蜂窝网络语音增强：去除背景噪声和传输噪声

Amanda Shu、*Hamza Khalid*、*Haohui Liu*、*Shikhar Agnihotri*、*Joseph Konan*、*Ojas Bhargave*

卡内基梅隆大学

{amshu、hkhalid、haohuil、sagnihot、jkonan、obhargav}@andrew.cmu.edu

arXiv:2301.09027v1 [cs.SD] 22 Jan 2023

摘要

语音增强的主要目标是在保留目标语音的同时降低背景噪声。当说话者身处嘈杂环境，并接到背景噪声和传输噪声都很大的电话时，常常会陷入两难。为解决这一问题，深度噪声抑制（DNS）挑战赛致力于利用新一代深度学习模型消除背景噪声，从而增强目标语音；然而，研究人员未能考虑网络电话（VoIP）应用中的传输噪声。我们的工作聚焦于 Google Meet 及其蜂窝网络应用，在 VoIP DNS 挑战赛的 Google Meet 转电话赛道上取得了最先进的性能。本文展示了如何超越业界表现，达到 1.92 的感知语音质量评估（PESQ）和 0.88 的短时客观可懂度（STOI），并且在多项指标上实现出色的声学保真度、感知质量和可懂度。

1. 引言

Google Meet、Zoom 和 Microsoft Teams 等在线网络电话（VoIP）会议平台允许用户通过音频会议功能，使用普通电话加入会议 [1]。经常使用此选项的用户包括正在外出、互联网连接不佳，或希望免提加入会议的人 [2]。鉴于这一选项带来的便利，它已成为不可或缺的功能，也促使我们聚焦于 VoIP DNS 挑战赛的 Google Meet 转电话赛道。 [1]

然而，音频会议会带来传输噪声问题。语音质量不仅会受到通话时存在的背景噪声影响，还会因传输信道而下降，无论传输是通过电话线路还是无线传输 [3]。移动电话传输中的这一问题尤为突出，网络拥塞和丢包等其他因素也会进一步降低传输质量 [4]。这些因素会导致通信过程中的信息丢失 [5]。

因此，重要的是引入语音增强方法来降低传输噪声，从而缓解信息丢失。现有的深度学习方法主要专注于去除背景噪声，Google Meet 本身也提供了“噪声消除”功能，利用深度学习模型解决这一问题。然而，用于去除传输噪声和背景噪声的语音增强仍鲜少受到探索。我们的目标是探索语音增强模型，提升通过 Google Meets 通话传输的音频的语音质量。所有实验和消融研究的代码可在 https://github. com/hamzakhalidhk/11785-project 找到

[1]https://github.com/deepology/VoIP-DNS-Challenge

2. 背景

在现有研究中，语音增强任务所用的数据集通常是合成的，即将噪声添加到干净音频中。这是因为监督优化需要噪声输入与纯净目标配对。 [6] 推出了 INTERSPEECH 2020 深度噪声抑制挑战赛（DNS），并采用可复现的合成方法。该方法使用 Librivox 语料库中的纯净语音，以及来自 Freesound 和 AudioSet 的噪声，能够合成超过 500 小时的含噪语音。VoIP DNS 挑战赛提供了 20 小时的 Google Meets 至电话中继通话数据，其中包括在信号源处添加的合成背景噪声，以及在接收端产生的真实传输噪声。

在 2020年 DNS 挑战赛中当前最先进的语音增强模型中，本研究重点关注 Demucs 和 FullSubNet。Demucs 在时域中处理波形输入，其实时降噪架构由 Defossez 等人提出 [7]。FullSubNet 在时频域中处理频谱图输入，由 Hao 等人开发 [8]。基线 Demucs 的感知语音质量评估得分为 1.73，短时客观可懂度得分为 0.86；基线 FullSubNet 的感知语音质量评估得分为 1.69，短时客观可懂度得分为 0.84。由于这两种基线模型都未考虑传输噪声，我们的研究将聚焦于微调并提升 VoIP DNS 挑战赛中的声学保真度、感知质量和可懂度。

3. 数据集

我们的数据集基于微软深度噪声抑制挑战赛（DNS）发布的开源数据集中的音频片段[2] 构建。具体而言，我们使用了 Reddy 等人所述 DNS 数据中不含混响的合成音频片段测试集 [6]。我们的新数据集包含这些音频片段，并加入了通过蜂窝设备拨打 Google Meet 通话时产生的传输噪声。该数据集按照以下步骤创建。

