训练动态感知的多因素课程学习用于目标说话人提取

Yun Liu[1,2], *Xuechen Liu*[1],*Xiaoxiao Miao*[3],*Junichi Yamagishi*[1,2]

[1]日本国立情报学研究所，东京，日本[2]综合研究大学院大学，神奈川，日本
[3]自然与应用科学部，昆山杜克大学，昆山，中国{yunliu, xuecliu, jyamagis}@nii.ac.jp

arXiv:2603.04943v1 [cs.SD] 5 Mar 2026

摘要

目标说话人提取（TSE）旨在从多说话人混合语音中分离出特定说话人的声音。尽管在基准测试中取得了优异结果，但由于不同交互因素的影响，实际应用中的性能往往会下降。以往针对TSE 的课程学习方法通常分别处理这些因素，无法捕捉它们复杂的相互作用，并且依赖预先设定的难度因素，而这些因素可能与模型实际的学习行为不一致。为应对这一挑战，我们首先提出一种多因素课程学习策略，联合安排信噪比阈值、说话人数量、重叠比例以及合成/真实数据比例，从而实现从简单场景到复杂场景的渐进式学习。然而，在不预先设定假设的情况下确定最优安排仍然具有挑战性。因此，我们引入 TSE-Datamap，这是一种可视化框架，通过跟踪训练轮次中的置信度和变异性，将课程设计建立在观测到的训练动态之上。我们的分析揭示了三个具有代表性的数据区域：（i）易学习样本，模型能够持续取得良好表现；（ii）模糊样本，模型会在不同预测结果之间摇摆；（iii）难学习样本，模型始终难以处理。在这些数据驱动洞见的指导下，我们的方法相比随机采样提升了提取效果，尤其在具有挑战性的多说话人场景中表现出显著增益。

索引词— 目标说话人提取、课程学习、数据选择、训练动态。

1. 引言

目标说话人提取（TSE）旨在从包含其他说话人和噪声的混合语音中分离出目标语音 [1, 2]。传统的目标说话人提取训练方案在所有训练数据中采用均匀随机采样 [3, 4], ，无论训练样本的学习难度如何，都将其一视同仁。然而，已有研究表明，不同样本对模型而言具有不同的学习难度，这推动了课程学习（CL） [5] 的采用，即根据难度因素，逐步引入从易到难的训练样本。广泛认可的难度因素包括信噪比（SNR） [6, 7], 、干扰说话人的数量 [8, 9],以及时间重叠 [10, 11]。随着合成数据越来越多地用于训练数据增强，近期研究还发现，干扰说话人的性质（实数还是合成）也是一种新出现的难度因素 [12, 13]。在现实场景中，这些因素并非彼此独立，其相互作用会带来无法通过单独分析各因素而预见的挑战。

以往用于目标说话人提取的持续学习方法通常应对这些因素分别 [14], 沿单一维度逐步增加难度。然而，此类单因素课程无法捕捉复杂的因素交互，并依赖于预定义的难度指标，而这些指标可能与模型在训练过程中实际感知任务难度的方式并不一致 [5, 15]。这种不匹配可能导致无效的课程调度，使得被预定义指标视为“简单”的样本，对模型而言实际上可能难以学习。

为解决这些局限，我们提出一种多因素课程学习策略，联合调度**信噪比阈值、说话人数量、重叠比例以及合成/真实数据比例**，使模型能够逐步从简单场景学习到复杂场景。这里的关键挑战在于，如何在不依赖对任务难度的预设假设的情况下，确定这些多因素的最优调度方式。我们不采用预先设定的方案，而是基于观测到的训练动态扩展课程设计，并提出***TSE-Datamap***，这是一种数据选择与可视化框架，可根据模型随时间实际学习训练样本的情况对其进行映射 [16]。通过跟踪每个样本在各训练轮次中的基本统计数据，我们构建了一个二维表征空间，其中包含若干不同区域，分别表示模型面临的不同学习难度。我们的分析表明，先让模型接触易学习样本，可使优化过程更加高效；这类样本能提供清晰的分离线索，并在处理更复杂的案例之前建立可靠的决策边界。该方法确保课程调度与模型实际的学习行为相一致，而非依赖对任务难度的预设假设。

