MEANFLOWSE：通过条件均值流实现一步生成式语音增强

Duojia Li*, Shenghui Lu†, Hongchen Pan*, Zongyi Zhan*, Qingyang Hong†*, Lin Li**

* School of Electronic Science and Engineering, Xiamen University, China
† School of Informatics, Xiamen University, China
liduojia@stu.xmu.edu.cn; qyhong@xmu.edu.cn; lilin@xmu.edu.cn

ABSTRACT

多步推理是实时生成性语音增强的瓶颈，因为基于流和扩散的系统学习瞬时速度场，因此依赖于迭代常微分方程（ODE）求解器。我们介绍了 MeanFlowSE，这是一种条件生成模型，它在一个轨迹上有限区间内的平均速度进行学习。使用雅可比-向量积（JVP）来实现 MeanFlow 恒等式，我们推导出一个局部训练目标，该目标直接监督有限区间的位移，同时保持与对角线上瞬时场约束的一致性。在推理过程中，MeanFlowSE 通过时间反向位移进行单步生成，消除了多步求解器的需求；可选的几步变体提供额外的优化。在 VoiceBank-DEMAND 上，单步模型实现了强大的可懂度、保真度和感知质量，并且计算成本远低于多步基线。该方法不需要知识蒸馏或外部教师，提供了实时生成性语音增强的有效高保真框架。

Index Terms— 语音增强，生成模型，流匹配，平均流，一步推理

1. 介绍

语音增强（SE）旨在从噪声信号中恢复清晰的语音，在通信系统和鲁棒的自动语音识别（ASR）中发挥着至关重要的作用 [1, 2] 。判别方法估计谱掩模或清晰语音特征 [3, 4] ，然而在恶劣条件下它们可能会产生过度平滑或失真的输出，从而降低感知质量和可懂度 [5] 。

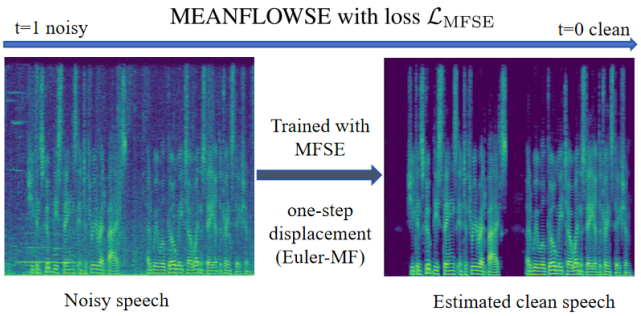

Fig. 1. 一步向后时间位移（欧拉–MF）。使用 MFSE 损失训练的模型将噪声频谱图在 t=1 映射到增强输出，通过沿着条件路径朝向 t=0 的单一位移实现。

生成模型通过学习清晰语音分布并反转噪声污染过程提供了另一种方法 [6, 7] 。在 STFT 域中，扩散或基于得分的方法已经取得了强大的性能，然而它们对大量函数评估（NFE）的依赖限制了其实时应用能力。几种努力旨在缓解这一问题：CDiffuSE 采用条件逆向采样以提高保真度，但仍依赖于长采样链 [8] ；SGMSE 使用预测-校正方法稳定复杂的频谱合成，但仍然保持高计算成本 [9] ；纠正逆过程（CRP）学习一个修正项以减少采样步骤，尽管还需要额外的微调 [10] ；薛定谔桥接（SB）形式将 SE 视为随机控制问题，但仍为多步且对离散化敏感 [11] 。在归一化流技术中，流动匹配（FM）提供了扩散的确定性替代方案 [12, 13, 14] 。条件流动匹配（CFM）进一步调节了沿预定路径的瞬时速度场 [13] 。FlowSE 实现了用于语音的 CFM，并通过线性-高斯条件路径实现了与扩散模型相当的性能，但步数更少；然而它仍然依赖于瞬时速度的迭代

*Corresponding authors: Lin Li and Qingyang Hong.

This work was supported in part by the National Natural Science Foundation of China under Grants 62371407 and 62276220.