首先，从 DNS 数据集中随机选取纯净语音和噪声音频。随后将这些音频文件混合，生成带有背景噪声的纯净语音噪声音频。接下来，为了加入 Google Meet 通话传输到手机会话中的传输噪声，我们启动一个 Google Meet 会话，并使用连接 T-Mobile 网络的手机拨入该会话。然后，我们在 Google Meet 中播放带噪语音音频，并将传输到手机的音频录制到音频接口中。

最终录制的音频同时包含 Google Meet 通话中的背景噪声和传输噪声。此外，我们还分别在开启和关闭 Google Meet 噪声消除功能的情况下录制音频，以便将我们的模型与行业标准进行比较。我们将关闭 Google Meet 语音增强时的音频称为 低，而将音-

[2]https://github.com/microsoft/DNS-Challenge/tree/interspeech2020/master/数据集

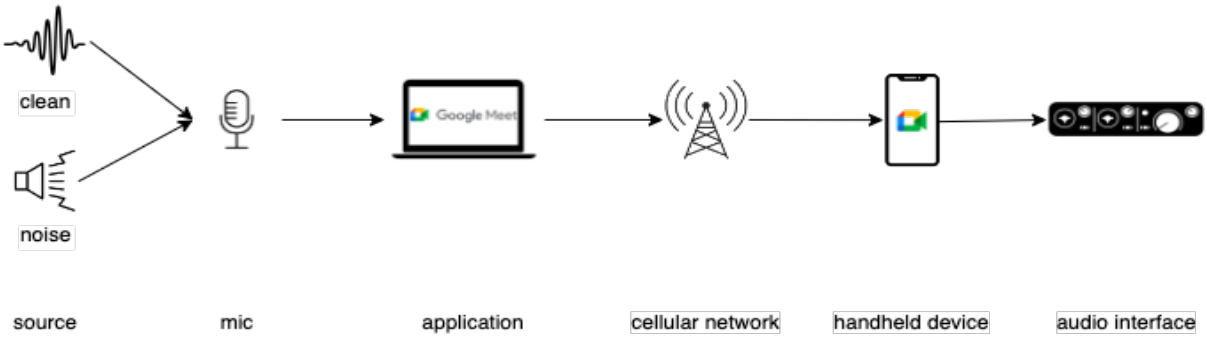

图 1：端到端数据处理流程图：从合成到录制。

频增强开启时的音频称为 自动。在我们的工作中，对于自动和低这两种数据来源，我们分别使用 400 个音频片段（每段 30 秒）作为训练样本，并使用 150 个 10 秒的音频片段作为测试样本。我们还对音频文件进行增益归一化，使其振幅范围一致，以便于训练。这一处理使音频文件的振幅范围保持一致，从而有助于模型训练。图 1 展示了数据集合成流程。

4. 模型

我们选用 Demucs 和 FullSubNet（见第 4.1 节和第 4.2 节）作为本文的基线模型，因为这两种模型在 DNS 挑战赛 [8, 9] 中都取得了最先进的结果。Demucs 在编码器-解码器结构之间包含 2 个长短期记忆网络层。FullSubNet 是一种融合模型，它结合了能够捕获全局频谱上下文的全频带模型，以及能够封装局部频谱模式的子频带模型，用于单通道实时语音增强。

4.1. Demucs

Demucs 的架构深受 SING [10], 和 Wave-U-Net [11]架构的启发。它由一个卷积编码器、一个长短期记忆网络和一个卷积解码器组成。编码器和解码器通过跳跃式 U-Net 连接相连。模型的输入是立体声混合信号 $s = \sum_i s_i$，输出则是每个声源的立体声估计值 $\hat{s}_i$。图 2 (a) 展示了完整模型的架构。

Demucs 的准则最小化波形之间的 L1 范数 $\mathcal{L}_1$ 与幅度频谱的多分辨率 STFT 损失 $\mathcal{L}_{STFT}$ 之和。

$$\begin{aligned}\mathcal{L}_{demucs} &= \frac{1}{T}[||y-\hat{y}||_1 + \sum_{i=1}^{M}\mathcal{L}_{STFT}^{(i)}(y,\hat{y})] \\ &\equiv \mathcal{L}_{L1} + \mathcal{L}_{STFT}\end{aligned}$$

