arXiv:2303.03732v1 [cs.SD] 7 Mar 2023

# 一种适用于噪声和混响环境的多阶段三路径语音分离方法

*Zhaoxi Mu*⋆ *Xinyu Yang*⋆ *Xiangyuan Yang*⋆ *Wenjing Zhu*†

⋆ 中国西安西安交通大学计算机科学与技术学院

† 中国北京度小满

**摘要**

在噪声和混响环境中，基于深度学习的语音分离方法的性能会大幅下降，因为以往的方法并非针对这类情况设计和优化。为解决这一问题，我们提出一种多阶段端到端学习方法，将噪声和混响环境下复杂的语音分离问题分解为三个子问题：语音去噪、分离和去混响。通过缩小解空间，提高了搜索语音分离模型最优解的概率和速度。此外，由于时域音频序列的通道信息对语音分离至关重要，我们提出一种能够对音频序列通道维度进行建模的三路径结构。实验结果表明，所提出的多阶段三路径方法只需少量增加模型参数，即可提升语音分离模型的性能。

索引词— 语音分离、多阶段学习、三路径模型

## 1. 引言

语音分离是一种常见的语音预处理方法，可从混合语音中分离出各个声源的纯净语音。本文重点研究噪声和混响环境下的单通道语音分离问题。随着深度学习在语音分离中的应用与发展，其性能得到了显著提升。例如，Luo 和 Mesgaran [1]提出了一种时域语音分离方法，以避免独立预测相位。Luo 等人 [2] 提出了一种双路径方法，可高效处理较长的语音序列。在此基础上，Subakan 等人 [3] 和 Chen 等人 [4] 利用 Transformer 强大的序列建模能力取代循环神经网络（RNN）或与其结合，进一步提升性能。

然而，在真实声学环境中，采集到的语音通常不仅是多个说话人声音的混合，还包含各种噪声和混响。噪声会掩盖语音信号并破坏相位信息。混响会导致源信号的频谱涂抹

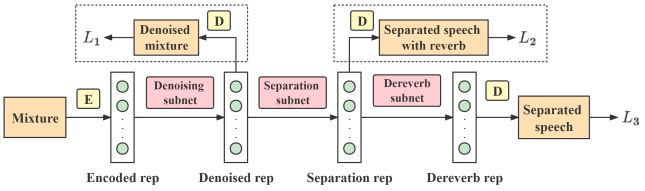



**图 1**：多阶段方法示意图。E、D 和 rep 分别是编码器、解码器和特征表示的缩写。虚线框内的模块仅在训练期间使用。

。面对噪声和混响时，大多数语音分离方法的性能都会急剧下降 [5]，因为它们并非针对这种情况进行设计和优化的。一种简单的解决方案是级联多个模型，以解决不同问题，并进行联合端到端优化[6, 7, 8, 9, 10]。然而，与单阶段方法相比，级联方法的训练过程较为繁琐，且内存消耗较大。此外，当使用尺度不变损失函数（如 SI-SNR） [11] 训练级联系统时，各网络的输出尺度可能不匹配，从而导致性能严重下降 [5]。为解决这些问题，受课程学习理念 [12, 13], 的启发，我们设计了一种单模型多阶段训练方法。该方法不仅能够训练端到端网络以简化训练过程，而且在模型规模相同的情况下，性能优于单阶段方法。

此外，广泛使用的时域双路径方法 [2, 3, 4] 通过将信号划分为较小的分块，并分别处理块内序列和块间序列，来考虑长序列的空间信息。然而，它没有考虑序列的通道（也称为特征图）信息，而这对语音分离至关重要。原因在于，信号的低频和高频信息对语音分离的重要性不同，分别对应时域特征表示中的不同通道 [1, 14]。低频带通常包含较高的能量、音调特征和持续时间较长的声音，而高频带通常包含较低的能量、噪声和快速衰减声音 [15]。为

弥补这一不足，我们使用通道注意力 [16, 17] 对特征表示的通道维度重新加权，以保留有价值的通道并忽略无价值的通道。

本文的主要贡献如下：(i) 我们提出了一种端到端多阶段语音分离方法，以更好地分离噪声和混响环境中的混合信号。(ii) 我们提出了一种三路径结构，可高效捕获有助于时域语音分离的通道信息。(iii) 结果表明，所提出的训练方法和模型结构能够在模型规模几乎不变的情况下提升语音分离性能。