数值积分，未能达到高效的一次推理 [15]。

在本文中，我们提出了 MeanFlowSE，一种生成式语音增强模型，该模型学习平均速度场，捕捉有限间隔的位移而不是瞬时斜率。我们的方法通过雅可比 – 向量积（JVP）目标利用了身份 [16]，特别适应于条件性语音增强。此目标监督了有限间隔的平均场，并自然地在对角线上简化为标准 CFM（$r{=}t$）。在推理过程中，学习到的位移替代迭代数值常微分方程（ODE）积分，允许在一个步骤生成并在一个统一框架下进行几步修正（图 1）。在 VoiceBank-DEMAND 上评估时，MeanFlowSE 达到了与强流和扩散基线相当或更优的性能，同时以显著更低的实时因子（RTF）运行。该模型从零开始训练，不使用知识蒸馏 [17]，并保持与其他领域推理加速技术如修正流和一致性模型的兼容性。[18, 19]

2. 相关工作

2.1. 流动 SE

FlowSE [15] 通过在指定的线性高斯路径上进行条件流匹配来学习确定性的瞬时速度场。令 x_1 表示清晰的语音信号，y 表示相应的噪声观测。一个时间变量 $t \in [0,1]$ 参数化了从 y 到 x_1 的插值，通过线性调度：

$$\mu_t(x_1, y) = t\, x_1 + (1-t)\, y, \tag{1}$$

$$\sigma_t = (1-t)\, \sigma, \tag{2}$$

其中 $\sigma > 0$ 控制噪声水平。沿着这条路径，对状态关于 t 进行微分并消去辅助变量得到闭式路径上瞬时速度目标：

$$v_t(x_t \mid x_1, y) = \frac{\sigma_t'}{\sigma_t}\big(x_t - \mu_t\big) + \mu_t' = \frac{x_1 - x_t}{1-t}, \tag{3}$$

其中包含 $\mu_t' = x_1 - y$ 和 $\sigma_t' = -\sigma$。条件流匹配损失回归一个网络 $v_\theta(x_t, t, y)$ 到目标：

$$\mathcal{L}_{\mathrm{CFM}} = \mathbb{E}_{(x_1,y),\,t}\Big[\big\|\, v_\theta(x_t,t,y) - v_t(x_t \mid x_1, y)\,\big\|_2^2\Big], \tag{4}$$

其中 x_t 位于由方程定义的路径上。(1) 和 (2) 和 t 在 $[0, 1-\delta]$ 上均匀采样，对于一个小的 $\delta > 0$，以避免奇异极限 $t \to 1$。推理从噪声侧的先验 $x_0 \sim \mathcal{N}(y, \sigma^2 I)$ 开始，并通过显式欧拉更新在一个网格 $0 = t_0 < \cdots < t_N = 1$ 上进行：

$$x_{t_i} = x_{t_{i-1}} + v_\theta\big(x_{t_{i-1}}, t_{i-1}, y\big)\, \Delta t(i), \tag{5}$$

其中步长定义为 $\Delta t(i) = t_i - t_{i-1}$。该规则实现了沿插值路径对学习到的向量场的正向积分。

2.2. 平均流场

平均流 [16] 引入了平均速度场来描述有限区间运动，与传统流匹配中使用的瞬时速度场相对。设动力学遵循边缘瞬时场 $v(z_t, t)$，即 $z_t' = v(z_t, t)$。其中 $z_t \in \mathbb{C}^D$ 表示时间 $t \in [0,1]$ 的插值路径，而 $(\cdot)'$ 是关于 t 的导数。对于任意 $r < t$，平均速度定义为：

$$u(z_t, r, t) = \frac{1}{t-r} \int_r^t v(z_\tau, \tau)\, d\tau, \tag{6}$$

其中位移除以经过的时间给出了定义 $u(z_t, t, t) = v(z_t, t)$。对 $(t-r)u$ 关于 t 求导得到的平均流身份为：

$$u(z_t, r, t) = v(z_t, t) - (t-r)\, \frac{d}{dt} u(z_t, r, t), \tag{7}$$

其中，全导数通过链式法则展开 [16] 为：

$$\frac{d}{dt} u(z_t, r, t) = v(z_t, t)\, \partial_z u + \partial_t u. \tag{8}$$

此恒等式产生了一个可计算的局部监督规则。通过网络 u_θ 对平均场进行参数化，并将 r 固定在 (z_t, t) 处，形成回归目标：

$$u_{\mathrm{tgt}} = v(z_t, t) - (t-r)\Big(v(z_t, t)\, \partial_z u_\theta + \partial_t u_\theta\Big), \tag{9}$$