如果没有 $\mathcal{L}_{STFT}$，我们会观察到音调伪影。我们将在第 5.2 节讨论消融 $\mathcal{L}_{STFT}$ 以及更新近的辅助损失。

4.2. FullSubNet

FullSubNet 是一种全频带与子频带融合模型，两者具有相似的拓扑结构。模型包含两个堆叠的单向长短期记忆网络层和一个线性（全连接）层。两者唯一的区别是，与全频带模型不同，子频带模型的输出层不使用任何激活函数。图 2（c）展示了完整的模型架构。

FullSubNet 采用复数理想比率掩码（$cIRM$）作为模型的学习目标。他们在训练中使用双曲正切函数压缩 $cIRM$，并在推理时使用反函数对掩码进行解压缩（$K = 10$，$C = 0.1$）。

5. 方法

5.1. 基线方法

在 theVoIP DNS 挑战赛的 Google MeetsTo Phone 赛道中，我们使用预训练的 Demucs 和 FullSubNet 确定基线性能。我们的工作在此基础上，通过消融研究多种评价标准来进一步提升性能。第 5.4 节的表 1 和表 2 展示了消融实验结果。

5.2. TAPLoss

我们在训练过程中引入 TAPLoss，使模型在我们的数据上超越当前最先进的语音识别模型。TAPLoss 涉及一组包含 25 个时域声学参数的指标，包括频率相关参数：基频、抖动、F1、F2、F3 频率和带宽；能量或振幅相关参数：微扰、响度、谐波噪声比（HNR）；频谱平衡参数：阿尔法比、Hammarberg 指数、频谱斜率、F1、F2、F3 相对能量、谐波差；以及其他时域参数：响度峰值率、有声区段/无声区段长度的均值和标准差，以及每秒连续有声区段数。

5.2.1. 增强的 *Demucs* 损失

以往研究表明，Demucs 模型容易产生音调伪影。多分辨率 STFT 损失会加剧这一问题，因为与时域相比，音调伪影引入的误差在时频域中更显著、更明显。因此，我们降低了 $\mathcal{L}STFT$ 对 Demucs 损失的影响，并额外引入 TAP损失 $\mathcal{L}_{\mathcal{TAP}}$。新的 Demucs 损失函数定义如下：

$$\mathcal{L}_{Demucs} = \mathcal{L}_1 + \lambda_1 \cdot \mathcal{L}_{\mathcal{TAP}} + \lambda_2 \cdot \mathcal{L}_{STFT}$$

5.2.2. 增强的 *FullSubNet* 损失

我们还通过引入 $\mathcal{L}_{\mathcal{TAP}}$ 损失扩展了 FullSubNet 损失。新的 FullSubNet 损失定义如下：

$$\mathcal{L}_{FullSubNet} = \mathcal{L}_{cIRM} + \gamma \cdot \mathcal{L}_{\mathcal{TAP}}$$

