处理通用性能基准”，*Interspeech*会议论文集，2021年。[14] A. Gulati、J. Qin、C. Chiu、N. Parmar、Y. Zhang、J. Yu、W. Han、S. Wang、Z. Zhang、Y. Wu 等，“Conformer：卷积增强型Transformer语音识别模型”，*Interspeech*会议论文集，2020年。[15] S. Chen、C. Wang、Z. Chen、Y. Wu、S. Liu、Z. Chen、J. Li、N. Kanda、T. Yoshioka、X. Xiao 等，“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》，2022年。[16] Z. Chen、T. Yoshioka、L. Lu、T. Zhou、Z. Meng、Y. Luo、J. Wu、X. Xiao 和 J. Li，“连续语音分离：数据集与分析”，*ICASSP*会议论文集，2020年。[17] Z. Huang, S. Watanabe, S. Yang, P. Garc´ıa, 和 S. Khudan- pur，《探索用于语音增强与分离的自监督学习》，*ICASSP*论文集，2022年。[18] J. Cosentino、M. Pariente、S. Cornell、A. Deleforge 和 E. Vin- cent，《LibriMix：一种用于可泛化语音分离的开源数据集》，*arXiv*预印本 *arXiv:2005.11262*，2020年。[19] W. Kraaij、T. Hain、M. Lincoln 和 W. Post，《AMI会议语料库》，国际行为研究方法与技术会议论文集，2005年。[20] A. T Liu、S. Li 和 H. Lee，《TERA：面向语音的Transformer编码器表征自监督学习》，*IEEE*音频、语音与语言处理汇刊，第29卷，第2351–2366页，2021年。[21] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，《Wav2Vec 2.0：语音表征自监督学习框架》，*NeurIPS*论文集，2020年。[22] W. Hsu、B. Bolte、Y. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 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*音频、语音与语言处理汇刊，第29卷，第3451–3460页，2021年。[23] S. Chen、Y. Wu、C. Wang、Z. Chen、Z. Chen、S. Liu、J. Wu、Y. Qian、F. Wei、J. Li 等，《UniSpeech-SAT：结合说话人感知预训练的通用语音表征学习》，*ICASSP*论文集，2022年。[24] H. Erdogan、J. R Hershey、S. Watanabe 和 J. Le Roux，《使用深度循环神经网络进行相位敏感与识别增强的语音分离》，*ICASSP*论文集，2015年。[25] M. Delcroix、K. Zmolikova、K. Kinoshita、A. Ogawa 和 T. Nakatani，《基于说话人波束的单通道目标说话人提取与识别》，*ICASSP*论文集，2018年。[26] Jont B Allen 和 David A Berkley，《用于高效模拟小房间声学的镜像法》，美国声学学会杂志，第65卷，1979年，第950、943页。[27] E. AP Habets 和 S. Gannot，《在各向同性噪声场中生成传感器信号》，美国声学学会杂志，第122卷，第6期，第3464–3470页，2007年。[28] S. Chen、Y. Wu、Z. Chen、J. Wu、J. Li、T. Yoshioka、C. Wang、S. Liu 和 M. Zhou，《使用Conformer进行连续语音分离》，*ICASSP*论文集，2021年。[29] X. Zheng、C. Zhang 和 P. C Woodland，《说话人日志化与语音识别的串联多任务训练，用于会议转录》，*Interspeech*论文集，2022年。[30] N. Dawalatabad、M. Ravanelli、F. Grondin、J. Thienpondt、B. Desplanques 和 H. Na，《用于说话人分离的 ECAPA-TDNN 嵌入》，*Interspeech*论文集，2021年。[31] W. Hsu、A. Sriram、A. Baevski、T. Likhomanenko、Q. Xu、V. Pratap、J. Kahn、A. Lee、R. Collobert、G. Synnaeve 等，《Robust Wav2Vec 2.0：分析自监督预训练中的域偏移》，*Interspeech*论文集，2021年。[32] J. J Godfrey、E. C Holliman 和 J. McDaniel，《Switchboard：用于研究与开发的电话语音语料库》，*ICASSP*论文集，1992年。