本文有两项主要贡献：1）我们提出了一种用于目标说话人提取的多因素课程学习策略，可联合安排多个复杂度因素，使模型能够从简单场景逐步学习到复杂场景；2）我们提出了TSE-Datamap，使所提出的课程学习设计以观测到的训练动态为依据，而非预先定义的难度指标。实验表明，我们的方法始终优于单因素课程，在复杂的多说话人场景中取得了显著的性能提升。

2. 方法

2.1. 目标说话人提取

目标说话人提取旨在从包含多个同时说话人和背景噪声的混合语音中恢复目标说话人的语音。给定单通道混合语音：

$$y(t) = s_{\text{tar}}(t) + \sum_{k=1}^{K} u_k(t) + n(t) \tag{1}$$

其中 $s_{\text{tar}}(t)$ 是目标说话人语音，$u_k(t)$ 是 K 个说话人中的第 k 个干扰说话人，$n(t)$ 表示背景噪声。目标说话人提取模型 $f_\theta(\cdot)$ 以混合语音 y 和来自目标说话人的参考话语 $s_{\text{tar}} = f_\theta\ y\ c$ 作为输入，生成估计 $\hat{}(,)$。该模型通过训练最小化估计结果与真实目标语音之间的回归损失。

2.2. 目标说话人提取中的多因素复杂性

如前所述，在多说话人混合中执行目标说话人提取的难度取决于多个以复杂方式相互作用的因素。本研究聚焦于四个关键因素，并探讨它们对语音提取性能的影响。下面将详细介绍这些因素。

信噪比（SNR）。根据式 1 中的混合信号公式，信噪比定义为 $\text{SNR} = 10\log_{10}(\|s_{\text{tar}}(t)\|^2)/(\|\sum_{k=1}^{K} u_k(t) + n(t)\|^2)$，其中较低的信噪比表示相对于目标信号而言干扰更强。当信噪比接近或低于 0 分贝时，干扰功率会超过目标语音，产生严重的掩蔽效应，显著削弱模型从背景干扰中辨别目标语音特征的能力。

干扰说话人的数量。传统的目标说话人提取仅涉及一个干扰说话人 [2, 3], ，模型需要在混合信号中的两位说话人之间进行区分。存在多个干扰说话人时，模型必须从重叠语音模式所产生的复杂频谱相互作用中分离出目标语音。随着式 1 中同时说话人数量K 的增加，声学混淆会以非线性方式加剧，因为每增加一位说话人，都会在不同频带 [8, 9]引入额外的频谱重叠和潜在的掩蔽效应。

时间重叠率。 时间重叠率用于量化目标说话人与干扰说话人同时发出语音的时间占比，这会对底层声学线索和说话人线索产生难以预测的影响。低重叠场景中有大量清晰的目标语音，可作为明确的声学参照。另一方面，高重叠比例可能会使任务转变为并发声源分离，且说话人切换并不频繁；现代目标说话人提取模型能够相对稳定地处理这种情况。中等重叠率最具挑战性，因为目标说话人与干扰说话人之间的快速交替要求持续跟踪说话人，并稳健地处理频繁的声学变化 [17, 9]。早期研究也已证实，现代目标说话人提取系统同样面临这一情况 [18, 11]。在本研究中，重叠率的范围为 0 到 1，其中 0 表示完全重叠，1 表示混合语音中的说话人之间没有重叠。

合成与真实干扰说话人。 使用合成干扰说话人会引入真实录音中可能不存在的变化。合成语音数据可能分布在声学空间中较为特殊的区域，从而带来不同于真实语音干扰的挑战。真实干扰说话人能够提供真实的声学模式，而合成数据则有助于模型学会处理边缘情况，并避免对特定说话人特征过拟合 [19, 12]。