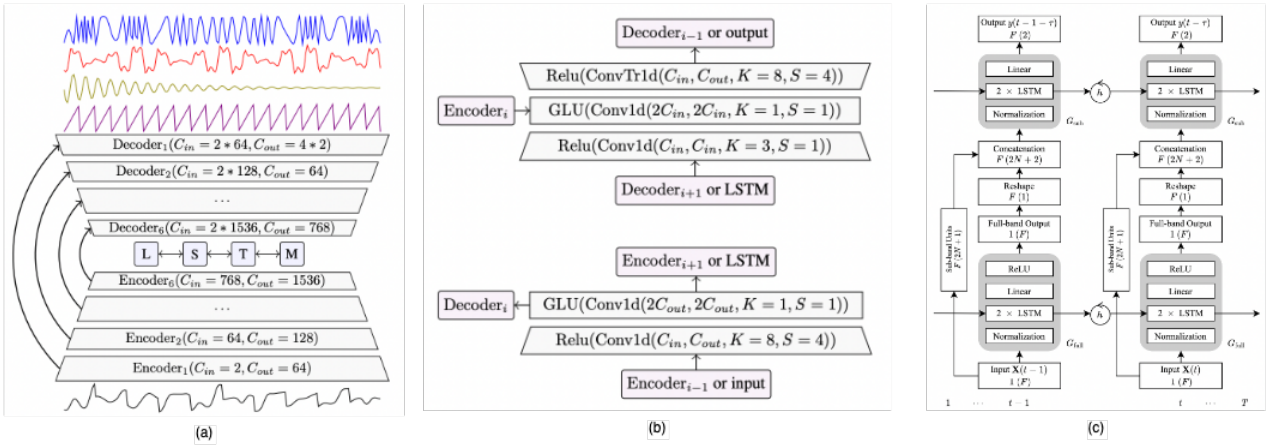

图 2： *(a) Demucs* 架构，以混合波形作为输入，并输出四个声源的估计结果。箭头表示 *U-Net* 连接。*(b)* 顶部 $Decoder_i$ 和底部 $Encoder_i$ 各层的详细视图。箭头表示与模型其他部分的连接。*(c) FullSubNet* 架构。矩形中的第二行描述当前阶段数据的维度，例如，*"1 (F)"*表示一个 *F* 维向量。*"F (2N + 1)"*表示 *F* 个相互独立的 *(2N + 1)* 维向量。

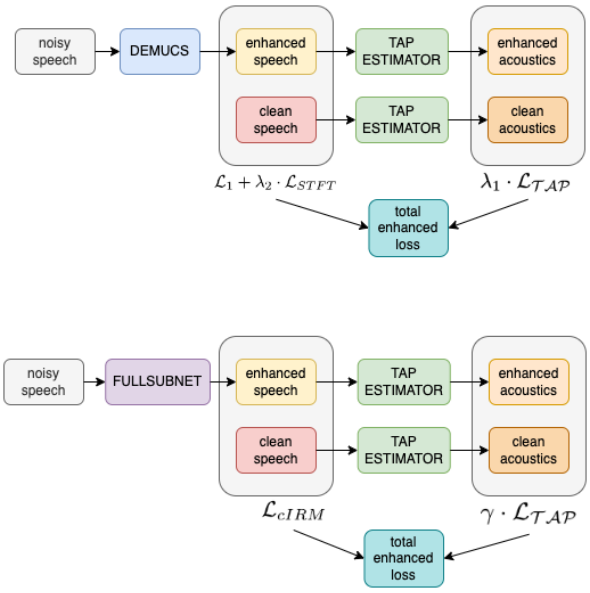

图 3： *Demucs* 和 *FullSubNet* 的训练流程

5.3. 消融实验

我们进行了消融实验，以确定 Demucs 的每个超参数 λ_1 和 λ_2 以及 FullSubNet 的 γ 的最佳取值。这些最佳取值是根据它们在感知语音质量评估（PESQ）和短时客观可懂度（STOI）指标上的表现确定的，这些指标的定义见表 4。表 1 和表 2 分别汇总了 Demucs 和 FullSubNet 的消融实验结果。

5.4. 评估指标

确定两个模型的最佳超参数值后，我们进一步使用客观指标和声学参数对各自的模型进行评估。

表 1： *Demucs* 消融实验

	行业语音增强开启					
λ_1	1.0	1.0	1.0	0.8	0.75	0.5
λ_2	0.8	0.5	0.0	0.5	0.5	0.5
STOI	0.832	0.852	0.681	0.867	**0.882**	0.870
PESQ	1.658	1.844	1.488	1.780	**1.921**	1.863

5.4.1. 客观评估

除了 PESQ 和 STOI，我们还使用另外三种客观指标对这些模型进行测试：对数似然比（LLR）、相干性和语音可懂度指数（CSII），以及归一化协方差测度（NCM）。PESQ 和 LLR 用于测量语音质量（SQ），而 STOI、CSII 和 NCM 用于测量语音可懂度（SI）。这四种指标旨在捕捉人类对语音录音质量的判断，这被视为评估语音增强模型的黄金标准 [12]。这些指标的定义见附录 A.2。

5.4.2. 声学评估

除客观指标外，我们还采用 [13] 提出的声学参数集，具体为 eGeMAPSv02 功能描述符，以进一步评估模型。该参数集包含 88 个频率、能量/振幅和频谱相关参数。我们使用 OpenSMILE（基于大空间提取的开源语音与音乐解释）[3] Python 包来处理这些参数。