## 2. 问题描述

在噪声和混响环境中，单声道混合信号可描述如下：

$$y(t) = \sum_i s_i(t) * h_i(t) + n(t) \tag{1}$$

其中，$t$ 表示长度索引，$s_i(t)$、$h_i(t)$和 $n(t)$分别表示无回声语音信号、第$i$位说话人的房间冲激响应（RIR）以及背景噪声。$*$表示卷积算子。目标是从 $y$中提取 $s_i$。

## 3. 多阶段方法

目标估计概率 $p(s, m, z \mid y)$可利用概率链式法则进行分解，

$$p(s, m, z \mid y) = p(z \mid y) \cdot p(m \mid z, y) \cdot p(s \mid m, z, y) \tag{2}$$

其中，$y$ 表示混合信号，$z$、$m$和 $s$ 分别表示去噪混合信号、带混响的分离语音和分离语音。原本复杂的问题被解耦为三个子问题：语音去噪、分离和去混响。我们称这种方法为多阶段学习，也可以将其视为多任务学习。其优点在于，通过缩小解空间，可以提高搜索最优解的概率，并加快模型训练的收敛速度。

多阶段方法的整体框架如图1所示。第一阶段中，去噪子网络将混合信号的编码特征表示转换为去噪后的特征表示。训练期间，解码器将其转换为去噪混合信号$\hat{z}$。以真实去噪混合信号 $z$ 为目标，计算损失 $L_1$，从而训练去噪子网络。与第一阶段类似，第二阶段中，分离子网络将去噪后的特征表示作为先验信息，用于预测每个说话人的分离特征表示。解码器将其转换为带混响的分离语音

$\hat{m}$在训练期间，通过计算损失 $L_2$来训练去噪子网络和分离子网络，其中以带混响的真实分离语音 $m$ 作为目标。在第三阶段，

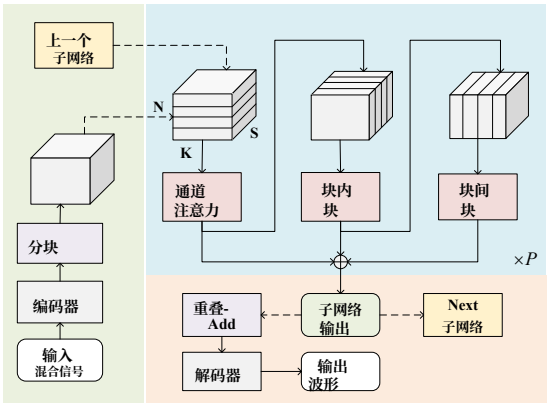



**图2**：模型结构示意图。右上方的三路径结构以蓝色显示。虚线箭头表示可能的输入或输出。

去混响子网络预测每个说话人的去混响特征表示。解码器将其转换为分离语音 $\hat{s}$。整个模型通过计算损失 $L_3$ 进行训练，并以真实分离语音$s$作为目标。总损失 $L_{\text{total}}$ 为：

$$L_{\text{total}} = \alpha L_1 + \beta L_2 + L_3 = \alpha \mathrm{S}(\hat{z}, z) + \beta \mathrm{S}(\hat{m}, m) + \mathrm{S}(\hat{s}, s) \tag{3}$$

其中，损失函数 S 为尺度不变信噪比（SI-SNR）。$\alpha$ 和 $\beta$ 用于平衡这三个损失函数的权重。

我们比较了两种策略：先去噪、分离、去混响，以及先去噪、去混响、再分离，发现第一种策略效果更好。原因可能是，如果先进行去混响再分离，由于多个声源具有不同的房间脉冲响应，去混响子网络必须充分补偿多种混响模式，这使问题变得更加复杂。我们不考虑分离后再去噪，因为目前尚不清楚网络应如何分配噪声，同时又不将其移除 [5]。