2.3. TSE-Datamap

本节介绍所提出的数据选择与可视化框架 TSE-Datamap。与传统持续学习不同，传统方法采用人工制定的规则，并忽略训练过程中模型的反馈；而在所提出的方法中，训练样本根据模型随时间的学习情况进行安排。通过追踪模型行为，TSE-Datamap 揭示了三个区域（简单、模糊、困难），可直接指导我们的采样计划。

2.3.1. 基本术语

对于每个训练样本 i，我们使用训练损失作为度量，记为 M_i。随后，我们在 E 个训练轮次内计算两个关键统计量，以分析整体训练动态：**置信度**$\mu_i = \frac{1}{E}\sum_{e=1}^{E} M_i^{(e)}$ 表示跨轮次的平均度量，其中 $M_i^{(e)}$ 是样本 i在第 e轮的度量。**变异性** $\sigma_i = \sqrt{\frac{1}{E}\sum_{e=1}^{E}(M_i^{(e)} - \mu_i)^2}$ 通过标准差（std）反映预测结果的一致性。

我们采用基于信噪比的损失，直接最大化信号质量。在每个训练轮次中，$M_i(s_{\text{tar},\,i}, \hat{}s_{\text{tar},\,i}) = -10\log_{10}\frac{\|s_{\text{tar},\,i}\|^2}{\|s_{\text{tar},\,i} - \hat{}s_{\text{tar},\,i}\|^2}$，其中$s_{\text{tar},\,i}$ 和 $\hat{}s_{\text{tar},\,i}$ 分别表示每个训练样本 i 的目标语音和预测语音。置信度和变异性构成一个二维 Datamap，每个训练样本根据其训练动态占据一个特定位置。这揭示了三种不同的学习行为：

- **易学习**：这类样本**置信度高**且**变异性低**，通常包含清晰的高信噪比语音，干扰极少。
- **模糊**：这类样本**变异性高**，模型会在不同假设之间摇摆，因此预测结果相当不稳定。这些样本包含丰富的判别信息，促使模型学习鲁棒边界并提升泛化能力。在目标说话人提取中，这种情况会出现在中度重叠或声学特征相似的说话人场景中。
- **难学习**：这类样本**置信度低**且**变异性低**，表明预测结果相当不精确。这种情况通常由极其不利的条件造成，例如信噪比极低、判别线索极少。

2.3.2. 构建与可视化

训练期间，我们通过对每个训练样本的四个因素进行均匀采样来生成数据：信噪比取自 $\{0, 5, 10, 15\}$ 分贝，重叠比例取自 $\{0, 0.2, 0.4\}$（其中 0 表示完全重叠），干扰说话人数量取自 $\{1, 2, 3\}$，干扰类型取自{真实，合成、真实/合成}；其中，“真实/合成”表示对于多说话人混合语音，每个干扰说话人都会随机选自真实语音池或合成语音池。

为了构建图1所示的 TSE-Datamap 可视化结果，我们跟踪每个样本在 $E = 50$ 个训练轮次中的训练动态。具体而言，对于每个训练样本 i，我们记录每个训练轮次 e的信噪比损失改善量（$\Delta\mathcal{L}_{\text{SNR}}$），其计算方式为输出信噪比与输入信噪比之差。随后，我们根据第 2.3 节中的定义，计算不同训练轮次间的置信度，并舍弃第一个训练轮次，以减轻初始训练不稳定性。在本研究中，为简便起见，我们使用 $\Delta\mathcal{L}_{\text{SNR}}$ ，因为我们的训练目标基于信噪比。需要指出的是，这一过程本质上与指标无关，任何能够反映提取质量的度量（例如信号失真比（SDR）、短时客观可懂度、感知语音质量评估、DNSMOS [20]）都可以替换使用。