以及对目标使用 stop-gradientsg$(\cdot)$ 的损失：

$$\mathcal{L} = \mathbb{E}\Big[\,\|\, u_\theta(z_t, r, t) - \mathrm{sg}(u_{\mathrm{tgt}})\,\|_2^2\Big]. \tag{10}$$

一旦学习了 u_θ，采样将用位移规则替换常微分方程的积分：

$$z_r = z_t - (t-r)\, u_\theta(z_t, r, t), \tag{11}$$

均值流提供了训练目标和基于位移的采样规则（方程 (11)）[16]。这种表述将状态从 t 单次更新到更早的 r，并在小步极限下恢复欧拉积分。

3. 方法

受平均流原理的启发，我们提出学习一个用于条件语音增强的平均速度场。所提出的方法在复数短时傅里叶变换域中操作，并使用配对样本 (x_1, y)，其中 x_1 表示干净的语音信号而 y 表示相应的噪声观测。为了确保训练和推理之间的一致性，我们采用双线性-高斯条件路径：

$$\mu_t = (1-t)\, x_1 + t\, y, \tag{12}$$

$$\sigma_t = (1-t)\, \sigma_{\min} + t\, \sigma_{\max}, \tag{13}$$

其中 $t \in [0,1]$，使得 $t = 0$ 是干净的端点 ($\mu_0 = x_1$，$\sigma_0 = \sigma_{\min}$) 和 $t = 1$ 是噪声端点 ($\mu_1 = y$，$\sigma_1 = \sigma_{\max}$)。这种双重参数化逆转了 FlowSE 的端点约定。训练点通过以下方式绘制在路径上：$x_t = \mu_t + \sigma_t\, z,$，其中 $z \sim \mathcal{N}(0, I)$。

沿路径的微分给出路径上的瞬时目标：

$$v_t(x_t \mid x_1, y) = \mu_t' + \sigma_t' z = \frac{\sigma_t'}{\sigma_t}\,(x_t - \mu_t) + \mu_t', \tag{14}$$

其中 $\mu_t' = y - x_1$ 和 $\sigma_t' = \sigma_{\max} - \sigma_{\min}$。该目标仅在路径上的采样点使用。

3.1. 平均速度和 MeanFlowSE 恒等式

迭代地整合局部斜率会在曲线轨迹上累积误差。相反，我们估计有限区间平均速度，定义为在两个时间点之间产生净位移的恒定速率。

令 $v(x, t \mid y)$ 表示控制条件动力学 $x_t' = v(x_t, t \mid y)$ 的边缘瞬时场。对于任意 $r < t$，定义平均速度为再现 $[r, t]$ 上位移的平均斜率：

$$u(x_t, r, t \mid y) = \frac{1}{t-r} \int_r^t v(x_\tau, \tau \mid y)\, d\tau, \tag{15}$$

因此 $u(x_t, t, t \mid y) = v(x_t, t \mid y)$。对 $(t-r)u$ 关于 t 进行微分得到平均流 SE 恒等式：

$$u(x_t, r, t \mid y) = v(x_t, t \mid y) - (t-r)\frac{d}{dt} u(x_t, r, t \mid y), \tag{16}$$

其中全导数沿着条件轨迹进行：

$$\frac{d}{dt} u(x_t, r, t \mid y) = v(x_t, t \mid y) \cdot \nabla_x u + \partial_t u. \tag{17}$$

方程 (15) 中的不可解路径积分在方程 (16)-(17) 中被计算于 (x_t, t) 处的局部项所替代。其中对角情况 $r = t$ 简化为 $u = v$。

3.2. 平均流场损失

我们方法的核心思想是训练一个单独的网络 $u_\theta(x, r, t, y)$ 使其满足在等式 (16)-(17) 中给出的身份关系在从等式 (12)-(13) 中的路径抽取的训练样本 (x_t, r, t) 上。通过将方程 (14) 中的闭式目标 v_t 代入方程 (16)，并利用方程 (17) 展开全导数，同时保持 (y, r) 不变，我们得到一个一阶目标训练目标：

$$u_{\text{tgt}} = v_t - c\,(t-r)[v_t \cdot \nabla_x u_\theta + \partial_t u_\theta], \tag{18}$$

