ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10448310

GTCRN：一种需要超低计算资源的语音增强模型

Xiaobin Rong[1,2], *Tianchi Sun*[1,2], *Xu Zhang*[3], *Yuxiang Hu*[2], *Changbao Zhu*[2], *Jing Lu*[1,2]

[1]南京大学现代声学教育部重点实验室，中国南京 210093

[2]南京大学地平线智能音频实验室，地平线机器人技术有限公司，中国北京 100094

[3]江苏星云信息技术有限公司，中国南京 210046

{xiaobin.rong, tianchi.sun}@smail.nju.edu.cn，zhangx@thingstar.cn，{yuxiang.hu, changbao.zhu}@horizon.cc，lujing@nju.edu.cn

摘要

尽管基于现代深度学习的模型在语音增强领域已显著超越传统方法，但它们通常需要大量的参数和强大的计算能力，这使得它们难以在实际应用中部署到边缘设备上。在本文中，我们引入了分组时域卷积循环网络（GTCRN），它采用分组策略对竞争模型 DPCRN 进行有效简化。此外，它还利用子带特征提取模块和时域循环注意力模块来提升性能。值得注意的是，所得到的模型对计算资源的需求极低，每秒仅需 23.7K 个参数和 39.6MMACs。实验结果表明，我们提出的模型不仅超越了 RNNoise 这种计算负担相似的典型轻量级模型，而且在与计算资源需求显著更高的近期基准模型的比较中也表现出色。

索引词 - 语音增强、轻量级模型、卷积循环网络

1. 简介

在语音增强（SE）领域已取得重大突破，这主要得益于深度神经网络（DNN）的快速发展。总体而言，基于 DNN 的 SE 算法可分为时频（T-F）域[1, 2, 3, 4]和时域[5, 6, 7]方法。基于 DNN 的方法在性能上往往远超传统 SE 算法，但通常伴随着较大的模型开销。大多数最先进的（SOTA）SE 模型需要大量的计算资源，从几个 GMAC 到几十个 GMAC 不等，这使得它们难以部署在边缘设备上用于实际应用。

一些近期的研究工作专注于探索轻量级的软件工程方法，这些方法在减少计算需求的同时，其性能可与最先进的模型相媲美。一种直接的解决办法是采用诸如剪枝和量化等技术对性能良好的模型进行压缩[8, 9]。另一类方法是高效模型设计，例如 TRU-Net [10]，它利用一维卷积将频率轴和时间轴上的计算解耦，并用深度卷积替代标准卷积操作。并行 GRU 和优化的跳跃连接[11]也可用于设计小型 SE 模型。第三类方法是将轻量级模型与适当的后处理相结合。在 RNNoise [12] 和 PercepNet [13] 中，先对低分辨率的频谱包络进行粗略增强，然后使用基音梳状滤波器在基音谐波之间执行更精细的噪声衰减。基于 PercepNet 的 DeepFilterNet [14] 首先采用类似 UNet 的更强大的 DNN 来增强频谱包络，并进一步利用深度滤波增强周期性成分。DPCRN-CF [15] 利用基于 DNN 的基音估计器和可学习的梳状滤波器来实现更出色的谐波增强。然而，尽管这些方法在计算开销方面取得了显著的降低，但对于诸如耳机和助听器这类对低功耗有严格要求的终端设备来说，它们的体积仍然过大，无法实际部署，只有 RNNoise 体积足够小，但性能有限。

在本文中，我们提出了分组时域卷积循环网络（GTCRN），这是一种计算资源需求极低的语音增强模型。以 DPCRN [3, 16] 为骨干网络，采用多种策略大幅压缩模型。使用等效矩形带宽（ERB）滤波器组来减少输入特征的冗余。采用分组卷积 [17] 和分组 RNN [18] 来降低模型复杂度。为了在不增加过多计算开销的情况下提升性能，我们进一步应用了子带特征提取（SFE）模块和时域循环注意力（TRA）模块。最终模型在 DNS3 和 VCTK-DEMAND 上的表现均显著优于 RNNoise。