随后，将每个训练样本在各训练轮次中的 $\Delta\mathcal{L}_{\text{SNR}}$ 绘制为该二维空间中的一个点，x轴表示变异性，y轴表示置信度。根据所得分布，我们将样本划分为三个区域：首先选取变异性最高的30%样本作为**模糊**，然后从剩余的70%中，将置信度最高的前50%归类为**易学习**，将最低的20%归类为**难学习**。这

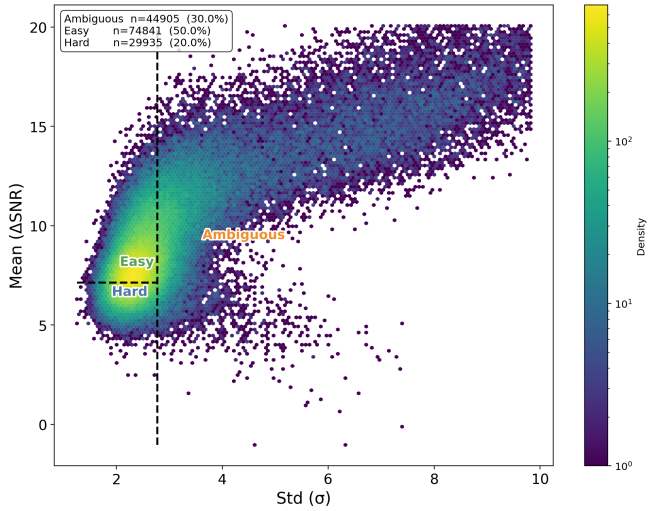

图1。计算得到的 TSE-Datamap。每个点表示一个样本，其中 x轴表示 Δ信噪比的标准差，y轴表示 Δ信噪比的均值。本研究的训练轮次为 50。

数据驱动分类与四个复杂度因素的跟踪元数据相结合，使我们能够分析不同因素组合在该图上的分布情况，并为我们的自适应课程学习策略提供依据，使其响应实际的模型行为，而非预先设定的难度假设。

3. 数据集与训练配置

3.1. 数据集组成

我们采用 *Libri2Vox* 作为主数据集[13], ，该数据集由 LibriTTS [21] （目标语音）和 VoxCeleb2 [22] （干扰语音）混合而成。官方划分规模如下：训练集包含 1,151 名 LibriTTS 说话人和 5,900 名 VoxCeleb2 说话人，共有 149,681 条语音，对应 250 小时的混合数据；验证集包含 40 名 LibriTTS 说话人和 94 名 VoxCeleb2 说话人，共有 4664 条语音，对应 7.77 小时；测试集包含 39 名 LibriTTS 说话人和 118 名 VoxCeleb2 说话人，共有 3931 条语音，对应 6.55 小时。原始混音规则将信噪比从 $[-5, 5]$ 分贝范围内均匀采样。该数据集还提供一个合成版本（*Libri2Vox-syn*），由 SALT [23] 和 SynVox2 [24]等语音生成模型制作。为进行评估，我们通过混合 LibriTTS 和 VoxCeleb2 的测试划分，构建包含 1、2 和 3 名干扰说话人的测试集。对于只有一名干扰说话人的混合语音，信噪比从 $[-5, 5]$ 分贝范围内均匀采样，遵循原始数据集 [13] 的设置。对于包含 2 名或 3 名干扰说话人的混合语音，信噪比范围调整为 $[0, 10]$ 分贝，以避免多个同时说话人可能造成的能量溢出和削波伪影。所有测试混合语音的重叠率均为 0.0，且仅包含实数音频，不含合成干扰说话人。

3.2. 目标说话人提取模型架构

我们采用一个两层、256 维的基于BLSTM的网络，后接一个全连接网络，用于处理频谱特征和说话人嵌入。该网络以 80 维对数梅尔滤波器组特征作为输入，说话人信息由预训练的 ECAPA-TDNN [25] 提供，并在训练期间保持冻结。BLSTM 以双向方式处理时间依赖关系，生成用于分离目标语音的时频掩码。模型详情见 [13]。此外，课程学习策略已展现出跨架构的可迁移性，这表明一旦在某个特定模型上取得成效，便可轻松应用于其他分离网络 [14]。

