基于扩散的无监督音视频语音分离在有噪声先验的嘈杂环境中

Yochai Yemini[1], Rami Ben-Ari[2], Sharon Gannot[1] and Ethan Fetaya[1]

[1]Faculty of Engineering, Bar-Ilan University, Ramat Gan, Israel
[2]OriginAI
yochai.yemini@biu.ac.il

ABSTRACT

在本文中，我们解决了存在环境噪声情况下的单麦克风语音分离问题。我们提出了一种生成性无监督技术，该技术直接建模清晰的语音和结构化的噪声成分，在训练过程中仅依赖于这些单独的信号而非嘈杂的混合信号。我们的方法利用了一个结合视觉线索的音视频评分模型，作为强大的生成式语音先验。通过明确地对噪声分布与语音分布进行建模，我们能够借助逆问题范式实现有效的分解。我们通过反向扩散过程从后验分布中采样来进行语音分离，直接估计并移除建模的噪声成分以恢复清晰的信号成分。实验结果展示了令人鼓舞的表现，突显了我们在具有挑战性的声学环境中直接噪声建模方法的有效性。

Index Terms— 生成模型，逆问题，视听语音先验

1. 介绍

语音处理中的一个基本挑战是从嘈杂环境中恢复清晰的语音记录。在过去十年中，这一领域主要由神经网络技术主导。大多数方法以监督方式进行噪声去除，即训练神经网络从其噪声版本预测出清晰的语音。尽管监督方法在其训练任务上取得了出色的结果，但它的主要缺点是缺乏灵活性。当遇到新的声学条件时，监督方法必须在新场景下重新训练以避免性能下降。

相比之下，无监督语音处理采取了不同的方法。特别是，无监督生成范式训练一个先验语音模型，该模型学习清洁语音信号的数据分布。域知识被用于从其噪声观测中估计目标语音的先验模型。这种方法的主要优点是它的灵活性。由于不需要为每个退化设置训练任何专用模型，可以使用任何语音或噪声模型以及任意数量的说话人。这种优点使得在统一框架内处理各种语音处理问题成为可能，例如语音增强和说话人分离。

我们专注于使用扩散模型 [1, 2] 作为语音先验的无监督算法，通过逆问题框架从受损录音中估计清晰语音。在逆问题公式化中，清晰语音是根据受损语音条件下的后验语音分布进行采样的。该技术已成功应用于各种语音处理任务，例如 [3–6]。

然而，将反问题范式应用于环境噪声抑制的适用性尚未得到充分探索。一种最近提出的语音增强方法 [7, 8] 使用non-negative matrix factorization (NMF) 建模噪声协方差矩阵。然后执行expectation-maximisation (EM) 次迭代，在从后验分布中采样干净信号和估计噪声的 NMF 矩阵之间交替进行。

噪声信号的频谱内容差异很大，呈现出复杂的分布。我们假设 NMF 噪声模型的表现力可能有限，导致性能不佳。因此，为了更好地捕捉噪声分布，我们建议部署一个单独的扩散模型作为噪声先验。对于语音先验，我们学习了一个由视觉线索增强的强大音频扩散模型。在这项研究中，我们利用diffusion posterior sampling (DPS) 框架 [9] 来展示我们的方法在存在环境噪声的情况下单麦克风语音分离这一具有挑战性的任务上的有效性。我们的 DPS 形式化将噪声信号视为一个额外的源，并与语音源一起进行估计。

我们将该方法命名为 DAVSS-NM，意为基于扩散的含噪声建模音频视觉语音分离。实验结果表明，DAVSS-NM 显著缩小了逆问题技术与最先进的预测基线之间的性能差距。因此，作为一种无监督技术，DAVSS-NM 在灵活性和性能之间提供了一个很好的权衡。

2. 问题表述

设 $\boldsymbol{y}$ 为一组混合信号，包括 K 个干净的语音信号 $\{\boldsymbol{x}_i\}_{i=1}^K \in \mathbb{R}^d$ 和噪声信号 $\boldsymbol{n}, \boldsymbol{z} \in \mathbb{R}^d$，使用单个麦克风录制：

$$\boldsymbol{y} = \sum_{i=1}^{K} \boldsymbol{x}_i + \boldsymbol{n} + \boldsymbol{z}. \tag{1}$$

