arXiv:2303.03005v1 [cs.SD] 6 Mar 2023

基于 Conv-Tasnet 的设备端低复杂度声源分离扩展策略

Mohamed Nabih Ali

意大利特伦托特伦托大学
mohamed.nabih@unitn.it

Francesco Paissan、*Daniele Falavigna*、*Alessio Brutti*

布鲁诺·凯斯勒基金会，意大利特伦托
(fpaissan,falavi,brutti)@fbk.eu

摘要

近年来，文献中提出了多种效果显著的单通道语音分离神经网络方法。然而，由于这些模型规模庞大、复杂度高，要在低资源设备（例如助听器和耳机）上使用仍然面临挑战，目前尚无成熟的解决方案。尽管已有基于剪枝或压缩神经网络模型的方法，但要设计出适用于特定应用领域的模型架构，往往需要采用启发式流程，而且这些流程难以直接迁移到不同的低资源平台。鉴于广为人知的 Conv-Tasnet 语音分离架构具有模块化特性，本文考察了三个直接控制模型总体规模的参数，即残差块的数量、分离块的重复次数，以及深度卷积中的通道数，并通过实验评估这些参数对语音分离性能的影响。具体而言，在 Libri2Mix 上开展的实验表明，扩张一维卷积块的数量是最关键的参数，而在残差块中使用额外扩张则有助于减小性能下降。

索引词— 语音分离，神经网络，深度学习

1. 引言

如今，各类应用都成功采用语音技术功能来提升整体终端用户满意度 [1, 2]。尽管许多应用领域已有稳健的解决方案和商业产品，但仍存在一些性能限制，尤其是在环境和噪声条件较难控制时。一个尚待解决的挑战是如何应对重叠语音出现的场景 [3, 4]。特别是得益于神经网络模型的应用，近年来单通道语音分离稳步发展，分离性能极为出色。尽管这些模型效果显著，但通常需要估算大量参数，因此会消耗可观的计算和内存资源，导致它们无法用于能力有限的设备，例如助听器或微控制器单元（MCU）。表1列出了文献中目前一些最受欢迎的分离方法在推理期间的参数数量和浮点运算次数（FLOPs）等特性。请注意，即使采用量化，所有模型运行时仍需数 MB 的随机存取存储器；其中 Dual-Path-RNN 尤其耗费计算资源。为便于理解上述数字，表2列出了一些当前处理平台的示例

表 1：近期用于声源分离的神经网络架构，以及参数总数（以百万计）和对一秒输入音频进行推理所需的十亿次浮点运算 [5]。

方法	# 参数（百万）	浮点运算次数（十亿）
Two-Step TDCN [6]	8.63	7.09
Conv-Tasnet [7]	5.10	5.23
Dual-Path-RNN [8]	2.63	48.89
SuDoRM-RF [5]	2.66	2.52

平台，涵盖从低端微控制器到单板计算机[1]。

表 2：嵌入式平台及其硬件能力示例，包括随机存取存储器和每秒百万条指令。

开发板名称	随机存取存储器[KB]	MIPS
STM32L476RG	128	80
TI MSP432P4111	256	58.56
BeagleBone Black	524288	1607
Raspberry Pi 3 B+	1048576	2800

表1中报告的最先进水平模型显然不适合在STM32L476RG或TI MSP432P4111等微控制器上运行，因为这些设备的随机存取存储器只有几百KB。在BeagleBone Black和Raspberry Pi等更强大的设备上部署这些模型也会存在问题。如果可用内存和计算能力足以运行表1中报告的某些模型（例如SuDoRM-RF，或许还有Conv-TAS-Net），那么设备上就无法运行任何其他服务。

这一问题在许多依赖神经网络模型的应用中都很常见。因此，已有大量研究致力于将大型神经网络部署到资源受限的设备上：量化 [9], 剪枝 [10], 通过知识蒸馏（KD）进行模型压缩 [11], 神经架构搜索（NAS） [12], 等等。尽管这些方法已成功应用于各种任务和领域，但它们存在一个局限：缺乏可扩展性。要获得满足特定资源受限设备要求，或更广泛地说，满足分配给当前任务的资源预算的精简模型，仍然需要进行专门设计和大量人工工作。相反，从应用角度来看，一种只需调整少量参数即可适配的可扩展模型非常具有吸引力。