3.3. 训练设置

优化采用 Adam 优化器，初始学习率为 2e-4，并结合热重启学习率调度。预热阶段持续 5,000 步，学习率呈指数增长至 1e-3，随后按照步数的平方根倒数进行 Noam 式衰减，最终降至 1e-5 [26]。该调度既能确保初期训练稳定，又能在后续阶段保持适应性。每个阶段均通过监测验证损失并设置 6 个训练轮次的耐心值来执行早停，从而防止过拟合，同时确保模型充分收敛。我们以信号失真比作为主要指标评估性能，并额外报告相对于输入混合信号的改进信号失真比（iSDR）；这一指标已被以往研究广泛采用 [2, 3]。

4. 结果

4.1. 单因素与多因素分析

为了解不同复杂度因素的相对贡献，并验证我们提出的多因素方法的有效性，我们对单因素课程与提出的多因素策略进行了系统比较。

基线采用均匀随机采样，不进行课程学习。训练样本的信噪比从 $[0, 10]$ 分贝中均匀采样，干扰说话人数从 $\{1, 2, 3\}$ 中采样，重叠率从 $\{0, 0.2, 0.4\}$ 中采样，干扰类型则从 {真实、合成、真实/合成}中采样。对于 2 人、3 人和 4 人混合语音，其 iSDR 分别为 12.38 分贝、8.56 分贝和 7.16 分贝。基线课程学习采用基本的三阶段课程，参数固定：各阶段的说话人数逐步从 1 人增加到 3 人，同时其他因素保持不变，使 iSDR 略有提升，分别达到 12.58 分贝、8.78 分贝和 7.40 分贝。

以基线课程学习为参照，我们通过仅改变一个复杂度因素、同时保持其他因素不变，系统地评估单因素课程。在单因素课程策略中，基于信噪比的课程表现突出，通过在各阶段逐步调整信噪比阈值，取得了 13.04 dB、9.71 dB 和 8.62 dB 的成绩。单因素说话人数量实验将所有阶段固定为仅使用 1 个干扰说话人；与默认的课程学习逐步将每个阶段的说话人数量从 1 个增加到 3 个不同，该实验取得了 12.80 dB、7.79 dB 和 6.73 dB 的成绩。仅使用单个干扰说话人进行训练，在双说话人混合场景中的表现优于基线课程学习，但在多说话人场景中性能显著下降。这表明，在训练中纳入多说话人案例需要权衡：双说话人场景的性能略有损失，但更复杂场景的表现得到提升。多因素策略取得了最佳的整体表现，分别为 13.22 dB、10.08 dB 和 9.21 dB。与基线相比，在 4 说话人场景中，相对提升最高达 24.5%；随着干扰

表 1。单因素和多因素持续学习的影响。每行代表一种不同的课程配置，包含不同的因子。课程阶段列展示了训练难度在三个阶段中的递进过程，参数格式为 [说话人数，信噪比下限，信噪比上限，重叠率，干扰源]。说话人数：干扰说话人的数量，[信噪比下限、信噪比上限]：以分贝为单位的信噪比范围，重叠率：时间重叠的比例，范围为 0.0 至 0.4，干扰源：干扰来自真实说话人还是合成（syn）说话人。

Exp.	课程阶段	iSDR 2说话人	iSDR 3说话人	iSDR 4说话人
基线	随机采样	12.38	8.56	7.16
基线课程学习	[[1,0,10,0,实数], [2,0,10,0,实数], [3,0,10,0,实数]]	12.58	8.78	7.40
单因素: #snr	[[1,5,10,0,实数], [2,0,10,0,实数], [3,0,5,0,实数]]	13.04	9.71	8.62
单因素: #重叠率	[[1,0,10,0,实数], [2,0,10,0.2,实数], [3,0,10,0.4,实数]]	12.62	9.87	8.76
单因素: #干扰源合成	[[1,0,10,0,合成], [2,0,10,0,合成], [3,0,10,0,合成]]	11.22	8.43	9.19
单因素: #干扰源 真实/合成	[[1,0,10,0,真实/合成], [2,0,10,0,真实/合成], [3,0,10,0,真实/合成]]	12.73	9.51	8.96
单因素: #说话人数量	[[1,0,10,0,实数], [1,0,10,0,实数], [1,0,10,0,实数]]	12.80	7.79	6.73
多因素	[[1,5,10,0,真实/合成], [2,0,10,0.2,真实/合成], [3,0,5,0.4,真实/合成]]	**13.22**	**10.08**	**9.21**