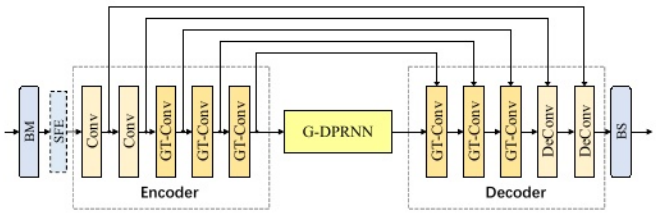

图 1：所提出的 GTCRN 模型的整体架构。

数据集。

2. 分组时间卷积循环网络

GTCRN 架构由频带合并（BM）和频带分割（BS）模块、一个可选的 SFE 模块、一个编码器、一个分组双路径 RNN（G-DPRNN）模块以及一个解码器组成，如图 1 所示。各模块的详细信息将在第 2.1 至 2.5 节中介绍。编码器由两个卷积（Conv）块和三个分组时间卷积（GT-Conv）块组成，将在第 2.3 节中讨论。每个 Conv 块由卷积层、批量归一化和 PReLU 激活函数组成，将输入频谱映射到高维嵌入，并降低频率轴的大小。采用跳跃连接来缓解编码阶段的信息损失。解码器是编码器的镜像版本，其中每个 Conv 块被替换为一个反卷积（DeConv）块，其组件与 Conv 块相同，只是将卷积层替换为转置卷积层以恢复原始大小。此外，最后一个 DeConv 块使用 tanh 而不是 PReLU 激活函数，以将输出值限制在 -1 到 1 之间。这些值被解释为估计的复比掩码（CRM）[19] 的实部和虚部。

2.1. 频段合并与拆分

我们可以通过 BM 操作对频谱特征进行下采样，并使用 BS 操作恢复原始分辨率。但需要注意的是，谐波更有可能出现在低频段，而在高频段则很少出现。因此，特征的合并仅在高于 2 千赫兹的高频段按照 ERB 规则进行。

2.2. 分组双路径 RNN

我们将分组循环神经网络（GRNN）[18]与双路径循环神经网络（DPRNN）[7]相结合，构建了 G-DPRNN。GRNN 利用一组较小的循环层来近似一个大的标准循环层。具体来说，输入特征和隐藏状态都被分成两个不相交的组，每组都输入到参数数量仅为原参数数量一半的循环层中，然后通过一个表示重组层来获得最终输出。DPRNN 最初被提出用于建模一维长序列，但它也

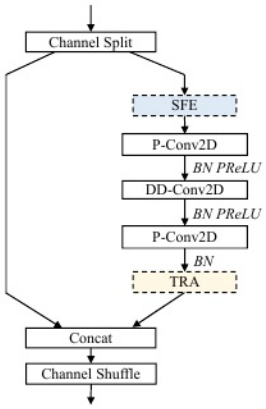

图 2：分组时间卷积块。

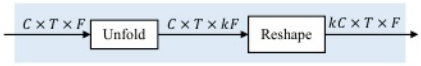

图 3：子带特征提取模块。

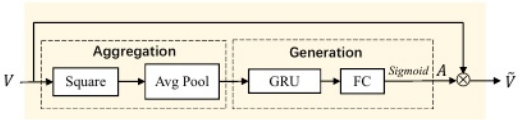

图 4：时间循环注意力模块。

非常适合用于时频域特征，如文献[3]中所述。帧内 RNN 可以对单帧中的频谱模式进行建模，而帧间 RNN 则对特定频率分量的时间相关性进行建模。我们使用分组双向 GRU 进行帧内建模，使用分组单向 GRU 进行帧间建模，从而保证模型的因果关系。

2.3. 分组时间卷积

GT-Conv 块以 ShuffleNetV2 [17] 单元为基础，在深度卷积中引入了时间膨胀，从而增强了其对长时序依赖关系的建模能力。GT-Conv 块的概览如图 2 所示。输入特征沿通道轴一分为二，形成两个分支。其中一个分支保持不变，而另一个分支则通过一系列卷积层进行高效的模式捕获和处理，这些卷积层由两个 2D 点卷积（P-Conv2D）层和一个 2D 膨胀深度卷积（DD-Conv2D）层组成。两个分支的输出最终被拼接起来以恢复原始大小。执行通道混洗操作以促进两个分支之间的信息交换。为了进一步提升模型性能，可选的 SFE 模块和 TRA 模块可以应用于第二个分支。

