用于歌声分离的混合式 Y-Net 架构

Rashen Fernando电气与电子工程系，佩拉德尼亚大学，佩拉德尼亚，斯里兰卡 e16103@eng.pdn.ac.lk

Pamudu Ranasinghe 电气与电子工程系，佩拉德尼亚大学，佩拉德尼亚，斯里兰卡 e16306@eng.pdn.ac.lk

Udula Ranasinghe 电气与电子工程系，佩拉德尼亚大学，佩拉德尼亚，斯里兰卡 e16309@eng.pdn.ac.lk

Janaka Wijayakulasooriya电气与电子工程系，佩拉德尼亚大学，佩拉德尼亚，斯里兰卡 jan@eng.pdn.ac.lk

Pantaleon Perera 工程数学系，佩拉德尼亚大学，佩拉德尼亚，斯里兰卡 pperera@pdn.ac.lk

arXiv:2303.02599v1 [cs.SD] 5 Mar 2023

摘要—本文提出一种名为 Y-Net 的新型深度学习神经网络架构，用于实现音乐源分离。该架构通过从频谱图域和波形域提取特征，执行端到端混合源分离。受 U-Net 架构启发，Y-Net 预测频谱图掩码，以便从混合信号中分离人声源。实验结果表明，所提出的架构能够以更少的参数有效地实现音乐源分离。总体而言，我们的工作为提高音乐源分离的准确性和效率提供了一种有前景的方法。

索引词—短时傅里叶变换、频谱图、波形、U-Net、源分离、Y-Net

I. 引言

音乐源分离是一项复杂任务，旨在将音频混合信号中的各个声源分离出来，可应用于制作卡拉 OK 伴奏、音乐转录和音乐制作。深度神经网络（DNN）已被用于有效执行音乐源分离，主要有两种方法，分别利用时频表示或时域表示。本文使用 DNN 从音频混合信号中分离歌声。

尽管时频表示中蕴含着许多有待发现的模式，但短时傅里叶变换（STFT）未必是音乐源分离的最佳选择 [1]。人声通常出现在低频带，因此这些区域需要更高的分辨率。然而，使用频谱图表示时，时间分辨率与频率分辨率之间存在权衡。将幅度频谱图输入网络是常见做法。这种方法可能无法完整表示音频信号，因为其中缺少相位信息。近期一篇题为《Investigating U-Nets with various intermediate blocks for spectrogram-based singing voice separation》[2] 的论文提出，在将相位频谱图和幅度频谱图输入网络前先将二者结合，以提供额外信息来提升分离效果。然而，对网络进行的权重分布分析表明，在分离任务中并未充分利用相位信息。

先前的研究探讨了利用音乐源分离端到端模型，将原始音频信号中的所有可用信息（包括相位）纳入处理的可行性。此方法旨在避免遗漏可能有价值的信息 [3], [4], [5]。尽管原始音频信号同时包含幅度和相位信息，但仅凭原始音频可能难以学习此类信号的内在结构。显然，不同的表示方式会产生不同的伪影。

本文提出一种新颖的架构，结合原始音频与频谱图表示的优势，以估计用于分离歌声的时频掩码。该架构由两个主要部分组成：可学习滤波器和 Y-Net。可学习滤波器的优势在于能够学习提取关键特征的滤波器，这些特征在短时傅里叶变换频谱图中并不存在，同时还能纳入相位信息。Y-Net 从两个域中提取特征，并重建歌声的频谱图。

II. 相关工作

我们的网络受 U-Net架构 启发 [6]。它使用一组卷积层将图像编码为紧凑且深层的表示，在底层进行合并，并在编码器和解码器中处于相同层级的各层之间添加额外的跳跃连接，使低层信息能够从高分辨率输入直接流向高分辨率输出。在音乐源分离领域，U-Net架构 被广泛用于估计掩码，从混合信号频谱图中分离出单个乐器 [7]。相比之下，频谱图通道 U-Net [8] 直接生成频谱图，而不是估计掩码。这种方法能够分离

多种乐器，但可能会在音频信号中产生不需要的伪影。

所提出的方法利用原始音频和频谱图数据，从混合音频中提取尽可能优质的特征，旨在通过在时频域生成掩码来分离歌声。在本研究中，我们引入了一种新技术，将原始音频数据转换为频谱图维度。具体而言，可学习频谱图经过设计用于学习滤波器系数，随后利用这些系数从混合音频波形中提取相关特征。该方法受到小波变换方法的启发 [9]。可学习频谱图基于 Spectnet [10]，，后者采用可学习的一维卷积滤波器从输入波形中提取不同频率，并创建频谱图像。