方程 (18) 中的首个目标与当 $c = 1$ 时的平均流恒等式一致；我们采用 $c = 0.5$ 作为稳定的一阶修正以增强稳定性而不改变不动点。受到原始平均流框架的启发，对目标应用了停止梯度操作以防止通过 JVP 进行更高阶的反向传播。

均值流 SE 损失函数通过停止梯度来训练网络以逼近此目标，从而避免高阶反向传播并保持良好的固定点：

$$\mathcal{L}_{\text{MFSE}} = \mathbb{E}\Big[\|u_\theta(x_t, r, t, y) - \text{sg}(u_{\text{tgt}})\|_2^2\Big]. \tag{19}$$

在边界 $r = t$ 处，方程 (18) 中的目标简化为 $u_{\text{tgt}} = v_t$，导致方程 (19) 完全与对角线上的条件流匹配目标一致。在实际操作中我们在训练期间加入一小部分边界样本，导数使用自动微分计算，并辅以数值稳定的中心有限差分方法作为后备；只有 (x, t) 被求导，而 (y, r) 则被视为常量。

3.3. 一步推理

由于网络学习了有限区间位移场，推理不再需要整合瞬时速度。我们应用反向欧拉更新，直接将噪声端点传输到增强估计。训练完成后，数值常微分方程积分被一个由学习到的场驱动的位移所替代：

$$x_{t_{k+1}} = x_{t_k} - \Delta_k\, u_\theta(x_{t_k},\, r = t_{k+1},\, t = t_k \mid y), \tag{20}$$

其中 k 表示时间步长索引 $(k = 0, \ldots, N-1)$，$\Delta_k = t_k - t_{k+1} > 0$，和 $\{t_k\}_{k=0}^N$ 是一个单调递减的网格 $T_{\text{rev}} = t_0 > t_1 > \cdots > t_N = t_\varepsilon$。

逆时间初始化使用噪声边界的边缘分布 $x_{T_{\text{rev}}} \sim \mathcal{N}\big(y, \sigma^2(T_{\text{rev}})I\big)$，其中 $T_{\text{rev}} \in (0,1]$ 是逆时间起点，$t_\varepsilon \in [0,1)$ 是用于数值稳定性的终端时间。

一个单步推理规则如下：

$$\hat{x}_{t_\varepsilon} \;=\; x_{T_{\text{rev}}} - (T_{\text{rev}} - t_\varepsilon)\, u_\theta(x_{T_{\text{rev}}}, r = t_\varepsilon, t = T_{\text{rev}} \mid y)\,. \tag{21}$$

4. 实验

我们在 VoiceBank – DEMAND (VB-DMD) 数据集上以 16 kHz 的采样率进行评估，使用标准的训练/验证/测试划分，其中验证说话人被排除在外，且测试说话人和信噪比 [20, 21] 均未见过。所有系统均在 STFT 域中运行，采用 Hann 窗口和居中文本框；波形通过噪声信号进行峰值归一化，并将复数谱映射到 $|z|^{0.5}\exp(j\angle z)$ 并按 0.15 进行缩放。

增强网络使用带有自注意力的 NCSN++ U-Net [22, 23, 24]。模型输入通过将 (x_t, y) 沿通道连接形成；时间条件使用高斯傅里叶嵌入主要时间 t 和跨度 $\Delta = t - r$ 实现，头部预测单个复向量场。训练目标结合了两个组成部分：一个即时分支，在 $\Delta = 0$ 时匹配封闭形式的目标；一个均流分支，使用一阶修正拟合有限区间目标。导数通过自动微分计算，并以中心有限差分作为备选；应用停止梯度，并采用每样本 ℓ_2 裁剪来稳定雅可比项。优化采用 Adam 在 10^{-4} [25]，指数移动平均 0.999，以及梯度裁剪 1.0；我们在四个 V100 设备上使用分布式数据并行进行训练，并在验证时使用 EMA 权重。一个简短的课程将均值分支权重预热到 0.25，使跨度采样指数从 8 退火到 1，并在百分之十的批次中注入 $r = t$。