2.4. 子带特征提取

如图 3 所示，SFE 模块旨在增强卷积层捕获和

通过利用频率信息来实现这一目标。它首先在频率维度上对输入特征执行展开操作，内核大小为 k，将每个频率带与其相邻的 k - 1 个带组合形成子带单元。随后，应用重塑操作，沿通道维度堆叠每个子带单元，从而形成子带交织特征。在整个过程中，SFE 模块将原本仅存在于频率维度的子带关系整合到通道维度，使后续的卷积层能够更有效地利用频率信息。

2.5. 时间循环注意力

TRA 模块旨在通过利用乘法注意力掩码对时间特征进行重新校准，从而有效地沿时间轴建模能量分布。注意力掩码的生成分为两步：全局信息聚合和注意力生成，如图 4 所示。给定输入特征 V $\in R^{C\times T\times F}$，首先通过全局平均池化得到时间能量表示 Z $\in R^{C\times T}$，其公式为 $Z(c,t)=\sum_{f=1}^{F} V(c,t,f)$，其中 C、T、F 分别表示通道——

其中，分别表示时间轴和频率轴的长度。然后，对时间能量表示进行处理，先通过一个门控循环单元（GRU），再通过一个全连接（FC）层，其中 GRU 将输入通道数翻倍，而 FC 层则恢复到原始通道数。随后，应用一个 S 型激活函数生成一个 1D 注意力掩码，然后沿频率轴复制以生成一个 2D 的时频掩码 A $\in R^{C\times T\times F}$。最终输出表示为 $\hat{V}=V\otimes A$，其中 ⊗ 表示逐元素乘法运算。

2.6. 损失函数

我们的损失函数同时应用于波形域和频谱图域：

$$\begin{aligned}\mathcal{L}=&\alpha\mathcal{L}_{\text{SISNR}}(\tilde{s},s)+(1-\beta)\mathcal{L}_{\text{mag}}(\tilde{S},S)\\&+\beta\left(\mathcal{L}_{\text{real}}(\tilde{S},S)+\mathcal{L}_{\text{imag}}(\tilde{S},S)\right)\end{aligned}\tag{1}$$

其中，\（\tilde{s}\） 和 \（s\） 分别为增强后的和干净的波形。\（\tilde{S}\） 和 \（S\） 分别为增强后的和干净的频谱图。α 和 β 分别设置为 0.01 和 0.3。上述公式中的每一项计算方式如下：

$$\mathcal{L}_{\text{SISNR}}=-\log_{10}\left(\frac{\|s_t\|^2}{\|\tilde{s}-s_t\|^2}\right);\ s_t=\frac{\langle\tilde{s},s\rangle s}{\|s\|^2}\tag{2}$$

$$\mathcal{L}_{\text{mag}}(\tilde{S},S)=\text{MSE}(|\tilde{S}|^{0.3},|S|^{0.3})\tag{3}$$

$$\mathcal{L}_{\text{real}}(\tilde{S},S)=\text{MSE}(\tilde{S}_r/|\tilde{S}|^{0.7},S_r/|S|^{0.7})\tag{4}$$

$$\mathcal{L}_{\text{imag}}(\tilde{S},S)=\text{MSE}(\tilde{S}_i/|\tilde{S}|^{0.7},S_i/|S|^{0.7})\tag{5}$$

3. 实验

3.1. 数据集

我们使用两个数据集来评估所提出的模型。第一个是 VCTK-DEMAND 数据集 [20]，其中包含干净语音和预先混合的噪声语音的配对数据。训练集和测试集分别包含来自 28 位说话人的 11572 个语音片段和来自 2 位说话人的 824 个语音片段。训练集中选取了 1572 个语音片段用于验证。所有语音片段均重采样至 16 千赫兹。

第二个数据集是大规模的 DNS3 数据集 [21]，其中包含各种各样的干净语音集、噪声集和房间脉冲响应（RIR）。此外，我们还纳入了来自 DiDiSpeech [22] 的普通话语料库。在混合过程中，干净语音与随机选取的 RIR 进行卷积，然后与随机选取的噪声片段在信噪比（SNR）范围为 -5 至 15 分贝的条件下进行混合。训练目标是通过保留前 100 毫秒的反射来获取。总共生成了 72 万对 10 秒的噪声 - 清晰语音数据用于训练，同时分别生成了 840 对和 800 对用于验证和测试。评估也是在 DNS 挑战赛 3 提供的盲测集上进行的。所有语音均以 16 千赫兹的采样率进行采样。