[1]由于其取决于固件以及模型的实际实现，因此无法在可用的每秒百万条指令（MIPS）与浮点运算次数之间建立精确的映射关系

近年来科学文献中广泛使用的一种分离模型是 Conv-Tasnet [7]。该模型的一个重要特性是其模块化结构，这使得模型能够根据可用资源轻松缩减。基本上，模型的计算复杂度由三个参数控制：膨胀残差块的数量、分离块的重复次数，以及深度卷积中的通道数。本文的主要贡献在于测量并分析上述三个参数如何影响整体分离性能，从而为未来在智能设备上实现该神经模型提供有用建议。我们通过实验观察到，残差块的数量是一个关键参数，因为它决定了最大的膨胀因子，进而决定感受野的大小。我们还观察到，增大膨胀步长能够显著减小减少块数量所导致的性能下降。

本文结构如下：第 2 节概述与本文工作相关的一些学术论文，第 3 节介绍为降低 Conv-Tasnet 模型复杂度而提出的方法，第 4 节报告并讨论实验结果，第 5 节对全文进行总结。

2. 相关工作

近年来，研究者采用深度学习进行单通道声源分离 [13, 14, 15]。具体而言，基于编码器、解码器和分离模块的早期架构，通过引入端到端神经网络模型得到了有效改进，例如前文提到的 Conv-Tasnet。在这类模型中，所有必要的处理都在单个阶段完成，并通过优化损失函数，尽可能减小模型输出与干净目标信号之间的失真。然而，由于当前模型的计算复杂度较高，将其部署在边缘设备上（例如嵌入式应用或智能移动便携设备中）颇具挑战，因为这会大量占用内存和计算资源。因此，已有多项研究致力于开发所需计算资源较少的神经网络模块和架构。一个广受欢迎的例子是深度可分离卷积（DWS） [16, 17]，它能够以较少的算术运算实现卷积运算，也就是音视频神经网络处理方案中采用的基本变换。

在语音领域，文献 [18] 将深度可分离卷积（DWS）应用于单声道歌声分离。Mobile-Nets 是一系列轻量级深度神经网络，文献 [19] 提出将类似策略用于边缘设备。然而，作者指出，可分离卷积中较大的膨胀因子往往会引入若干伪影。近期，越来越多的研究提出元学习算法，以优化架构配置，从而高效利用计算资源并取得良好性能 [20, 21]。

关于专门针对语音领域的研究，文献 [22] 介绍了用于处理语音合成任务的模型“WaveRNN”。采用该模型的动机是，在保证可接受质量的前提下缩短生成合成样本所需的时间。这个紧凑型模型能够在 GPU 上以快于实时的速度生成 24 kHz、16 位音频 4× 。作者还提出使用剪枝技术，大幅减少模型权重数量，从而在浮点运算次数和延迟方面实现出色的性能。

[23] 的作者提出了一种基于 U-Net [24] 架构的轻量级模型，可在边缘计算设备上有效实现语音分离。为此，他们在每个卷积层中采用了类似 Inception 的多通道降维技术。

在 [25], 提出的组通信（GroupComm）中，一个N维输入特征向量被拆分为 K 个组，每组由一个维度为 M的小型特征向量表示。随后，所有组共享一个小型分离模块。采用这种方法，该模型的性能与原始双路径 RNN 模型相近，同时参数减少了 35.6 倍，乘加运算减少了 2.3 倍。

文献 [26]提出了一种两步轻量级在线语音分离方法。首先，使用基于相位的波束形成器估计目标声源。随后，使用基于双向长短期记忆网络（BLSTM）单元的时频二值掩码估计器，计算每个声源的二值掩码。

最后，在文献 [27], 中，作者采用了基于二值化和蒸馏方法的两步模型压缩方案：首先对模型进行二值化，然后在训练阶段使用知识蒸馏来提升分离性能。

需要指出的是，专门针对声源分离问题设计的方法数量有限。此外，这些方法
这些方法是针对特定资源量身定制的，因此难以轻松适配不同设备。相反，下一节所述的分析能够帮助我们确定最优网络架构，使其在给定设备可用资源的情况下，将性能下降幅度降至最低。