比较包括 CDiffuSE [8]，SGMSE [9]，反向过程校正 [10]，薛定谔桥 [11] 和 FlowSE [15]。我们遵循作者的设置，报告函数评估次数作为每个语音片段的前向传递计数。我们的模型默认在单次评估设置下报告；所有系统共享相同的前端和归一化。度量标准包括 ESTOI [26]，SI-SDR [27]，DNSMOS P.835 (SIG/BAK/OVRL) [28]，以及说话人相似度 (SpkSim) [29]。实时因子 (RTF) 是从端到端测量的，包括 STFT 和逆 STFT，在单个 V100 上以相同的精度和 I/O 设置下，批处理大小为一。

5. 结果

表 1 在相同的前端和归一化条件下展示了 VB-DMD 的系统级比较。仅通过一次函数评估，MeanFlowSE 就达到了最高的整体质量，ESTOI 值为 0.881，SI SDR 值为 19.975 dB。它还获得了最佳的 BAK 值 4.073，以及具有竞争力的 OVRL 分数 3.207 和说话人相似度 0.892。值得注意的是，在比较中，MeanFlowSE 实现了最低的 RTF 值 0.11。相比之下，基于扩散、桥梁和流的先前系统需要 5 到 200 个评估步骤，在相同的测量协议下产生了从 0.23 到 6.94 的 RTF 范围。表 2 在相同的前端和归一化条件下展示了 NFE 研究。相比之下，MeanFlowSE 仍然在单次函数评估中实现了最高的 ESTOI 值、SI SDR 值、说话人相似度和最低的实际时间因子，将采样计划简化为单一更新。

这些结果强调了基础建模策略的影响。SGMSE、StoRM、Schrödinger' s Bridge、CDiffuSE 和 FlowSE 等方法将瞬时场与多步 ODE 求解器结合在一起，质量的提升是以推理时间增加为代价的。相比之下，MeanFlowSE 学习一个平均速度场，直接预测有限区间位移，并在推理过程中允许一次逆向更新。通过将生成轨迹压缩到仅一步，MeanFlowSE 在保持高信号保真度和可理解性的同时，提供最佳的背景噪声抑制和最低的计算成本。总体而言，结果表明，平均速度学习推进了无蒸馏或外部教师情况下的生成语音增强的质量与效率前沿。

6. 结论

我们提出了 MeanFlowSE，这是一种用于语音增强的单步生成框架，其学习一个平均速度场。通过在一个条件路径上实例化 MeanFlow 恒等式，我们推导出一个可追踪的训练目标，该目标的最大值保证了有限区间位移，从而使得基于位移的推理可以替代数值常微分方程积分。实验结果表明，该方法在远低于多步基线计算成本的情况下实现了具有竞争力的表现。一个限制是假设了一条线性-高斯路径并使用一阶导数估计；未来的工作将研究更灵活或数据驱动的路径、更高阶的校正以及在实际条件下的评估。

Table 1. 与最先进的语音增强系统在 VOICEBANK – DEMAND 数据集上的性能比较。

系统	NFE	DNSMOS			估计标准误差 ↑	SI_SDR ↑	SpkSim ↑	准备就绪 ↓
		信号 ↑	备份 ↑	总体 ↑				
Noisy	N/A	3.346	3.126	2.697	0.787	8.445	0.888	N/A
SGMSE	30	3.488	3.985	3.176	0.863	17.396	0.891	1.81
FlowSE	5	3.478	4.051	3.202	0.873	19.145	0.889	0.23
Schrödinger's Bridge	30	3.486	4.062	3.216	0.872	19.448	0.886	1.07
CDiffuSE	200	3.434	3.727	2.994	0.798	13.665	0.812	6.94
StoRM	50	3.487	4.031	3.204	0.868	18.518	0.891	2.61
MeanFlowSE (ours)	1	3.471	4.073	3.207	0.881	19.975	0.892	0.11

注意：在本文中，粗体表示每列中的最佳分数，下划线表示第二佳分数。

Table 2. 质量-效率对比 NFE：FLOWSE 与 MEANFLOWSE

系统	NFE	估计标准误差 ↑	SI_SDR ↑	SpkSim ↑	准备就绪 ↓
FlowSE	5	0.873	19.145	0.889	0.23
FlowSE	10	0.870	18.428	0.891	0.38
FlowSE	20	0.868	18.099	0.890	0.71
MeanFlowSE (Ours)	1	0.881	19.975	0.892	0.11

7. REFERENCES

[1] P. C. Loizou, Speech Enhancement: Theory and Practice, CRC Press, Boca Raton, FL, 2007.