## 4. 三路径结构

这三个子问题在从混合信号中提取目标信号方面具有相似的内在特性，因此所有子网络都采用相同的结构。我们提出了一种三路径结构，可对时域音频序列的通道信息进行建模。整体模型结构如图2所示。卷积编码器对混合信号进行编码，得到二维编码特征表示；该表示沿空间维度划分为多个分块，重叠因子为50%，并作为第一个子网络的输入。

通道注意力 [16] 首先用于在三路径子网络中特征表示的通道维度上自适应地预测关键信息。假设 $X$ 是通道注意力的输入，其权重 $\phi(X)$计算如下

$$\phi(X) = \sigma(W_1 f(W_0(X_{\text{avg}})) + W_1 f(W_0(X_{\max}))) \tag{4}$$

我们对 $X$ 进行平均池化和最大池化，得到 $X_{\text{avg}}$和 $X_{\text{max}}$。随后将它们输入两个共享线性层，其权重分别为 $W_0$ 和 $W_1$。 $f$ 是 ReLU。 $\sigma$ 是 Sigmoid 函数。通道注意力的输出为 $\phi(X) \otimes X$，其中 $\otimes$ 表示逐元素乘法。池化操作可以提取通道维度上的整体信息，并屏蔽空间维度上的信息，从而有效利用通道间的相关性 [18]。

接下来，块内和块间模块分别从全局和局部角度对序列的空间维度进行建模，与通道注意力模块形成互补。此过程重复 $P$ 次。三条路径的结果相加后作为子网络的输出，并输入下一个子网络。同时，通过重叠相加法将其还原为二维特征表示，再由转置卷积解码器转换为音频波形。所有子网络的解码器均采用参数共享。

所有子网络都预测特征表示而非掩码，因为预测得到的特征表示可作为下一个子网络的先验信息。此外，前几个阶段的处理可能导致纯净语音信息损失，并在预测的特征表示中引入伪影。然而，预测掩码方法只能处理加性噪声，并且由于编码器和解码器的线性特性，难以处理此类干扰 [19, 20]。

## 5. 实验

### 5.1. 实验设置

使用 WHAM! [21] 和 WHAMR! [5]数据集训练和评估模型。WHAM! 数据集由 WSJ0-2mix 数据集中的两个说话人混合信号 [22]，以及独特的真实环境噪声样本组成。WHAMR! 数据集是在 WHAM! 数据集基础上扩展混响而成，进一步为含噪语音数据添加人工混响。使用 SI-SNRi 和 SDRi 评估语音分离方法。

为验证三路径结构和多阶段方法的有效性，我们在原有的双路径模型 DPRNN [2], 、DPTNet [4], 和 Sep-former [3], 中加入通道注意力路径，并分别采用单阶段和多阶段方法进行训练。为保持模型参数数量不变，由于需要三个层数相同的子网络，多阶段方法中每个子网络的重复组数量 $P$ 设为原始模型的 $1/3$ 。在 WHAM! 数据集上进行实验时， $P$ 设为原始模型的 $1/2$ 。由于混合信号中没有混响，此时只需要去噪和分离子网络。

所有子网络均采用端到端方式联合训练。在训练开始时，$L_1$ 和 $L_2$ 的权重至关重要，以确保学到的去噪和分离特征表示能够真正达到预期效果。

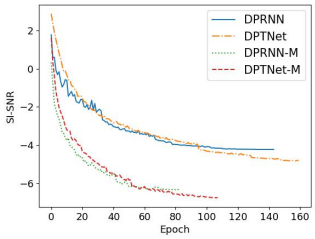



**图 3**：WHAMR! 数据集上单阶段和多阶段方法的学习曲线。纵坐标表示验证集上的尺度不变信噪比值。

随后逐渐增大 $L_3$ 的权重，以稳定训练过程。具体而言，$\alpha$ 和 $\beta$ 初始值设为 1，并且每训练 20 轮就减半。对于级联系统，首先分别训练去噪、分离和去混响网络，然后再采用端到端方式联合优化。每个网络的结构均与原始双路径模型相同。