3.2. 实施细节

短时傅里叶变换（STFT）采用长度为 32 毫秒的平方根汉宁窗，跳帧长度为 16 毫秒，快速傅里叶变换（FFT）长度为 512。输入特征是将含噪频谱的实部和虚部以及其幅度按通道级联而成。对于 BM，我们将 192 个高频频带映射到 64 个耳等效带宽（ERB）频带，同时保持 65 个低频频带不变，从而得到一个 129 维的压缩特征图。对于所有可选的 SFE 模块，我们统一使用 3×3 的卷积核。两个卷积块的输出通道数均为 16，卷积核大小为 （1, 5），步长为（1, 2）。第二个卷积层的分组大小设置为 2，以减少参数和计算量。三个 GT-Conv 块中的 DD-Conv2D 层共享相同的通道数 16 和相同的卷积核大小 （3, 3），时间膨胀率分别为 1、2 和 5。对于整个模型，参数数量为 23.7K，每秒的计算成本为 39.6MMACs。

模型通过 Adam 优化器 [23] 进行训练，初始学习率为 0.001。如果验证损失连续 5 个周期没有下降，学习率将减半。对于 VCTK-DEMAND 数据集，我们使用 4 的批量大小；对于 DNS3 数据集，批量大小为 16。在 DNS3 数据集的训练过程中，将话语切分为 8 秒长的片段，并且每个周期随机选取 40,000 对有噪声和干净的语音对。

3.3. 结果

3.3.1. 消融研究

我们验证了 SFE 的有效性，并在从该数据集中抽取的一个相对较小的训练集（约 100 小时）上，将我们的 TRA 与 [24] 中提出的时间维度注意力（TA）进行了比较。

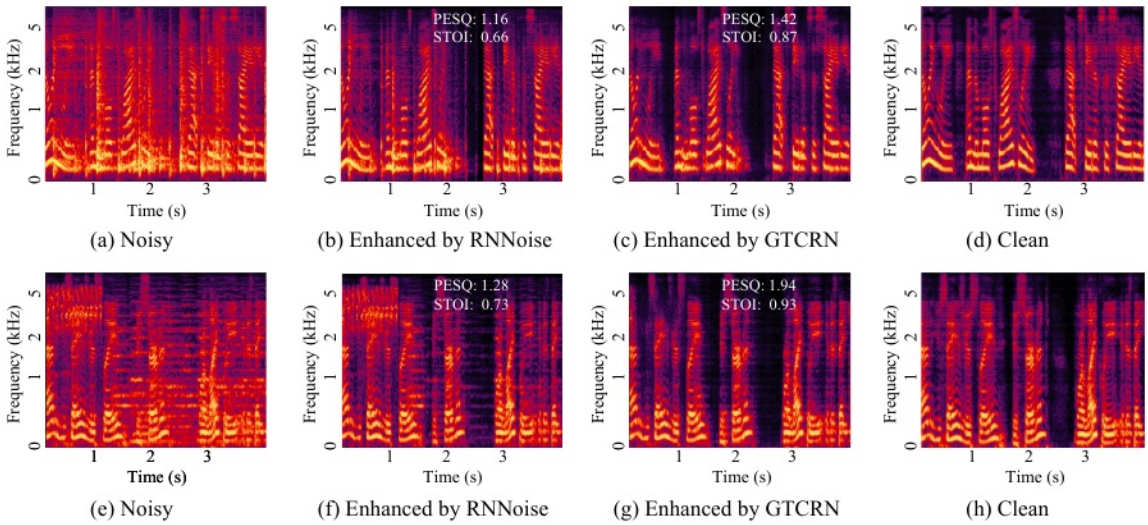

图 5：来自 DNS3 测试集的典型频谱图。（a、e）噪声语音，（b、f）RNNoise 增强后的语音，（c、g）GTCRN 增强后的语音，（d、h）干净的参考语音。

表 1：在 DNS3 测试集上的消融研究结果。

SFE	他/她/它TRA	第（K）段	每秒百万次操作（MACS）	SISNR	PESQ	STOI
-	-	-	-	3.92	1.30	0.789
☒	☒	13.35	33.91	9.87	1.87	0.834
☒	√	14.84	34.00	10.00	1.89	0.838
☒	☒	21.65	34.47	10.25	1.91	0.840
√	☒	15.37	39.07	10.10	1.90	0.838
√	√	16.86	39.16	10.29	1.92	0.841
√	☒	23.67	39.63	10.39	1.94	0.844