波形域源分离最早由“端到端音乐源分离” [3]，进行探索，该方法采用扩张卷积，但与当时由频谱图驱动的模型相比表现不佳。Convtasnet [4]，是一种基于掩码的模型，通过使用扩张卷积在学习到的时频表示上进行掩码处理；2019年，由于端到端波形模型能够纳入相位信息，其性能超过了当时基于频谱域的现有最先进模型。但根据人工评估结果，该模型存在严重的音频伪影问题。在 2021 [11]，中，有研究提出了一种改进的新型波形到波形模型，该模型采用 U-Net 结构和双向长短期记忆网络，并以 Wave-U-Net [5]，为基础构建。Wave-U-Net 是将 U-Net 适配到一维时域的模型，其性能优于 Convtasnet。然而，Wave-U-Net 使用速度更快的步幅卷积和更多的通道，这可能会引入混叠伪影。

E.M. Grais [12] 强调了频谱图中时频分辨率对特征提取的重要性，其中低频和高频频带采用不同尺寸的卷积核进行处理。这些论文质疑了频谱图的可靠性。与传统频谱图相比，可学习频谱图能够以更高分辨率更可靠地捕捉低频，因为权重偏置已证明在这方面颇为有效。这是音乐源分离技术的一项重要进展。J. Paulus 和 M. Torcoli [13] 指出，以 8kHz 或 44kHz 对输入混合音频进行采样，并不能显著提升人声分离质量，因为神经网络能够在 U-Net核心中提取相同的特征。尽管语音信号存在于低频（8 kHz），演唱人声有时也能达到 17 kHz。可训练的频谱图能够确定从原始音频中提取哪些最佳特征以分离人声，而 Y-Net 则可以从输入频谱图中恢复缺失的特征。混合频谱图与波形源分离 [14] 扩展了原始 U-Net架构，并提供两个并行分支：一个位于时间（时域），另一个位于频率（频域）；两个分支在 U-Net核心中共享特征，同时各自配有一个解码器。Hybrid Transformer Demucs [15] 将两个 U-Net架构最内层的层替换为 Transformer层，这就需要大量数据用于训练。

与我们提出的 Y 形架构不同，以往的研究采用了 X 形架构，为每个输入分别设置分支 [15], [14]。我们的方法仅使用一个解码器分支，因此与上述研究相比，参数量更少、训练速度更快、计算成本更低，且对数据的需求更少。