[2] Y. Ephraim and D. Malah, "Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 32, no. 6, pp. 1109–1121, Dec. 1984.

[3] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in Proc. Interspeech, Virtual Only Conference, 2020.

[4] Y. Luo and N. Mesgarani, "Conv-tasnet: Surpassing ideal time-frequency magnitude masking for speech separation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 8, pp. 1256–1266, Aug. 2019.

[5] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in Proc. Interspeech, Stockholm, Sweden, 2017, pp. 1–5.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in 34th Conference on Neural Information Processing Systems (NeurIPS 2020), June 2020, pp. 6840–6851.

[7] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in International Conference on Learning Representations (ICLR), Virtual Only Conference, 2021.

[8] Y. J. Lu, Z. Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 7402–7406.

[9] S. Welker, J. Richter, and T. Gerkmann, "Speech enhancement with score-based generative models in the complex STFT domain," July 2022, arXiv:2203.17004.

[10] B. Lay, J. M. Lermercier, J. Richter, and T. Gerkmann, "Single and few-step diffusion for generative speech enhancement," in ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2024, pp. 626–630.

[11] A. Jukić, R. Korostik, J. Balam, and B. Ginsburg, "Schrödinger bridge for generative speech enhancement," July 2024, arXiv:2407.16074.

[12] D. Rezende and S. Mohamed, "Variational inference with normalizing flows," in Proceedings of the 32nd International Conference on Machine Learning (ICML), July 2015, pp. 1530–1538.

[13] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le, "Flow matching for generative modeling," in The Eleventh International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.

[14] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud, "Neural ordinary differential equations," in 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Dec. 2018, pp. 1–13.

[15] S. Lee, S. Cheong, S. Han, and J. W. Shin, "Flowse: Flow matching-based speech enhancement," in ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Apr. 2025, pp. 1–5.

[16] Z. Geng, M. Deng, X. Bai, J. Z. Kolter, and K. He, "Mean flows for one-step generative modeling," May 2025, arXiv:2505.13447.

[17] T. Salimans and J. Ho, "Progressive distillation for fast sampling of diffusion models," in International Conference on Learning Representations (ICLR), Virtual Only Conference, 2022.

[18] X. Liu, C. Gong, and Q. Liu, "Flow straight and fast: Learning to generate and transfer data with rectified flow," in International Conference on Learning Representations (ICLR), Kigali, Rwanda, 2023.

[19] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever, "Consistency models," in Proceedings of the 40th International Conference on Machine Learning (ICML), July 2023, pp. 32211–32252.

[20] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large

regional accent speech database," in 2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE), Nov. 2013, pp. 1–4.

[21] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database (demand): A database of multichannel environmental noise recordings," Proceedings of Meetings on Acoustics, vol. 19, no. 1, pp. 035081, May 2013.

[22] Y. Song and S. Ermon, "Improved techniques for training score-based generative models," in 34th Conference on Neural Information Processing Systems (NeurIPS 2020), Dec. 2020, pp. 12438–12448.

[23] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Oct. 2015, pp. 234–241.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in 31st Conference on Neural Information Processing Systems (NIPS 2017), Dec. 2017, pp. 1–11.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in International Conference on Learning Representations (ICLR), San Diego, CA, USA, 2014.

[26] J. Jensen and C. H. Taal, "An algorithm for predicting the intelligibility of speech masked by modulated noise maskers," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 11, pp. 2009–2022, Nov. 2016.

[27] J. Le Roux, S. Wisdom, H. Erdogan, and J. R. Hershey, "Sdr – half-baked or well done?," in ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, pp. 626–630.

[28] C. K. A. Reddy, V. Gopal, and R. Cutler, "DNSMOS p.835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors," in ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2022, pp. 886–890.

[29] H. Wang, C. Liang, S. Wang, Z. Chen, B. Zhang, X. Xu, Y. Deng, and Y. Qian, "Wespeaker: A research and production oriented speaker embedding learning toolkit," in ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2023, pp. 1–5.