6. 结果

6.1. 声学改进

为了量化第 5.4 节中提到的声学参数的改进，我们按照 [14] 的定义测量声学改进。首先，我们计算时间轴上的平均绝对误差（平均绝对误差）

[3]https://github.com/audeering/opensmile-python

表 2： *FullSubNet* 消融实验

行业语音增强开启						
γ	1.00	0.30	0.10	0.03	0.01	0.00
STOI	**0.864**	0.863	0.860	0.861	0.858	0.861
PESQ	1.843	1.829	1.805	**1.867**	1.787	1.808
行业语音增强关闭						
γ	1.00	0.30	0.10	0.03	0.01	0.00
STOI	0.722	0.731	0.735	**0.740**	0.739	0.739
PESQ	1.427	1.437	1.470	1.572	**1.621**	1.617

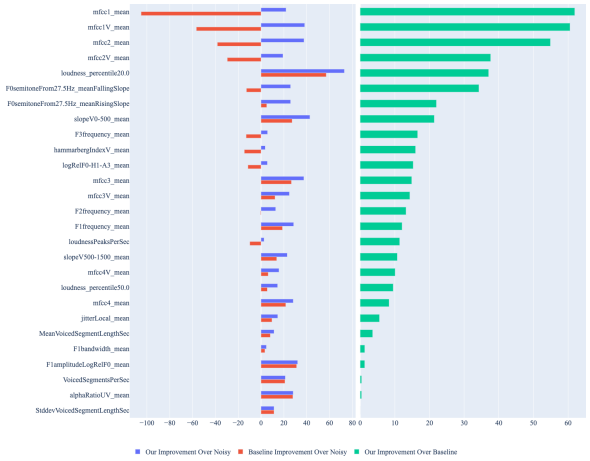

图4： 微调后的 *Demucs* 声学改进。左侧以红色表示 *Demucs* 相较于噪声信号的改进，以紫色表示微调后模型相较于噪声信号的改进。右侧展示了我们微调后的 *Demucs* 模型相较于基线模型的相对改进



图5： 微调后的 *FullSubNet* 模型的声学改进。左侧以红色表示 *FullSubNet* 相较于噪声信号的改进，以紫色表示微调后模型相较于噪声信号的改进。右侧展示了我们微调后的 *FullSubNet* 模型相较于基线模型的相对改进

表3： 结果表

	Demucs		FullSubNet	
	基线	微调后	基线	微调后
$PESQ$	1.727	**1.921**	1.694	**1.822**
LLR	1.432	**1.645**	1.630	**1.833**
$STOI$	0.864	**0.883**	0.841	**0.860**
$CSII_{high}$	0.665	**0.667**	0.652	**0.653**
$CSII_{mid}$	0.535	**0.572**	0.514	**0.539**
$CSII_{low}$	0.327	**0.333**	0.299	**0.334**
NCM	0.692	**0.756**	0.650	**0.675**

。新数据集中带噪音频文件与干净音频文件之间的平均绝对误差记为 MAE_N。基线增强音频（未经过微调的 FullSubNet 或 Demucs 的输出）与干净音频文件之间的平均绝对误差记为 MAE_B。最后，微调模型输出的增强音频与干净音频文件之间的平均绝对误差记为 MAE_F。随后，改进定义如下，其中 I_B 是基线模型的改进， I_F 是微调模型的改进。

$$I_B = 1 - \frac{MAE_B}{MAE_N} \tag{1}$$

$$I_F = 1 - \frac{MAE_F}{MAE_N} \tag{2}$$

图4和图5比较了我们微调后的 Demucs 和 FullSubNet 模型与基线模型在声学改进方面的表现。我们发现，微调模型在 14 个以上的声学参数上优于基线模型。

6.2. 客观指标结果

客观指标的结果如表3所示。我们发现，微调模型在所有指标上的表现均优于 Demucs 和 FullSubNet 基线模型。

7. 结论

在本研究中，我们超越了行业标准性能和最先进的基线架构，并指出传输噪声是当前语音增强研究中缺失的一个组成部分。我们在 VoIP DNS 挑战赛中取得了最佳表现，在 Google Meet 转电话赛道上同时降低了音频录音中的传输噪声和背景噪声。我们在新数据集上评估了 Demucs 和 FullSubNet 基线模型，为语音增强树立了新的基准。此外，我们还证明，在训练过程中引入 TAPLoss 并对这些模型进行微调，可以进一步提升性能。未来，我们计划将训练数据从 400 个样本增加到 1200 个样本，以进一步提升模型表现。我们相信，本研究成果可应用于电信行业，也可直接应用于手机制造商的产品中。