所有信号均以时域表示，并且统计独立。当 $\boldsymbol{n}$ 是来自结构化噪声分布的环境噪声时，$\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{z}}^2 \boldsymbol{I})$ 是一种人为添加的具有任意低方差的噪声。$\boldsymbol{z}$ 被添加以确保整个空间的概率非零，以便进行数学处理。在整个论文中，我们将使用所有语音源的连接表示 $\boldsymbol{x}_{1:K} \in \mathbb{R}^{Kd}$，而不是 $\{\boldsymbol{x}_i\}_{i=1}^K$。

3. 背景

3.1. 基于分数的扩散模型

扩散生成模型旨在通过最初从已知噪声分布中采样，然后迭代去噪样本，从而从数据分布 $p_{\text{data}}(\boldsymbol{x})$ 中采样。在去噪过程结束时，获得来自 $p_{data}(\boldsymbol{x})$ 的样本。去噪过程可以被公式化为一个ordinary differential equation (ODE) [2] ，该过程结合了中间扩散状态的得分函数。Karras 等人。[10] 建议了以下方差爆炸 ODE:

$$d\boldsymbol{x}^\tau = -\dot{\sigma}(\tau)\sigma(\tau)\nabla_{\boldsymbol{x}^\tau} \log p_\tau(\boldsymbol{x}^\tau) d\tau \tag{2}$$

其中 $p_\tau(\boldsymbol{x}^\tau|\boldsymbol{x}^0) = \mathcal{N}(\boldsymbol{x}^0, \sigma^2(\tau)\boldsymbol{I})$,$\tau$ 是扩散步骤，点表示关于 τ 的导数。ODE 将来自 $\mathcal{N}(\mathbf{0}, \sigma^2(T_{\max})\boldsymbol{I})$ 的噪声样本传输到来自 $p_{\text{data}}(\boldsymbol{x})$ 的样本 $\boldsymbol{x}^0$。选择 $\sigma(\tau)$ 的值使得 $\sigma(\tau)$ 单调递增。在 [10] 中选择了 $\sigma(\tau) = \tau$，我们在推导中遵循这一选择。为了确保数值稳定性，最小扩散步骤是 $\tau = T_{\min}$ 而不是 $\tau = 0$，对于任意小的 $T_{\min}$。

得分函数 $\nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau)$ 在推理时是不可计算的，因为 $\boldsymbol{x}^0$ 是未知的。然而，它可以被近似为一个扩散去噪器 $D_\theta(\boldsymbol{x}^\tau, \tau)$，该去噪器经过训练以输出 $\hat{\boldsymbol{x}}^0 \approx \boldsymbol{x}^0$ 使用以下 L_2 损失：

$$\mathbb{E}_{\tau, \boldsymbol{x} \sim p_{\text{data}}, \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})} \left\| D_\theta(\boldsymbol{x} + \sigma(\tau)\boldsymbol{\epsilon}, \tau) - \boldsymbol{x} \right\|^2. \tag{3}$$

最后，利用 Tweedie 公式 [11]，得分函数可以被估计为：

$$\nabla_{\boldsymbol{x}^\tau} \log p(\boldsymbol{x}^\tau) \approx \boldsymbol{s}_\theta(\boldsymbol{x}^\tau, \tau) = \frac{D_\theta(\boldsymbol{x}^\tau, \tau) - \boldsymbol{x}^\tau}{\sigma^2(\tau)}. \tag{4}$$

3.2. 扩散后验采样

上述采样程序描述了一种无条件的基于扩散的采样程序。这个预训练的扩散模型可以用来估计 $\boldsymbol{x}$，当有一个退化观测 $\boldsymbol{r} = \mathcal{H}(\boldsymbol{x}) + \boldsymbol{w}$ 可用时，其中 $\mathcal{H}(\cdot)$ 是一个腐蚀算子且 $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \sigma_{\boldsymbol{w}}^2 \boldsymbol{I})$ 是一个噪声项。逆问题框架建议通过从 $p(\boldsymbol{x}|\boldsymbol{r})$ 采样来估计 $\boldsymbol{x}$。这是通过用 $p_\tau(\boldsymbol{x}^\tau|\boldsymbol{r})$ 替换逆 ODE(2) 中的先验 $p_\tau(\boldsymbol{x}^\tau)$ 实现的。根据贝叶斯定理，$p_\tau(\boldsymbol{x}^\tau|\boldsymbol{r}) \propto p_\tau(\boldsymbol{r}|\boldsymbol{x}^\tau) p_\tau(\boldsymbol{x}^\tau)$，因此后验采样 ODE 为：

$$d\boldsymbol{x}^\tau = -[\nabla_{\boldsymbol{x}^\tau} \log p_\tau(\boldsymbol{r}|\boldsymbol{x}^\tau) + \nabla_{\boldsymbol{x}^\tau} \log p_\tau(\boldsymbol{x}^\tau)]\tau d\tau. \tag{5}$$