3. 缩放方法

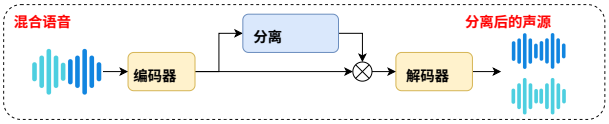

图1：Conv-TAS-Net 网络架构概念框图

图1展示了 Conv-Tasnet 网络架构的概念框图 [7]；该架构由 3 个模块组成：编码器、分离模块和解码器。在本研究中，我们重点关注分离模块，因为它是计算量最大的部分。我们不考虑缩减编码特征的维度，因为即使分离准确，这也会显著影响解码器的信号重建质量。图2展示了分离器模块：它由一系列 1-D 空洞残差卷积块和一个掩码生成器组成。分离模块的模块化特性使我们能够根据目标内存占用以及乘加次数（MAC）或浮点运算次数（FLOPs），轻松缩放模型。

具体而言，图2重点展示了本分析考虑的 3 个缩放参数，即：*i)* 膨胀块的数量B；*ii)* 膨胀块序列的重复次数 R，以及*iii)* 每个残差块中深度卷积所使用的通道数 C（见图2右侧）。尤其是，减少后者会采用倒置残差卷积块，这种做法已在 [19, 28] 中得到成功应用。基线模型中这些参数的取值分别为：$B = 8$、$R = 3$和 $C = 512$。如图所示，残差块中的膨胀率呈指数递增，具体为 $\delta = 2^{(i-1)}$，其中 i为残差块的索引（$i = 1, .., B$）。因此，B的取值尤为关键，因为它决定了模型中应用的最大膨胀率。与我们此前将 PhiNets 用于声音事件检测（SED）的研究 [29], 一致，我们发现，通过对这三个缩放参数采用不同配置，可以实现相近的内存占用量和浮点运算次数，但分离性能却可能存在很大差异。