说话人数量增加，提升幅度也随之增大，验证了协调的多因素渐进策略比单因素优化能够实现更有效的学习。

4.2. TSE-Datamap分析

在此，我们展示基于TSE-Datamap的结果与发现。为确定持续学习中难度区域的最优排序，我们对简单（E）、模糊（A）和困难（H）样本的所有可能排列进行了实验。每种排列代表一种不同策略，训练以等长的三个阶段推进。各区域的生成遵循第2.4节所述方法。

多种TSE-Datamap变体的结果见表2，并与基线以及在表1中取得最佳性能的多因素系统一并给出。简单／模糊／困难日程在所有日程中表现最佳，相对基线分别取得0.77、1.29和2.16分贝的绝对增益，并比精心设计的多因素方案高出0.11分贝。收益随干扰说话人增多而增大，表明TSE-Datamap在应对日益增加的复杂度方面具有潜力。在先易策略中，简单／模糊／困难优于简单／困难／模糊，说明在困难样本之前引入模糊数据能更好地校准决策边界。相比之下，将困难样本提前（模糊／简单／困难、模糊／困难／简单）或以其开头（困难／简单／模糊、困难／模糊／简单）的顺序，通常因早期优化不稳定而表现较差。总体而言，简单／模糊／困难提供了最有效的递进路径，尤其是在具有挑战性的多说话人场景中。

我们还进行了 E / A / H（遗忘）实验，每个阶段使用的数据与 E / A / H 相同，但每个阶段仅使用其目标区域数据，不保留前一阶段的数据。这种设置显示出显著的性能下降，证实了

表 2。课程学习中不同阶段顺序的性能比较。每个阶段占总训练轮次的三分之一。基线采用随机采样，不使用课程学习。

顺序	iSDR 2说话人	iSDR 3说话人	iSDR 4说话人
基线	12.38	8.56	7.16
多因素	**13.22**	**10.08**	9.21
E / A / H	13.15	9.85	**9.32**
E / H / A	12.93	9.63	9.18
A / E / H	12.82	9.61	9.22
A / H / E	12.96	9.72	9.17
H / E / A	12.77	9.54	**9.32**
H / A / E	12.90	9.63	9.10
E / A / H（遗忘）	8.83	5.43	5.52

表3。从 Datamap 区域中抽取固定数量的数据时的 iSDR（分贝）。所有设置均使用训练集的 70% 进行等量比较。

设置	2spk	3spk	4spk
ambi70%	**11.67**	**8.67**	**8.61**
baseline70%	11.37	8.19	7.17
hard70%	10.10	6.75	6.99
easy70%	9.71	6.50	6.40

在课程推进过程中，若不保留先前学到的知识，就会发生灾难性遗忘。

4.3. Datamap区域的固定数量消融 ns

为了将不同TSE-Datamap区域的学习贡献与数据数量效应分离开来，我们在固定数据数量下进行了受控消融实验。我们将所有方法约束为恰好使用全部训练数据的70%。all70%基线从整个训练分布中均匀采样，而easy70%、ambi70%和hard70%则将各自70%的数据分配给相应的目标区域。在所有情况下，剩余30%的数据均匀分配到另外两个区域。