表 2：在 VCTK-DEMAND 测试集上的性能。

	段落（M）	每秒百万次操作（G）	SISNR	PESQ	STOI
嘈杂的	-	-	8.45	1.97	0.921
RNNoise（2018 年）	0.06	0.04	-	2.29	-
PercepNet（2020）	8.00	0.80	-	2.73	-
DeepFilterNet（2022 年）	1.80	0.35	16.63	2.81	0.942
S-DCCRN（2022）	2.34	-	-	2.84	0.940
GTCRN（提议的）	0.02	0.04	18.83	2.87	0.940

DNS3 数据集。评估是在测试集上使用客观评估指标进行的，包括 SISNR [25]、PESQ [26] 和 STOI [27]。消融测试结果见表 1。可以看出，我们提出的 TRA 在计算资源仅略有增加的情况下就优于 TA。表 1 中 SFE 的优势也很明显，通过将 SFE 与 TRA 集成，实现了最佳性能指标。

3.3.2. 与基准模型的比较

我们将我们的模型与 RNNoise [12]、PercepNet [13]、DeepFilterNet [14] 和 S-DCCRN [28] 进行了比较。表 2 展示了在 VCTK-DEMAND 测试集上的客观结果。显然，GTCRN 不仅在计算量相当且参数更少的情况下大幅优于 RNNoise，而且在 SISNR 和 PESQ 指标上也显著超越了其他参数和 MACs 多得多的基线模型。

在表 3 中，我们展示了我们的模型与其他模型的比较。

表 3：DNS3 盲测数据集上的性能。

	段落（M）	每秒百万次操作（G）	DNSMOS-P.808	DNSMOS-P.835		
				BAK	信号	OVRL
嘈杂的	-	-	2.96	2.65	3.20	2.33
RNNoise[1]（2018）	0.06	0.04	3.15	3.45	3.00	2.53
S-DCCRN（2022）	2.34	-	3.43	-	-	-
GTCRN（提议的）	0.02	0.04	3.44	3.90	3.00	2.70

在 DNS3 盲测集上对 RNNoise 和 S-DCCRN 进行评估。评估使用的是 DNSMOS P.808 [29] 和 DNS-MOS P.835 [30] 标准。结果始终表明，我们的模型大幅优于 RNNoise，并且也超过了大规模的 S-DCCRN 模型。图 5 展示了来自我们测试集的两个典型示例，清楚地表明 GTCRN 在噪声抑制方面优于 RNNoise。源代码和音频示例可在 https://github.com/Xiaobin-Rong/gtcrn 获取。

4. 结论

在本文中，我们提出了 GTCRN，这是一种仅需 23.7K 参数和每秒 39.6MMAC 的语音增强模型。我们对 DPCRN 应用了多种策略，有效地减少了模型规模，同时保持了语音增强性能。实验表明，我们的模型不仅在 VCTK-DEMAND 和 DNS3 数据集上大幅优于 RNNoise，而且与几个计算开销显著更高的基线模型相比，也取得了具有竞争力的性能。

5. 致谢

本研究得到了国家自然科学基金（项目编号：12274221）的支持。

[1] Metrics are measured with source code provided at https://github.com/xiph/rnnoise/

6. 参考文献

[1] K. Tan and D. Wang, "A Convolutional Recurrent Neu-ral Network for Real-Time Speech Enhancement," in *Proc. Interspeech 2018*, 2018, pp. 3229–3233.

[2] Y. Hu, Y. Liu, S. Lv, M. Xing, et al., "DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement," in *Interspeech*, 2020.

[3] X. Le, H. Chen, K. Chen, and J. Lu, "DPCRN: Dual-Path Convolution Recurrent Network for Single Chan-nel Speech Enhancement," in *Proc. Interspeech 2021*, 2021, pp. 2811–2815.

[4] Z.-Q. Wang, S. Cornell, S. Choi, Y. Lee, et al., "Tf gridnet: Integrating full-and sub-band modeling for speech separation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[5] D. Stoller, S. Ewert, and S. Dixon, "Wave-u-net: A multi-scale neural network for end-to-end audio source separation," in *Proc. International Society for Music Information Retrieval Conference*, 2018, pp. 334–340.