8. 致谢

我们谨向卡内基梅隆大学、Bhiksha Raj 教授，以及我们的导师 Joseph Konan 和 Ojas Bhar- gave 致谢，感谢他们在整个项目期间给予我们坚定的支持、鼓励和指导。

9. 参考文献

[1] B. N. Ilag，“Teams 音频会议和电话系统管理”，载于《*Understanding Microsoft Teams Administration*》。Springer，2020年，第263–339页。

[2] B. N. Ilag 和 A. M. Sabale，“Microsoft Teams 概述”，载于*Troubleshooting Microsoft Teams*. Springer，2022，第 17–74 页。

[3] J. Kaiser 和 T. Bořil，“GSM AMR 编解码器对自动化元音共振峰测量在 Praat 和 VoiceSauce 中的影响”，载于*2018*年第*41*届国际电信与信号处理会议（*TSP*）。IEEE，2018年，第1–4页。

[4] K. McDougall、F. Nolan 和 T. Hudson，“电话传输和听觉目击者：针对声音相似度控制的声音列队辨认表现”，*Phonetica*，第72卷第4期，第257–272页，2015年。[5] S. Lawrence、F. Nolan 和 K. McDougall，“电话传输对元音音质的声学和感知影响。”国际言语、语言与法律期刊，第15卷，2008年第2期。

[6] C. K. Reddy、V. Gopal、R. Cutler、E. Beyrami、R. Cheng、H. Dubey、S. Matusevych、R. Aichner、A. Aazami、S. Braun 等，“Interspeech 2020年深度噪声抑制挑战赛：数据集、主观测试框架和挑战赛结果”，*arXiv* 预印本 *arXiv:2005.13981*，2020年。

[7] A. Défossez, N. Usunier, L. Bottou 和 F. Bach，“波形域中的音乐源分离”，*arXiv* 预印本 *arXiv:1911.13254*，2019年。

[8] X. Hao、X. Su、R. Horaud 和 X. Li，“Fullsubnet：用于实时单通道语音增强的全频带与子频带融合模型”，载于*ICASSP 2021-2021 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021，页码 6633–6637。

[9] A. Defossez、G. Synnaeve 和 Y. Adi，“波形域中的实时语音增强”，*arXiv* 预印本 *arXiv:2006.12847*，2020年。

[10] A. Défossez, N. Zeghidour, 和 Bach, “SING：将乐器声音转换为歌声”，*arXiv* 预印本 *arXiv:1810.09281*，2018。

[11] D. Stoller, S. Ewert 和 S. Dixon，“Wave-u-net：一个多尺度神经网络，用于端到端音频源分离，”*arXiv* 预印本 *arXiv:1806.03185*，2018。

[12] C. K. Reddy, V. Gopal 和 R. Cutler，“Dnsmos：一种非侵入式感知客观语音质量指标，用于评估噪声抑制器”，载于 *ICASSP 2021-2021 IEEE* 声学、语音与信号处理国际会议 *(ICASSP)*。IEEE，2021，页 6493–6497。

[13] F. Eyben、K. R. Scherer、B. W. Schuller、J. Sundberg、E. Andr´e、C. Busso、L. Y. Devillers、J. Epps、P. Laukka、S. S. Narayanan 等，“用于语音研究和情感计算的日内瓦极简声学参数集（GeMAPS）”，*IEEE* 情感计算汇刊，第 7 卷，第 2 期，第 190–202 页，2015 年。[14] Zeng、Konan、Han、Bick、Yang、Kumar、Watanabe 和 Raj，“Ta-ploss：一种用于语音增强的时序声学参数损失”，未发表手稿，2022 年。[15] A. W. Rix、J. G. Beerends、M. P. Hollier 和 A. P. Hekstra，“感知语音质量评估（PESQ）——一种用于评估电话网络和编解码器语音质量的新方法”，载于*2001* 年 *IEEE* 国际声学、语音与信号处理会议论文集（目录号 *01CH37221*），第 2 卷。IEEE，2001 年，第 749–752 页。[16] R. Crochiere、J. Tribolet 和 L. Rabiner，“将对数似然比解释为波形编码器性能的度量”，*IEEE* 声学、语音与信号处理汇刊，第 28 卷，第 3 期，第 318–323 页，1980 年。[17] C. H. Taal、R. C. Hendriks、R. Heusdens 和 J. Jensen，“一种用于时频加权噪声语音的短时客观可懂度度量”，载于*2010* 年 *IEEE* 国际声学、语音与信号处理会议。IEEE，2010 年，第 4214–4217 页。