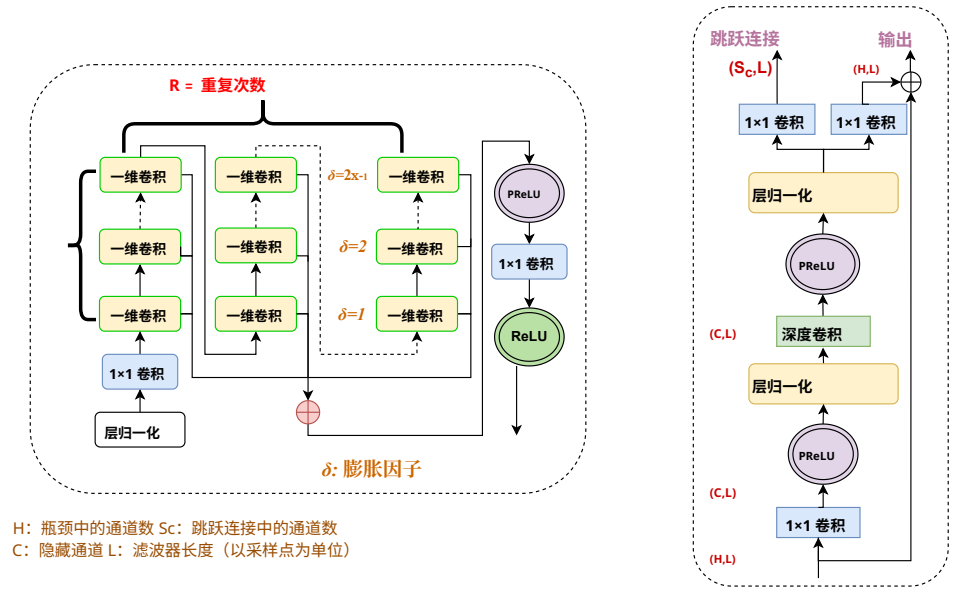

图 2：Conv-Tasnet 分离块（左）和单个一维扩张卷积块（右）的框图

4. 实验分析与结果

我们的实验基于 Asteroid 工具包 [30] 提供的模型和方案，并使用通过上述工具包获得的 Libri2Mix 数据集 [31], 进行。我们采用“train-100”训练集和“sep-clean”任务，即输入信号为两个干净说话人源的混合音频，不含噪声；采用“min”模式，即混合信号以较短的源为结尾。我们使用代码库中定义的训练参数：学习率10^{-3}、训练 200 个 epoch 并启用早停，耐心值为 30，每批包含 24 个片段。按照声源分离领域的惯例，我们使用尺度不变信号失真比（SI-SDR） [32]来衡量性能。

如上所述，我们通过改变以下参数进行了实验：Conv-Tasnet 模型中的残差块数量 B：2，4，6，8；重复块的数量 R（$R=1$，2，3）；深度可分离一维卷积块中的通道数（$C=512$，128，64）。

图3展示了各配置的性能与相应模型参数量之间的关系。正如预期，模型参数量减少会导致性能下降。不过，有些模型的参数量相近，性能却相差很大；甚至有些参数更少的模型性能还明显更好。这表明，三个参数 B、R 和 C 对性能的影响各不相同。另一个值得注意的结果是，只有当模型参数量少于 1500K 时，性能才开始明显下降，这说明模型规模可以轻松缩减至原来的 1/5，而不会导致性能大幅下降。

上述分析在表3中得到了定量体现。表3列出了各配置的总体结果，并按模型参数总量排序。总体最佳结果以粗体显示。参数量相近但分离性能迥异的模型被归为一组以便比较。例如，三元组 $B=6$，$R=3$，$C=64$（对应 $\#params=972$）的尺度不变信号失真比显著高于三元组 $B=2$，$R=2$，$C=512$（对应

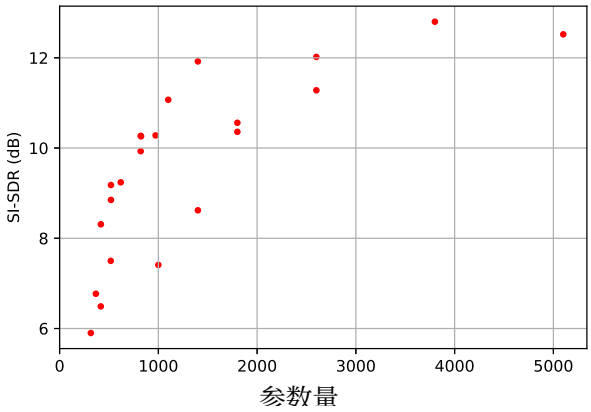

图 3：不同参数数量的分离模型所提供的尺度不变信号失真比

$\#params=1000$。其他值得关注的情况是两个拥有 1400 个参数的模型（(8,3,128)；(2,3,512)），它们的性能相差超过 3 dB，以及三个拥有 520 个参数的模型（(6,2,64)、(4,3,64)、(2,3,128)）。

观察上述情况以及表3中的总体趋势，可以明显看出，对分离性能影响最大的参数是残差块的数量 B，尤其是对于 R而言。其原因在于，膨胀率在后续残差块中呈指数级增长。使用较少的残差块会限制膨胀率，从而限制感受野的时间跨度。相反，减少通道数 C，尤其是在 B 和 R值较大时，可以显著缩小模型规模，同时将性能下降限制在较小程度。

表3：采用不同精简配置的 Conv-Tasnet 基线模型（第一行）的分离性能。B：残差块数；R：重复次数；C：深度卷积一维卷积层中的通道数。

B	R	C	参数量（K）	SI-SDR（dB）
8	3	512	5100	**12.52**
6	3	512	3800	**12.80**
6	2	512	2600	12.02
4	3	512	2600	11.28
8	1	512	1800	10.36
4	2	512	1800	10.56
8	3	128	1400	11.92
6	1	512	1400	9.77
2	3	512	1400	8.62
6	3	128	1100	11.07
2	2	512	1000	7.41
6	3	64	972	10.28
8	3	64	825	10.26
6	2	128	821	10.27
4	3	128	821	9.93
8	1	128	619	9.24
6	2	64	520	9.18
4	3	64	520	8.85
2	3	128	518	7.50
8	1	64	418	8.31
2	2	128	417	6.49
6	1	64	367	7.64
2	3	64	367	6.77
2	2	64	316	5.90