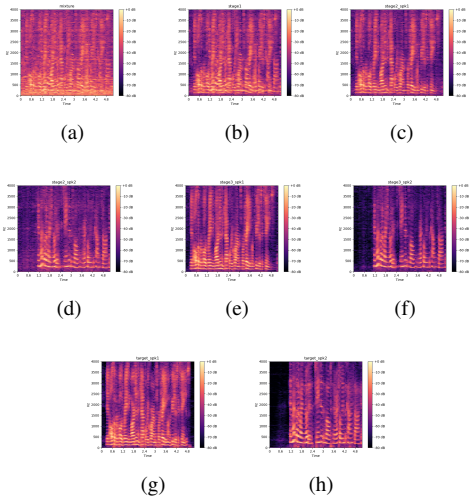



**图 4**：混合信号的频谱图（(a)）、各阶段输出波形的频谱图（(b) 至 (f)）以及目标的频谱图（(g) 和 (h)）。

### 5.2. 结果与讨论

WHAM! 和 WHAMR! 数据集上的结果如表 1 所示。我们的方法以粗体显示。-Ca、-T、-M 和 DM 分别表示级联模型、三路径模型、多阶段学习和动态混合 [24]。与原始单阶段方法 DPRNN、DPTNet 和 Sepformer 相比，多阶段方法 DPRNN-M、DPTNet-M 和 Sepformer-M 显著提升了分离效果，同时参数数量并未大幅增加，甚至优于参数量高出数倍的级联方法 DPRNN-Ca、

**Tab表 1**：WHAM! 和 WHAMR! 数据集上的模型规模、SI-SNR 和 SDR 提升（dB） t.

| 方法 | WHAM! # 参数 | WHAM! SI-SNRi | WHAM! SDRi | WHAMR! # 参数 | WHAMR! SI-SNRi | WHAMR! SDRi |
|---|---|---|---|---|---|---|
| DPRNN [2] | 2.6M | 13.9 | 14.3 | 2.6M | 10.6 | 10.0 |
| DPRNN-Ca | 5.2M | 14.5 | 14.8 | 7.8M | 11.7 | 10.8 |
| DPRNN-SRSSN [23] | 7.5M | 15.7 | 16.1 | 7.5M | 12.3 | 11.4 |
| **DPRNN-T** | **2.7M** | **14.4** | **14.8** | **2.7M** | **11.1** | **10.4** |
| **DPRNN-M** | **2.7M** | **15.6** | **16.0** | **2.7M** | **12.1** | **11.2** |
| **DPRNN-T-M** | **2.8M** | **16.2** | **16.6** | **2.8M** | **12.8** | **11.9** |
| DPTNet [4] | 2.7M | 14.9 | 15.3 | 2.7M | 11.2 | 10.6 |
| DPTNet-Ca | 5.4M | 15.5 | 15.8 | 8.1M | 11.9 | 11.0 |
| DPTNet-SRSSN [23] | 5.7M | 16.1 | 16.5 | 5.7M | 12.3 | 11.3 |
| **DPTNet-T** | **2.8M** | **15.3** | **15.7** | **2.8M** | **11.7** | **10.8** |
| **DPTNet-M** | **2.8M** | **16.0** | **16.3** | **2.8M** | **12.5** | **11.5** |
| **DPTNet-T-M** | **2.9M** | **16.6** | **16.9** | **2.9M** | **12.9** | **12.3** |
| Sepformer [3] | 25.7M | 15.5 | 15.8 | 25.7M | 13.7 | 12.7 |
| Sepformer +DM [3] | 25.7M | 16.4 | 16.7 | 25.7M | 14.0 | 13.0 |
| Sepformer-Ca | 51.4M | 16.1 | 16.5 | 77.1M | 14.3 | 13.5 |
| **Sepformer-T** | **25.8M** | **15.9** | **16.2** | **25.8M** | **14.1** | **13.1** |
| **Sepformer-M** | **25.8M** | **16.6** | **16.9** | **25.8M** | **14.7** | **14.0** |
| **Sepformer-T-M** | **25.9M** | **16.9** | **17.2** | **25.9M** | **15.2** | **14.4** |
| Wavesplit [24] | 29M | 15.4 | 15.8 | 29M | 12.0 | 11.1 |
| Wavesplit + DM [24] | 29M | 16.0 | 16.5 | 29M | 13.2 | 12.2 |