[18] J. M. Kates 和 K. H. Arehart，“相干性与语音可懂度指数”，美国声学学会杂志，第 117 卷，第 4 期，第 2224–2237 页，2005 年。[19] J. F. Santos、S. Cosentino、O. Hazrati、P. C. Loizou 和 T. H. Falk，“复杂聆听环境中人工耳蜗使用者的客观语音可懂度测量”，语音通信，第 55 卷，第 7–8 期，第 815–824 页，2013 年。

A. 附录

A.1. 计算 TAPLoss

为计算 TAPLoss，我们定义时域声学参数估计器 A_y。 $A_y(t,\ p)$ 表示离散时间帧 t 处的声学参数 p。为预测 A_y，我们定义该估计器：

$$\hat{A}_y = \mathcal{TAP}(y)$$

$\mathcal{TAP}$ 函数接收信号输入 y。它计算出一个包含 $F = 257$ 个频率箱的复数频谱图 Y，然后将该复数频谱图传递给循环神经网络，以输出时序声学参数 $\hat{} A_y$。TAP损失定义为实际估计值与预测估计值之间的平均绝对误差。

$$MAE(A_y, \hat{A}_y) = \frac{1}{TP}\sum_{t=0}^{T-1}\sum_{p=0}^{P-1}|A_y(t,p) - A_{\hat{y}}(t,p)| \in R$$

在训练过程中，$\mathcal{TAP}$ 参数通过 Adam优化学习，以最小化 $MAE(A_s,\ \hat{} A_s)$的差异。由于这种损失函数端到端可微，且仅以波形为输入，因此能够对任何具有干净参考语音的语音模型和任务进行声学优化。

A.2. 指标定义

表 4：评估指标

指标	描述
PESQ	反映人类对语音质量的主观评分 由网络条件导致的可懂度下降，包括—— 模拟连接和丢包 [15]
LLR	通过评估干净语音与增强语音之间的拟合程度来评估语音质量 [16]
STOI	衡量受损语音信号的可懂度 这些信号可能由加性噪声等因素导致 [17]
CSII	通过测量带宽降 低和加性噪声条件下的语音可懂度， 并分别测量各个频带中的信噪比。 计算范围分为高、中、低三个频段， 其中高频段测量不低于语音均方根 （RMS）分贝级的片段，中频段 测量范围为 (RMS-10, RMS], 及 范围低于 (RMS-30, RMS-10] [18]
NCM	根据以下协方差估算语音可懂度： 干净语音与含噪语音 信号的包络 [19]

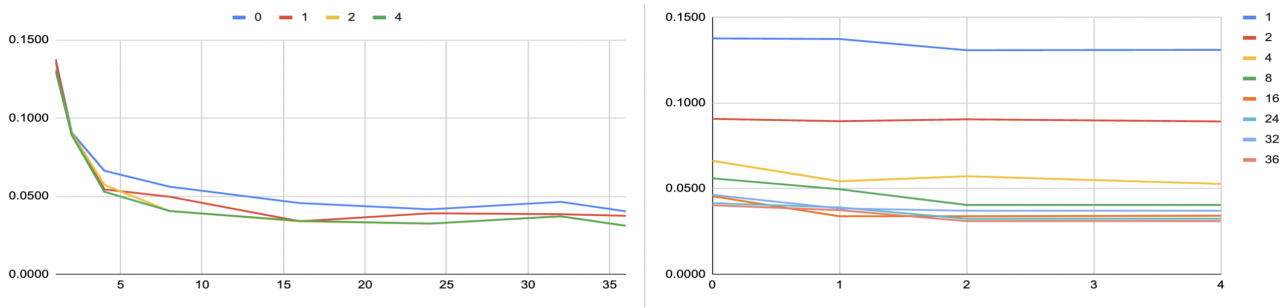

图 6：*Demucs*：左图显示每个工作进程在不同批次大小下所需的时间。右图显示不同批次大小下，各工作进程数对应的耗时。这有助于确定 *Demucs* 的最佳批次大小。