先验项可以很容易地使用 $\boldsymbol{s}_\theta(\boldsymbol{x}^\tau, \tau)$ 进行估计。然而，似然得分 $p_\tau(\boldsymbol{r}|\boldsymbol{x}^\tau)$ 是不可计算的。DPS 算法 [9] 提出通过将条件 $\boldsymbol{x}^\tau$ 替换为 $\hat{\boldsymbol{x}}^0 = D_\theta(\boldsymbol{x}^\tau, \tau)$ 来近似似然项，即对扩散终端状态 $\boldsymbol{x}^0$ 进行单步估计。最后，由于 $\boldsymbol{w}$ 是高斯噪声，(5) 可以写成：

$$d\boldsymbol{x}^\tau = -\left[\nabla_{\boldsymbol{x}^\tau} \log p_\tau(\boldsymbol{x}^\tau) - \zeta \nabla_{\boldsymbol{x}^\tau} ||\boldsymbol{r} - \mathcal{H}(\hat{\boldsymbol{x}}^0)||_2^2\right] \tau d\tau \tag{6}$$

其中 ζ 是一个用于提高采样质量的超参数，如在 [9] 中提出的。

4. 方法

给定混合信号 $\boldsymbol{y}$，我们提出通过从以下后验分布中采样来恢复混合成分：

$$p(\boldsymbol{x}_{1:K}, \boldsymbol{n}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{x}_{1:K}, \boldsymbol{n}) p(\boldsymbol{n}) \prod_{i=1}^{K} p(\boldsymbol{x}_i) \tag{7}$$

其中我们使用了源的统计独立性。

我们现在希望利用逆问题框架，通过使用扩散反向过程从后验分布中采样 $\boldsymbol{x}_{1:K}$ 和 $\boldsymbol{n}$。为此，由于源的独立性，需要求解以下 $K+1$ ODEs 以得到 $i=1,\ldots,K$:

$$d\boldsymbol{x}_i^\tau = -[\nabla_{\boldsymbol{x}_i^\tau} \log p(\boldsymbol{y}|\boldsymbol{x}_{1:K}^\tau, \boldsymbol{n}^\tau) + \nabla_{\boldsymbol{x}_i^\tau} \log p_\tau(\boldsymbol{x}_i^\tau)] \cdot \tau d\tau \tag{8a}$$

$$d\boldsymbol{n}^\tau = -[\nabla_{\boldsymbol{n}^\tau} \log p(\boldsymbol{y}|\boldsymbol{x}_{1:K}^\tau, \boldsymbol{n}^\tau) + \nabla_{\boldsymbol{n}^\tau} \log p_\tau(\boldsymbol{n}^\tau)] \cdot \tau d\tau \tag{8b}$$

为了计算公式 (8a) 和 (8b) 中的先验项，我们分别在干净语音和噪声先验上定义了两个独立的扩散模型。噪声先验由分数函数 $\boldsymbol{s}_\phi(\boldsymbol{n}^\tau, \tau)$ 和扩散去噪器 $G_\phi^N(\boldsymbol{n}^\tau, \tau)$ 表示。对于语音先验，我们假设可以从对应于每位发言人的唇部视频中提取出视觉特征 V_i。视觉模态有助于提高分数估计的准确性。因此，使用了音频-视觉评分函数 $\boldsymbol{s}_\theta(\boldsymbol{x}_i^\tau, \tau, V_i)$ 和与无分类器引导 [12] 结合的去噪器 $F_\theta^S(\boldsymbol{x}_i^\tau, \tau, V_i)$。请注意，所有语音信号共享相同的语音先验。