如表3所示，在相同的数据数量约束下，ambi70%在所有混合复杂度上均持续优于all70%基线。easy70%设置在四种方法中表现最差，这可能是因为这些样本被迅速处理，随着训练的推进，所提供的信息性梯度逐渐减弱。相比之下，模糊样本因其固有的学习难度，在整个训练过程中保持稳定的信息量，促使模型形成更稳健的决策边界，从而在同等数据数量下增强泛化能力。

5. 结论

在本研究中，我们提出了 TSE-Datamap，这是一个通过置信度和变异性指标识别易、模糊和难区域，从而分析目标说话人提取训练动态的框架。我们的多因素课程学习策略联合安排信噪比、说话人数量、重叠比例以及合成/真实数据比例，在多说话人场景中实现了最高 24.5% 的相对 iSDR 增益。在不同的排序方案中，易-模糊-难排序效果最佳，这验证了模型在处理复杂情况之前，先建立可靠决策边界会从中受益。

6. 致谢

本研究部分获得了日本文部科学省科学研究费助成金（24K21324）以及日本科学技术振兴机构“为创造科学技术创新而设立大学奖学金”项目（JPMJFS2136）的支持。

7. 参考文献

[1] Katerina Zmolikova、Marc Delcroix、Tsubasa Ochiai、K. Ki- noshita、Jan Honza Cernocky 和 Dong Yu，《神经网络目标语音提取：综述》，*IEEE*《信号处理杂志》，第 40 卷，第 8–29 页，2023 年。[2] Katerina Zmolikova、Marc Delcroix、K. Kinoshita、Tsub- asa Ochiai、T. Nakatani、L. Burget 和 Jan Honza Cer- nocky，《Speakerbeam：用于语音混合中目标说话人提取的说话人感知神经网络》，*IEEE*《信号处理领域精选主题期刊》，第 13 卷，第 800–814 页，2019 年。[3]Chenglin Xu、Wei Rao、E. Chng 和 Haizhou Li，《Spex：多尺度时域说话人提取网络》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，第 1370–1384 页，2020 年。[4] Ruijie Tao、Xinyuan Qian、Yidi Jiang、Junjie Li、Jiadong Wang 和 Haizhou Li，《基于选择性听觉注意的视听目标说话人提取》，*IEEE*《音频、语音与语言处理汇刊》，第 33 卷，第 797–811 页，2025 年。[5] Xin Wang、Yudong Chen 和 Wenwu Zhu，《课程学习综述》，*IEEE*《模式分析与机器智能汇刊》，第 44 卷，第 4555–4576 页，2021 年。[6] Tian Gao、Jun Du、Lirong Dai 和 Chin-Hui Lee，《基于信噪比的深度神经网络语音增强渐进式学习》，*Interspeech2016*，第 3713–3717 页，2016 年。[7] Yanhui Tu、Jun Du、Tian Gao 和 Chin-Hui Lee，《一种用于基于回归的语音增强的多目标信噪比渐进式学习方法》，*IEEE/ACM* 音频、语音与语言处理汇刊，第 28 卷，1619，2020 年 1608 页。[8] Jiahe Wang、Shuai Wang、Junjie Li、Ke Zhang、Yan min Qian 和 Haizhou Li，《通过纠正目标混淆增强说话人提取》，*2024 IEEE* 口语技术研讨会（*SLT*），第 349–356 页，2024 年。[9] Sanyuan Chen、Yu Wu、Zhuo Chen、Jinyu Li、Chengyi Wang、Shujie Liu 和 M. Zhou，《基于 Conformer 的连续语音分离》，*ICASSP*，第 5749–5753 页，2020 年。[10] Qingjian Lin、Lin Yang、Xuyang Wang、Luyuan Xie、Chen Jia 和 Junjie Wang，《时域稀疏重叠语音训练：目标语音分离与个性化语音活动检测的联合学习带来益处》，*2021* 亚太信号与信息处理协会年度峰会暨会议（*APSIPA ASC*），第 689–693 页，2021 年。

[11] Zexu Pan、Meng Ge 和 Haizhou Li，“Usev：利用视觉线索进行通用说话人提取”，*IEEE/ACM* 音频、语音与语言处理汇刊，第 30 卷，第 3032–3045 页，2021 年。