DPTNet-Ca 和 Sepformer-Ca。结果表明，多项任务间共享信息比单独处理各项任务能带来更好的性能。

与另一种多阶段方法逐步细化语音分离网络（SRSSN）相比， [23], 由于我们的方法在每个阶段都有更具体的目标，因此其学习效率更高。结果表明，当参数量分别为 DPRNN-SRSSN 的 $1/3$ 和 DPTNet-SRSSN 的 $1/2$ 时，我们的方法 DPRNN-M 和 DPTNet-M 仍能取得相近的性能。

从表中还可以看出，三路径模型 DPRNN-T、DPTNet-T 和 Sepformer-T 的表现始终优于双路径模型 DPRNN 、 DPTNet 和 Sepformer，而模型参数量仅略有增加，这验证了增加通道注意力路径的有效性。此外，多阶段三路径模型 DPRNN-T-M、DPTNet-T-M 和 Sepformer-T-M 的表现优于所有基线模型。

我们还比较了单阶段方法 DPRNN、DPTNet 和多阶段方法 DPRNN-M、DPTNet-M 的训练收敛速度，如图 3 所示。多阶段方法的学习曲线斜率大于单阶段方法，因此收敛速度更快。这验证了多阶段方法能够通过缩小解空间，提高找到最优解的概率和速度。

为阐明多阶段方法在各阶段的状态，我们将每个阶段输出的频谱图可视化。图 4 (a) 至 (h) 展示了混合信号、各子网络的输出以及目标。在两个说话人混合的情况下，去噪子网络输出一个信号，而分离和去混响子网络输出两个信号。图 4 (a) 和 (b) 显示，经过去噪阶段后，混合信号中的背景噪声有所降低。图 4 (b)、(c) 和 (d) 显示，经过分离阶段后，混合信号中的两个声源彼此分离。图 4 (c)、(d)、(e) 和 (f) 显示，经过去混响阶段后，各声源的混响干扰被去除。结果验证了多阶段方法在每个阶段都实现了相应目标。与以往以黑箱方式运行的语音分离模型不同，所提出的框架为各个模块提供了可解释性。多阶段方法的另一项优势是，各子网络可以单独用于相应任务。例如，经过多阶段训练后，去噪子网络可以单独用于语音增强。

## 6. 结论

在本研究中，我们通过概率链式法则推导出一种新颖的多阶段方法，用于解决噪声和混响环境中的语音分离问题。此外，我们提出了一种三路径结构，可在时域中自适应地建模音频序列的通道信息。实验结果表明，所提出的训练方法和模型结构显著提升了语音分离模型在噪声和混响环境中的性能，同时未大幅增加模型规模。

## 7. 参考文献


[1] Yi Luo 和 Nima Mesgarani，《Conv-TasNet：超越理想时频幅度掩蔽的语音分离方法》，*IEEE ACM* 音频、语音与语言处理汇刊，第 27 卷第 8 期，第 1256–1266 页，2019年。

[2] Yi Luo、Zhuo Chen 和 Takuya 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”，载于*ICASSP*。2020年，第46–50页，IEEE。

[3]Cem Subakan、Mirco Ravanelli、Samuele Cornell、Mirko Bronzi 和 Jianyuan Zhong，“语音分离中只需注意力”，载于*ICASSP*。2021年，第21–25页，IEEE。[4] Jingjing Chen、Qirong Mao 和 Dong Liu，“双路径 Transformer 网络：用于端到端单通道语音分离的直接上下文感知建模”，载于*IN-TERSPEECH*。2020年，第2642–2646页，ISCA。[5] Matthew Maciejewski、Gordon Wichern、Emmett Mc- Quinn 和 Jonathan Le Roux，“WHAMR!：噪声和混响单通道语音分离”，载于*ICASSP*。2020年，第696–700页，IEEE。[6] Ke Tan 和 DeLiang Wang，“采用门控残差网络的噪声同频语音分离两阶段方法”，载于 *INTERSPEECH*。2018年，第3484–3488页，ISCA。[7] Yan Zhao、Zhong-Qiu Wang 和 DeLiang Wang，“用于噪声混响语音增强的两阶段深度学习”，*IEEEACM Trans. Audio SpeechLang.Process.*，第27卷，第1期，第53–62页，2019年。[8]Masood Delfarah、Yuzhou Liu 和 DeLiang Wang，“混响环境下说话人无关的说话人分离”，载于*ICASSP*。2020年，第8723–8727页，IEEE。[9]Yuzhou Liu、Masood Delfarah 和 DeLiang Wang，“用于说话人无关的单声道语音分离的 Deep Casa”，载于 *ICASSP*。2020年，第6354–6358页，IEEE。