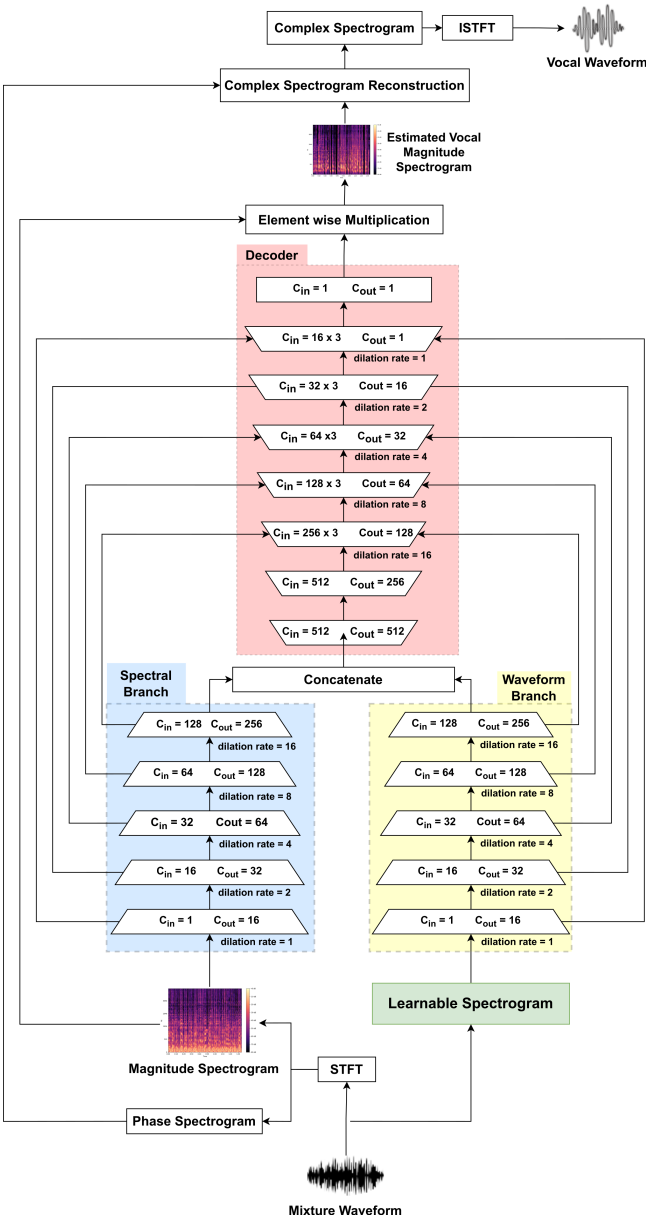


图 1：Y-Net 架构：以原始音频波形和幅度频谱图为输入，并输出分离的人声的幅度频谱图

III. 架构

本文提出了一种神经网络架构，以原始音频波形和频谱图作为输入，并估计出一个与输入频谱图形状相似的掩码，该掩码可以

用于从混合音频中分离并提取歌声。具体而言，将估计出的掩码与原始混合频谱图相乘，即可获得歌声的幅度频谱图。

值得注意的是，我们的网络不会改变原始混合信号的相位。这种方法与 C. Sun 等人提出的、用于单声道录音语音分离的注意力框架卷积循环神经网络相似 [16]。为了获得最终预测的源信号，我们将估计的幅度频谱图与原始混合信号的相位频谱图结合，生成复数频谱图。如图1所示，我们通过逆短时傅里叶变换（ISTFT）获得人声波形。通过保留相位信息，预测的源信号得以保留原始时间结构和声音连贯性。

我们的网络由两个编码器分支组成，分别是波形分支和频谱分支，它们分别接收原始音频和频谱图。原始音频编码器分支的前端包含一个可学习频谱图模块，可将原始音频转换为频谱图格式。随后，两个编码器分别从两种输入中提取特征，并在网络核心处合并，再传送至解码器。解码器由一组上采样层组成，可将特征解码至频谱图的原始尺寸，如图1所示。值得注意的是，两个编码器与解码器的结构是对称的。

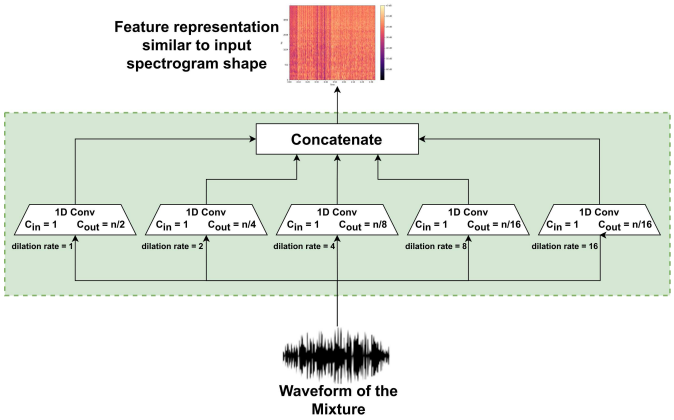
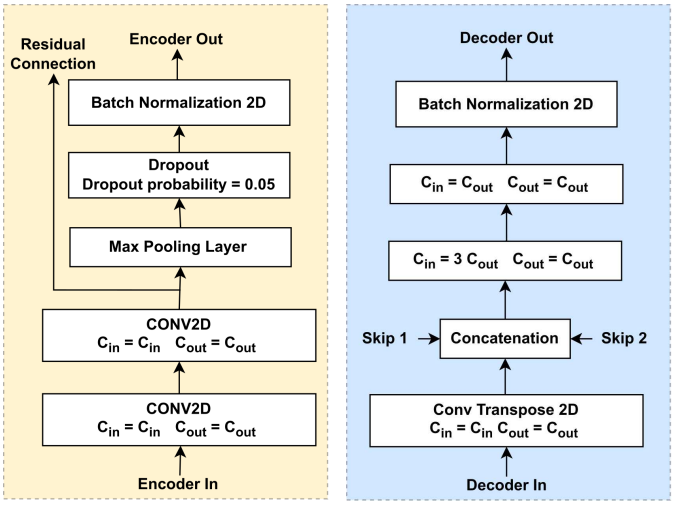