[6] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpass-ing ideal time–frequency magnitude masking for speech separation," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 27, no. 8, pp. 1256–1266, 2019.

[7] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation," in *ICASSP*, 2020, pp. 46–50.

[8] I. Fedorov, M. Stamenovic, C. R. Jensen, L.-C. Yang, et al., "TinyLSTMs: Efficient Neural Speech Enhance-ment for Hearing Aids," in *Interspeech*, 2020.

[9] K. Tan and D. Wang, "Towards model compression for deep learning based speech enhancement," *IEEE/ACM transactions on audio, speech, and language process-ing*, vol. 29, pp. 1785–1794, 2021.

[10] H.-S. Choi, S. Park, J. H. Lee, et al., "Real-time denoising and dereverberation wtih tiny recurrent u-net," in *ICASSP*, 2021, pp. 5789–5793.

[11] S. Braun, H. Gamper, C. K. Reddy, and I. Tashev, "Towards efficient models for real-time deep noise suppression," in *ICASSP*, 2021, pp. 656–660.

[12] J.-M. Valin, "A hybrid DSP/deep learning approach to real-time full-band speech enhancement," in *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*. IEEE, 2018, pp. 1–5.

[13] J.-M. Valin, U. Isik, N. Phansalkar, R. Giri, et al., "A Perceptually-Motivated Approach for Low-Complexity, Real-Time Enhancement of Fullband Speech," in *Proc. Interspeech 2020*, 2020, pp. 2482–2486.

[14] H. Schroter, A. N. Escalante-B, T. Rosenkranz, and A. Maier, "DeepFilterNet: A low complexity speech enhancement framework for full-band audio based on deep filtering," in *ICASSP*, 2022, pp. 7407–7411.

[15] X. Le, T. Lei, L. Chen, Y. Guo, et al., "Harmonic enhancement using learnable comb filter for light-weight full-band speech enhancement model," in *Proc. INTERSPEECH 2023*, 2023, pp. 3894–3898.

[16] X. Le, T. Lei, K. Chen, and J. Lu, "Inference skipping for more efficient real-time speech enhancement with parallel RNNs," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2411–2421, 2022.

[17] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 116–131.

[18] F. Gao, L. Wu, L. Zhao, T. Qin, X. Cheng, and T.-Y. Liu, "Efficient sequence learning with group recurrent networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2018, pp. 799–808.

[19] D. S. Williamson, Y. Wang, and D. Wang, "Com-plex ratio masking for monaural speech separation," *IEEE/ACM transactions on audio, speech, and lan-guage processing*, vol. 24, no. 3, pp. 483–492, 2015.

[20] C. Valentini-Botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech.," in *SSW*, 2016, pp. 146–152.

[21] C. K. A. Reddy, H. Dubey, K. Koishida, et al., "Interspeech 2021 Deep Noise Suppression Challenge," 2021.

[22] T. Guo, C. Wen, D. Jiang, N. Luo, et al., "Didispeech: A Large Scale Mandarin Speech Corpus," in *ICASSP*, 2021, pp. 6968–6972.

[23] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6980, 2014.

[24] Q. Zhang, Q. Song, et al., "Time-Frequency Attention for Monaural Speech Enhancement," in *ICASSP*, 2022, pp. 7852–7856.

[25] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "SDR–half-baked or well done?," in *ICASSP*, 2019, pp. 626–630.

[26] A. W. Rix, J. G. Beerends, et al., "Perceptual evalua-tion of speech quality (PESQ)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001, vol. 2, pp. 749–752.

[27] C. H. Taal, R. C. Hendriks, et al., "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *ICASSP*, 2010, pp. 4214–4217.

[28] S. Lv, Y. Fu, M. Xing, et al., "S-DCCRN: Super Wide Band DCCRN with Learnable Complex Feature for Speech Enhancement," in *ICASSP*, 2022, pp. 7767–7771.

[29] C. K. Reddy, V. Gopal, and R. Cutler, "DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in *ICASSP*, 2021, pp. 6493–6497.

[30] C. K. Reddy, V. Gopal, and R. Cutler, "Dnsmos P.835: A Non-Intrusive Perceptual Objective Speech Quality Metric to Evaluate Noise Suppressors," in *ICASSP*, 2022, pp. 886–890.