[12] Yun Liu、Xuechen Liu 和 Junichi Yamagishi，“利用合成说话人改进目标说话人提取的课程学习”，*2024 IEEE* 口语语言技术研讨会 *(SLT)*，第 364–370 页，2024年。

[13] Yun Liu、Xuechen Liu、Xiaoxiao Miao 和 Junichi Yamagishi，《Libri2vox 数据集：具有多样说话人条件和合成数据的目标说话人提取》，*ArXiv*，第 abs/2412.12512 卷，2024 年。[14] Yun Liu、Xuechen Liu、Xiaoxiao Miao 和 Junichi Yamagishi，《采用课程学习的目标说话人提取》，*ArXiv*，第 abs/2406.07845 卷，2024 年。[15] Mengya Xu、Mobarakol Islam、Ben Glocker 和 Hongliang Ren，《通过标签平滑实现置信度感知的节奏式课程学习，以进行手术场景理解》，*IEEE Trans- actions on Automation Science and Engineering*，第 21 卷，第 3170–3181 页，2022 年。[16] Swabha Swayamdipta、Roy Schwartz、Nicholas Lourie、Yizhong Wang、Hannaneh Hajishirzi、Noah A. Smith 和 Yejin Choi，《数据集制图：利用训练动态对数据集进行映射与诊断》，*EMNLP*，第 9275–9293 页，2020 年。[17] H. Taherian 和 Deliang Wang，《基于重叠率的时域损失调制，用于单声道对话说话人分离》，*ICASSP*，第 5744–5748 页，2021 年。[18] Jun Yu Li、Ruijie Tao、Zexu Pan、Meng Ge、Shuai Wang 和 Haizhou Li，《针对稀疏重叠的多说话人语音的视听主动说话人提取》，*ICASSP*，第 10666–10670 页，2023 年。[19] Wupeng Wang、Zexu Pan、Xinke Li、Shuai Wang 和 Haizhou Li，《采用预训练前端的语音分离，以最小化领域失配》，*IEEE/ACM Transactionson Audio, Speech, and Language Processing*，第 32 卷，第 4198 页，2024 年。[20] Chandan K. A. Reddy、Vishak Gopal 和 Ross Cutler，《DNSMOS：用于评估噪声抑制器的非侵入式感知语音质量客观指标》，*ICASSP*，第 6493–6497 页，2020 年。[21] Heiga Zen、Viet Dang、Rob Clark、Yu Zhang、Ron J Weiss、Ye Jia、Zhifeng Chen 和 Yonghui Wu，《LibriTTS：源自 LibriSpeech 的文本转语音语料库》，*Interspeech*，2019 年 4 月。[22] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，《VoxCeleb2：深度说话人识别》，*Interspeech*，第 1086–1090 页，2018 年。[23] Xiaoxiao Miao、Xin Wang、Erica Cooper、Junichi Yamagishi、Nicholas Evans、Massimiliano Todisco、Jean-François Bonastre 和 Mickael Rouvier，《SynVox2：迈向隐私友好的 VoxCeleb2 数据集》，发表于 *ICASSP*，2024 年，第 11421–11425 页。[24] Yuanjun Lv、Jixun Yao、Peikun Chen、Hongbin Zhou、Heng Lu 和 Lei Xie，《SALT：通过潜在空间变换实现可区分说话人匿名化》，发表于*2023 IEEE Automatic SpeechRecognitionandUnderstanding(ASRU)*，2023 年。[25] Brecht Desplanques、Jenthe Thienpondt 和 Kris De muynck，《ECAPA-TDNN：基于 TDNN 的说话人验证中的增强型通道注意力、传播与聚合》，*ArXiv Inter- speech2020*，第 3830–3834 页，2020 年。[26] Ashish Vaswani、Noam M. Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser 和 I. Polosukhin，《Attention Is All You Need》，神经信息处理系统进展，第 5998–6008 页，2017 年。