图 2：模型以原始音频波形为输入，并使用一维卷积层提取特征。输出为特征表示，其大小与输入频谱图相近。

A. 可学习频谱图

可学习频谱图的架构如图2所示。为了生成可学习频谱图，首先，将单通道音频依次输入五组具有不同膨胀率（1、2、4、8、16）的一维卷积，使网络能够捕获长期依赖关系。对于每一组，膨胀率较低的卷积层会分配更多卷积核，因为局部特征对人声的影响比全局特征更显著。因此，我们优先考虑膨胀率为1的卷积组。五组卷积层中的卷积核总数等于输入频谱图中的频率分箱数，在我们的模型中为1024。

长度为 2048 的卷积核采用步幅，在连续帧之间形成不同的重叠，并生成长度等于输入频谱图中时间分箱数量的向量。最后，将所有卷积信号拼接起来，得到形状与输入频谱图相近的输出。每个一维卷积后均使用 LeakyReLU 作为激活函数。此处，卷积核充当滤波器，类似于小波变换中的母小波。



(a) 编码器层　　(b) 解码器层

图 3：(a)：编码器层包含二维卷积层和最大池化层，并在最大池化层之前引出一条跳跃连接。(b)：解码器层包含一个上采样层，后接两个二维卷积层；来自频谱分支和波形分支中编码器层的两条跳跃连接在卷积层之前拼接

B. 编码器

随后，可学习频谱图的表示经过五个编码器层处理：随着膨胀率的增加，通道数逐层加倍；同时，输入频谱图像的尺寸在网络逐层深入时减半。每个编码器层由一个批归一化层、两个卷积核尺寸为 3x3 的卷积层、一个最大池化层和一个 dropout 层组成。每个卷积层之后均使用 LeakyReLU 作为激活函数。单个编码器层的架构如图 3(a)所示。

在输入频谱分支之前，我们使用短时傅里叶变换将原始音频混合波形转换为频谱图。在我们的实验中，输入频谱图包含 1024 个频率分箱和 128 个时间分箱。频谱图经过与波形分支类似的五编码器结构处理，不同之处在于，编码器中每次卷积后的激活函数都替换为 ReLU。在第五个编码器块之后，

网络将从频谱分支和波形分支提取的特征进行拼接，前提是它们的尺寸相同。此拼接操作在网络的核心部分进行，随后接入两个卷积层和一个批归一化层。然后，使用 ReLU 激活这些拼接后的特征，再将其传递给网络的解码器组件。

C. 跳跃连接

跳跃连接是我们架构中的一个重要组成部分，因为它们能够将来自两个编码器分支的信息传递至共享解码器分支，绕过网络中的中间处理阶段。使用跳跃连接有助于减轻下采样和上采样操作带来的影响，否则这些操作可能会导致音频信号中的重要信息丢失。此外，跳跃连接在网络的编码器层与解码器层之间提供了直接连接，使模型能够更好地保留输出音频信号中的高频细节和细粒度结构。

D. 解码器

编码器输出依次经过 5 个解码器层，每个解码器层由一个上采样层、两个经 ReLU函数激活且卷积核尺寸为 3x3 的卷积层，以及一个批归一化层组成。每个解码器层有三个输入，分别来自前一个解码器层，以及两个来自频谱分支和波形分支中编码器层的跳跃连接。最后一个解码器层的输出随后经过一个卷积核尺寸为 1x1 的卷积层，再由 Hardtanh 激活，以增强人声特征并抑制频域中的其他特征。单个解码器层的架构如图3 (b) 所示。

最后，解码器的输出给出估计的掩码。将解码器的输出与输入混合频谱图进行逐元素乘法，即可得到歌声的频谱图。

总而言之，编码器-解码器架构提供了一种方法，可从两个输入中提取有意义的特征，然后利用这些特征生成掩码，以分离歌声。

IV. 结果与实验

A. 数据集

MUSDB18 是音乐源分离领域研究中常用的数据集 [17]。它汇集了各种流派和风格的专业制作音乐曲目。这些曲目包含多个声源，例如人声、鼓、贝斯及其他乐器。我们从 MUSDB18 数据集中总共使用了 150 首歌曲，其中 100 首用于训练、25 首用于验证，另有 25 首用于测试。首先，我们将以 44千赫采样的单声道人声和混合音频曲目切分为长度为 67072 个采样点的时间窗口，每个窗口约为 1.5 秒。