4.1. 额外扩张

增大膨胀步长是减小 B缩减影响的一种可行方法，因为膨胀率不会增加模型参数的数量，同时对计算需求的影响也较小。

表4报告了采用较大的指数膨胀步长基数（即 4 和 8）时，我们最小模型的分离性能。当 $B = 4$ 采用 $\delta = 4^{i-1}$（其中 i 是分离块的索引）时，性能与基线（表格第一行）非常接近。类似地，当 $B = 2$ 采用 $\delta = 8^{i-1}$ 时，性能最佳，但与基线仍相差甚远。这些结果证实了上述关于B 对感受野影响的观察。请注意，对 $B = 4$应用 $\delta = 8^{i-1}$ 会导致性能下降，这是合理的，因为数值 $B = 8$和 $\delta = 2^{i-1}$ 是针对这项任务专门优化的。作为比较，最后一行报告了 SuDoRM-RF 在相同任务上的性能。规模相近的 Conv-Tas-Net 缩放并采用额外扩张后的版本，性能略逊于 SuDoRM-RF，但可扩展性更高。

4.2. 知识蒸馏

为了进一步提升最小模型的性能，我们尝试从基线模型进行知识蒸馏，但未能奏效。我们采用教师模型生成的信号和掩码，尝试了多种蒸馏策略，并使用了尺度不变信号失真比和均方误差损失。不过，我们只观察到训练期间模型收敛速度更快，最终性能并无明显提升。

表 4：采用不同膨胀率时最小模型的分离性能。上排表示基线模型，下排报告 SuDoRM-RF 的性能以供比较。

B	R	C	# 参数量 (K)	尺度不变信号失真比 (dB)		
				$\delta = 2^{i-1}$	$\delta = 4^{i-1}$	$\delta = 8^{i-1}$
8	3	512	5100	12.52	-	-
4	3	512	2600	11.28	**12.02**	11.10
4	3	128	821	9.90	**10.17**	9.66
4	3	64	520	8.85	**9.21**	8.51
2	3	512	1400	8.62	8.94	**9.44**
2	3	128	518	7.50	7.83	**8.02**
2	3	64	367	6.77	7.09	**7.26**
SuDoRM-RF			≈ 2660	13.13		

5. 结论与未来工作

本文研究了缩放方法对 Conv- Tasnet 架构的影响，旨在构建适用于设备端部署的可扩展声源分离模型。我们通过实验评估了 3 个缩放参数对缩小版 Conv-TAS-Net 模型的影响，并为相关领域的从业者提供指导，以便根据目标平台简化紧凑模型的设计。我们发现，Conv-Tasnet 架构中的残差块数量对分离性能影响最大；采用更大的卷积膨胀步长，可能是弥补相应性能下降的一种可行方案。

未来的工作将进一步探究知识蒸馏或其他训练策略，以及能够减轻模型压缩对分离性能不利影响的架构修改。

6. 参考文献

[1] 李金宇等人，《稳健的自动语音识别：通往实际应用的桥梁》，*Academic Press*，2015年。[2] Jayashri Vajpai 和 Avnish Bora，《自动语音识别系统的工业应用》，《国际工程研究与应用期刊》，第88–95页，2016年3月。[3] 陆亮等人，《流式端到端多说话人语音识别》，*IEEE*信号处理快报，第28卷，第803–807页，2021年。[4] Samuele Cornell等人，《使用远距离麦克风阵列进行重叠语音检测和说话人计数》，计算机语音与语言，第72卷，第101306页，2022年。[5] Efthymios Tzinis等人，《Sudo RM-RF：用于通用音频源分离的高效网络》，载于信号处理机器学习国际研讨会，2020年，第1–6页。[6] Efthymios Tzinis等人，《两步声音源分离：基于学习得到的潜在目标进行训练》，载于*IEEE*国际声学、语音与信号处理会议，2020年，第31–35页。[7] Yi Luo和Nima Mesgarani，《Conv-tasnet：超越语音分离中的理想时频幅度掩蔽》，*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。[8] Yi Luo等人，《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*论文集，2020年，第46–50页。