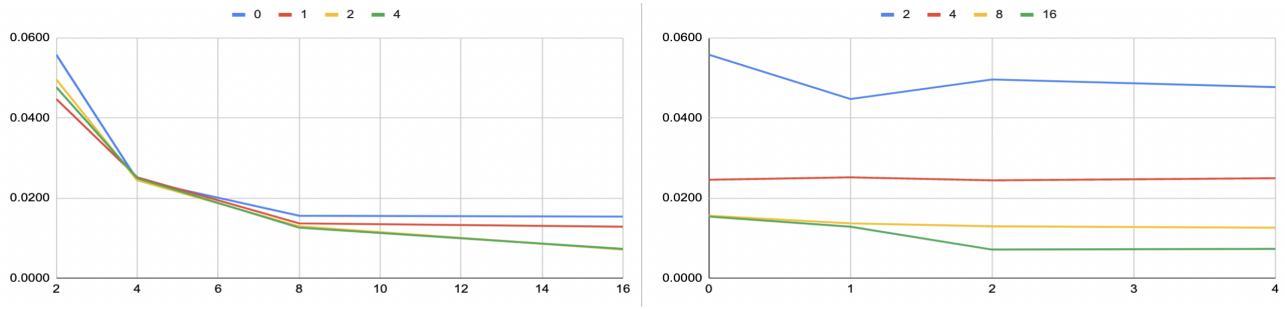

图 7：*FullSubNet*：左图显示每个工作进程在不同批次大小下所需的时间。右图显示不同批次大小下，各工作进程数对应的耗时。这有助于确定 *FullSubNet* 的最佳批次大小。

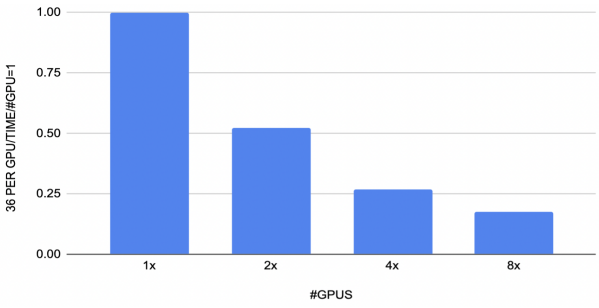

图 8：*DEMUCS* 的 *GPU* 扩展。该图展示了近乎完美的扩展性：计算能力翻倍，吞吐量也随之翻倍。不过，从 *4* 倍 *GPU* 增至 *8* 倍时，收益会递减。

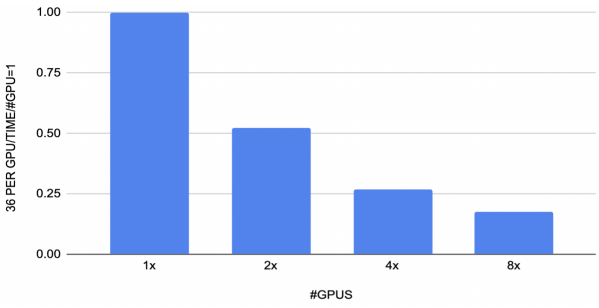

图 9：*FullSubNet* 的 *GPU* 扩展。该图同样展示了近乎完美的扩展性：计算能力翻倍，吞吐量也随之翻倍。与 *Demucs* 类似，从 *4* 倍 *GPU* 增至 *8* 倍时，收益会递减。

A.3. 最佳批次大小和工作进程数

图 6 和图 7 分别展示了我们为 DEMUCS 和 FullSubNet 训练寻找最优批次大小的实验。最终，我们为 DEMUCS 选择了 36 的批次大小和 2 个工作进程。对于 FullSubNet，我们选择了 16 的批次大小和 2 个工作进程。

A.4. GPU 扩展

确定批次大小和工作进程数后，我们将实验扩展到多个图形处理器上，以找到最适合训练的耗时，如图 8 和图 9 所示。