B. 实验设置

我们使用 PyTorch 框架开发了该模型。神经网络在 MUSDB18 数据集上训练了 100 个 epoch，使用配备 32GB 内存的 Nvidia V100 GPU。每个 epoch 中，整个数据集都以批量大小为 16 的批次输入网络，共计 497 个批次。训练输入为混合音频波形和 STFT幅度频谱图，相应的人声频谱图则作为训练过程中的真实值。波形域中的单个数据元素包含 67072 个数据点。我们以 2048 个采样点为窗口，并采用 75% 的重叠率进行分帧，以获取幅度频谱图，得到 128 个时间分箱和 1024 个频率分箱。我们使用均方误差（MSE）作为损失函数，训练目标是最小化真实值频谱图与估计频谱图之间的 MSE。我们使用 Adam优化器 [18]，学习率为 0.0001。

C. 评估指标

为客观评估源分离的性能，我们采用信号失真比（SDR） [19], 尺度不变信噪比（SI-SNR） [4], 和短时客观可懂度（STOI） [20] 作为模型的评估指标。SDR、SI-SNR 和 STOI 的值越高，表示质量越好。SDR 和 SI-SNR 的值可以超过 1，而 STOI 的值介于 0 到 1 之间。表 I 使用上述指标评估了网络改动，包括 U-Net、U-Net + 可学习频谱图以及 Y-Net。

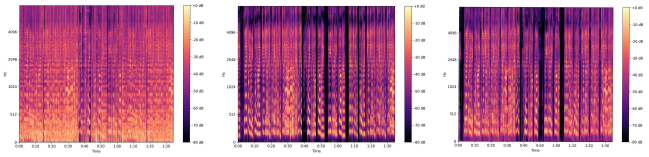
D. U-Net

起初，我们将 U-Net 用作歌声分离的基线架构，其输入为混合音频频谱图。然而，正如本文前文所述，我们发现传统频谱图的性能存在问题。因此，我们探索了其他策略。U-Net 会在输出端生成一个掩码，将其与输入混合频谱图相乘即可得到歌声频谱图。尽管掩码估计方法在客观评估指标上表现尚可，但我们发现，直接估计人声频谱图能够有效抑制泄漏到人声频谱图中的伴奏部分，从而提升主观性能，但会导致 SDR 降低。为弥补研究中缺少受试者的不足，我们选择仅关注客观指标。

E. U-Net + 可学习频谱图

我们的方案是在 U-Net架构的前端加入可学习频谱图，将原始音频转换为类似频谱图的格式，其中包含适合源分离的特征。我们将这一方法与 U-Net架构相结合，以混合波形作为输入，并在 U-Net末端直接预测人声频谱图。然而，该方法的性能不如

以频谱图为输入的 U-Net。具体而言，使用可学习频谱图技术导致信号失真比下降了 1.86 dB。另一方面，我们观察到，该方法需要更大的数据集，因为理解音频的底层结构具有挑战性。



(a) **(b)** **(c)**

图4： (a) 混合音频信号 (b) 歌声（真实值）(c) Y-Net估计的歌声的梅尔频谱图

F. Y-Net

我们进一步研究了一种Y形网络，该网络将从频谱域和波形域提取的特征进行拼接。与基线相比，这种方法显著提升了源分离任务的性能，使信号失真比提高了0.95 dB。此外，我们对网络进行了改进：波形分支采用带泄漏的ReLU激活函数，而频谱分支采用ReLU函数。我们尝试了多种输出层激活函数，例如阈值函数、ReLU、Sigmoid函数和Hardtanh。通过实证分析和实验，我们发现，在输出层中加入Hardtanh激活函数显著提升了性能，而且所需训练量更少。图4展示了混合音频信号、真实值（人声）以及Y-Net的输出（分离的人声）的梅尔频谱图。

表 I：U-Net 模型、可学习频谱图 + U-Net 与 Y-Net 在人声源分离性能上的比较。信号失真比越高越好

方法	**SDR**	尺度不变信噪比	**STOI**
U-Net	6.61	5.82	0.55
U-Net+可学习频谱图	4.75	3.93	0.54
Y-Net	**7.56**	**6.13**	**0.62**