[10] Ke Tan、Yong Xu、Shi-Xiong Zhang、Meng Yu 和 Dong Yu，“采用两阶段多模态网络的视听语音分离与去混响”，*IEEE* 信号处理专题精选期刊，第 14 卷，第 3 期，第 542–553 页，2020年。

[11] Jonathan Le Roux、Scott Wisdom、Hakan Erdogan 和 John R. Hershey，“SDR——半生不熟还是恰到好处？”，载于 *ICASSP*。2019年，第 626–630 页，IEEE。

[12] Yoshua Bengio、J´erˆome Louradour、Ronan Collobert 和 Jason Weston，“课程学习”，载于*ICML*。2009年，第382卷《*ACM*国际会议论文集》，第41–48页，ACM。

[13] Andong Li、Wenzhe Liu、Xiaoxue Luo、Guochen Yu、Chengshi Zheng 和 Xiaodong Li，“一种采用目标解耦的同时去噪与去混响框架”，载于*INTERSPEECH*。2021年，第2801–2805页，ISCA。[14] Zequn Qin、Pengyi Zhang、Fei Wu 和 Xi Li，“FcaNet：频率通道注意力网络”，载于*ICCV*。2021年，第763–772页，IEEE。[15] Naoya Takahashi 和 Yuki Mitsufuji，“用于音频源分离的多尺度多频带密集网络”，载于*WASPAA*。2017年，第21–25页，IEEE。[16] Sanghyun Woo、Jongchan Park、Joon-Young Lee 和 In So Kweon，“CBAM：卷积块注意力模块”，载于*ECCV (7)*。2018年，《计算机科学讲义》第11211卷，第3–19页，Springer。[17]Hyun Joon Park、Byung Ha Kang、WooSeok Shin、Jin Sob Kim 和 Sung Won Han，“MANNER：用于噪声消除的多视图注意力网络”，载于*ICASSP*。2022年，第7842–7846页，IEEE。[18]Jie Hu、Li Shen 和 Gang Sun，“挤压与激励网络”，载于*CVPR*。2018年，第7132–7141页，计算机视觉基金会 / IEEE计算机学会。[19] DeLiang Wang 和 Jitong Chen，“基于深度学习的监督式语音分离：综述”，*IEEE ACM*音频、语音与语言处理汇刊，第26卷，第10期，第1702–1726页，2018年。[20] Jingjing Chen、Qirong Mao 和 Dong Liu，“时域监督式单通道语音分离的合成问题”，载于*INTERSPEECH*。2020年，第2627–2631页，ISCA。[21]Gordon Wichern、Joe Antognini、Michael Flynn、Licheng Richard Zhu、Emmett McQuinn、Dwight Crow、Ethan Manilow 和 Jonathan Le Roux，“WHAM!：将语音分离扩展至噪声环境”，载于*INTERSPEECH*。2019年，第1368–1372页，ISCA。[22] John R. Hershey、Zhuo Chen、Jonathan Le Roux 和 Shinji Watanabe，“深度聚类：用于分割与分离的判别式嵌入”，载于*ICASSP*。2016年，第31–35页，IEEE。[23] Zengwei Yao、Wenjie Pei、Fanglin Chen、Guangming Lu 和 David Zhang，“通过高阶潜在域中的细粒度编码逐步细化的语音分离网络”，*IEEE ACM*音频、语音与语言处理汇刊，第30卷，第378–393页，2022年。[24] Neil Zeghidour 和 David Grangier，“Wavesplit：通过说话人聚类实现端到端语音分离”，*IEEE ACM*音频、语音与语言处理汇刊，第29卷，第2840–2849页，2021年。