[9] Jiwei Yang等人，《量化网络》，载于《IEEE/CVF计算机视觉与模式识别会议论文集》，2019年，第7308–7316页。[10] Lorenzo Valerio等人，《在互联网边缘动态硬剪枝神经网络》，arXiv预印本 arXiv:2011.08545，2020年。[11] Geoffrey Hinton等人，《神经网络中的知识蒸馏》，arXiv预印本 arXiv:1503.02531，2015年。[12] Hyunho Mo等人，《用于剩余使用寿命预测的进化神经架构搜索》，《应用软计算》，第108卷，第107474页，2021年。[13] Weninger Felix等人，《用于单通道语音分离的判别式训练循环神经网络》，载于GlobalSIP论文集，2014年，第577–581页。[14] Yuxuan Wang和DeLiang Wang，《用于时域信号重建的深度神经网络》，载于ICASSP论文集，2015年，第4390–4394页。[15] Erdogan Hakan等人，《采用深度循环神经网络的相位敏感且识别增强型语音分离》，载于ICASSP论文集，2015年，第708–712页。[16] Laurent Sifre和St´ephane Mallat，《用于纹理分类的刚性运动散射》，arXiv预印本 arXiv:1403 1687，2014年。。[17] Franc¸ois Chollet，《Xception：采用深度可分离卷积的深度学习》，载于CVPR，2017年，第1251–1258页。[18] Pyry Pyykk¨onen等人，《用于单声道歌声分离的深度可分离卷积与循环神经网络比较》，载于MMSP，2020年，第1–6页。[19] Andrew G Howard等人，《MobileNets：面向移动视觉应用的高效卷积神经网络》，arXiv预印本 arXiv:1704.04861，2017年。[20] Jiahui Yu和Thomas S Huang，《通用可伸缩网络与改进的训练技术》，载于ICCV，2019年，第1803–1811页。[21] Han Cai等人，《Once-for-all：训练一个网络并针对高效部署进行专门化》，arXiv预印本 arXiv:1908 09791，2019年。。[22] Nal Kalchbrenner等人，《高效神经音频合成》，载于ICML，2018年，第2410–2419页。[23] Kwang Myung Jeon等人，《面向边缘计算设备的轻量级U-Net单声道语音源分离》，载于IEEE国际消费电子会议，2020年，第1–4页。[24] Daniel Stoller等人，《Wave-U-Net：用于端到端音频源分离的多尺度神经网络》，arXiv预印本 arXiv:1806.03185，2018年。[25] Yi Luo等人，《通过组通信实现超轻量级语音分离》，载于ICASSP论文集，2021年，第16–20页。[26] Alejandro Maldonado等人，《基于双向长短期记忆网络的二值掩码，实现目标声源的轻量级在线分离》，《计算与系统》，第卷、期、页码未注明。[27] Xiuyi Chen等人，《用于单声道语音分离的蒸馏二值神经网络》，载于国际神经网络联合会议，2018年，第1–8页。[28] Francesco Paissan 等人，《PhiNets：面向边缘端低功耗人工智能的可扩展骨干网络》，*ACM* 嵌入式计算系统汇刊，2022年2月。[29] Alessio Brutti 等人，《针对低端设备上的城市声音检测优化 PhiNet架构》，收录于*EUSIPCO*，2022年，第1121–1125页。[30] Manuel Pariente 等人，《Asteroid：面向研究人员的基于 PyTorch 的音频源分离工具包》，收录于国际语音通信协会年会论文集，2020年。[31] Joris Cosentino 等人，《LibriMix：用于通用语音分离的开源数据集》，2020年。[32] Emmanuel Vincent 等人，《盲音频源分离中的性能测量》，*IEEE* 音频、语音与语言处理汇刊，第14卷，第4期，第1462–1469页，2006年。