V. 结论

本研究旨在通过将 U-Net架构改造为 Y-Net，提升 MUSDB18 数据集中的音乐源分离效果。我们的改进包括引入并行的波形和频谱图编码器分支，并使其汇聚到一个公共核心，从而通过一种基于频谱图的原始音频特征提取方法，利用频谱图特征和原始音频特征。然而，在根据估计的频谱图还原波形时，ISTFT 会导致信号失真比降低，这是我们方法的一项局限。解决这一问题的一种可能方案是使用端到端波形网络。

参考文献

[1] M. Goto 和 H. Hashiguchi，“基于掩蔽原理的改进型谐波-打击乐分离及其在真实世界音频信号中的应用”，载于 IEEE 声学、语音与信号处理国际会议（ICASSP）论文集，2003 年，第 105–108 页。[2] W. Choi、M. Kim、J. Chung、D. Lee 和 S. Jung，“研究用于基于频谱图的歌声分离、具有不同中间模块的 U-Net”，ISMIR 2020。[3] F. Llu´ıs、J. Pons 和 X. Serra，“端到端音乐源分离：波形域中是否可行？”，载于 Interspeech，2018。[4] Y. Luo 和 N. Mesgarani，“Conv-TasNet：超越语音分离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，1266，2019 年 8 月，第 1256 页。[5] D. Stoller、S. Ewert 和 S. Dixon，“Wave-U-Net：用于端到端音频源分离的多尺度神经网络”，ismir2018。[6] A. Jansson、E. Humphrey、N. Montecchio、R. Bittner、A. Kumar 和 T. Weyde，“利用深度 U-Net 卷积网络进行歌声分离”，2017。[7] V. S. Kadandale、J. F. Montesinos、G. Haro 和 E. G´omez，“用于音乐源分离的多通道 U-Net”，载于 2020 年 IEEE 第 22 届多媒体信号处理国际研讨会（MMSP），2020 年，第 1–6 页。[8] J. Oh、D. Kim 和 S.-Y. Yun，“频谱图通道 U-Net：一种将每个通道视为各个源的频谱图的源分离模型”，arXiv 预印本 arXiv:1810. 11520，2018。[9] P. M. Bentley 和 J. T. E. McDonnell，“小波变换：简介”，《电子与通信工程期刊》，第 6 卷，第 4 期，第 175–186 页，1994。[10] M. Ansari、T. Hasan 等人，“SpectNet：使用可学习频谱图进行端到端音频信号分类”，arXiv 预印本 arXiv:2211. 09352，2022。[11] A. D´efossez、N. Usunier、L. Bottou 和 F. Bach，“波形域中的音乐源分离”，ICLR 2020。[12] E. M. Grais、F. Zhao 和 M. D. Plumbley，“用于歌声分离的多频带多分辨率全卷积神经网络”，载于 2020 年第 28 届欧洲信号处理会议（EUSIPCO），2021 年，第 261–265 页。[13] J. Paulus 和 M. Torcoli，“采样频率无关的对话分离”，载于 2022 年第 30 届欧洲信号处理会议（EUSIPCO），2022 年，第 160–164 页。[14] A. D´efossez，“混合频谱图与波形源分离”，arXiv 预印本 arXiv:2111. 03600，2021。[15] S. Rouard、F. Massa 和 A. D´efossez，“用于音乐源分离的混合 Transformer”，arXiv 预印本 arXiv:2211. 08553，2022。[16] C. S un 等人，“一种用于单声道录音语音分离的注意力框架卷积循环神经网络”，《科学报告》，第 11 卷，第 1 期，第 1434 页，2021。[17] Z. Rafii、A. Liutkus、F.-R. St¨oter、S. I. Mimilakis 和 R. Bittner，“用于音乐分离的 MUSDB18 语料库”。2017 年 12 月。[18] Diederik P. Kingma 和 Jimmy Ba，“Adam：一种随机优化方法”，2014
[19] Emmanuel Vincent、Remi Gribonval 和 C ´edric ´ Fevotte，“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。[20] M. Kolbæk、Z.-H. Tan 和 J. Jensen，“通过最大化短时客观可懂度度量，使用深度神经网络进行单声道语音增强”，载于 2018 年 IEEE 声学、语音与信号处理国际会议（ICASSP），2018 年，第 5059–5063 页。