由于 (8a) 和 (8b) 中的似然项依赖于 $\boldsymbol{x}_{1:K}^\tau$ 和 $\boldsymbol{n}^\tau$，因此它们是不可处理的。因此我们依赖于算法 DPS 的 [9] 形式，并将其原始形式从单一信号扩展到多个信号和两种不同的先验分布，即语音和环境噪声。类似于 (6)，它源于 $\boldsymbol{z}$ 的正态性，似然得分项变为:

$$\nabla_{\boldsymbol{x}_i^\tau} \log p(\boldsymbol{y}|\hat{\boldsymbol{x}}_{1:K}^0, \hat{\boldsymbol{n}}^0) = -\zeta_{\boldsymbol{x}}(\tau) \nabla \boldsymbol{x}_i^\tau \mathcal{L}_{\text{rec}}^t \tag{9a}$$

$$\nabla_{\boldsymbol{n}^\tau} \log p(\boldsymbol{y}|\hat{\boldsymbol{x}}_{1:K}^0, \hat{\boldsymbol{n}}^0) = -\zeta_{\boldsymbol{n}}(\tau) \nabla \boldsymbol{n}^\tau \mathcal{L}_{\text{rec}}^t. \tag{9b}$$

其中 $\mathcal{L}_{\text{rec}}^\tau = ||\boldsymbol{y} - \sum_i \hat{\boldsymbol{x}}_i^0 - \hat{\boldsymbol{n}}^0||_2^2$ 是时域中的混合重构损失，而 $\zeta_{\boldsymbol{x}}(\tau), \zeta_{\boldsymbol{n}}(\tau)$ 是归一化标量。为了获得更精确的梯度，我们遵循 [3]，通过用其频域对应项替换时域混合重构误差:

$$\mathcal{L}_{\text{rec}}^f = \left\| S(\boldsymbol{y}) - S\left(\sum_{i=1}^K \hat{\boldsymbol{x}}_i^0 + \hat{\boldsymbol{n}}^0 \right) \right\|_2^2 \tag{10}$$

其中 $S(\boldsymbol{y}) = |\text{STFT}(\boldsymbol{y})|^{\frac{2}{3}} \exp j\angle \text{STFT}(\boldsymbol{y})$，和 $\text{STFT}(\cdot)$ 是short-time Fourier Transform (STFT)。

我们遵循 [4, 5]，使用 $\zeta_{\boldsymbol{x}}(\tau), \zeta_{\boldsymbol{n}}(\tau)$ 来提高采样稳定性:

Algorithm 1 后验评分估计

Require: $\boldsymbol{y}, \boldsymbol{x}_{1:K}^\tau, \boldsymbol{n}^\tau, \tau, \zeta, d, F_\theta^S, G_\phi^N$

$\hat{\boldsymbol{n}}^0 \leftarrow G_\phi^N(\boldsymbol{n}^\tau, \tau)$

$\boldsymbol{s}_\phi(\boldsymbol{n}^\tau, \tau) \leftarrow \frac{\hat{\boldsymbol{n}}^0 - \boldsymbol{n}^\tau}{\tau^2}$ ▷ 噪声先验得分

for $i = 1, \ldots, K$ **do**

  $\hat{\boldsymbol{x}}_i^0 \leftarrow F_\theta^S(\boldsymbol{x}_i, \tau, V_i)$

  $\boldsymbol{s}_\theta(\boldsymbol{x}_i^\tau, \tau) \leftarrow \frac{\hat{\boldsymbol{x}}_i^0 - \boldsymbol{x}_i^\tau}{\tau^2}$ ▷ 语音先验得分

end for

$\mathcal{L}_{\text{rec}}^f \leftarrow \left\| S(\boldsymbol{y}) - S(\sum_{i=1}^K \hat{\boldsymbol{x}}_i^0 + \hat{\boldsymbol{n}}^0) \right\|_2^2$

$\zeta_{\boldsymbol{n}}(\tau) = \frac{\zeta\sqrt{d}}{\tau \left\| \nabla_{\boldsymbol{n}^\tau} \mathcal{L}_{\text{rec}}^f \right\|_2}$

$\boldsymbol{p}_{\boldsymbol{n}^\tau} \leftarrow \boldsymbol{s}_\phi(\boldsymbol{n}^\tau, \tau) - \zeta_{\boldsymbol{n}}(\tau) \nabla_{\boldsymbol{n}^\tau} \mathcal{L}_{\text{rec}}^f$ ▷ 噪声后验得分

$\zeta_{\boldsymbol{x}}(\tau) = \frac{\zeta\sqrt{d}}{t \left\| \nabla_{\boldsymbol{x}_{1:K}^\tau} \mathcal{L}_{\text{rec}}^f \right\|_2}$

for $i = 1, \ldots, K$ **do**

  $\boldsymbol{p}_{\boldsymbol{x}_i^\tau} \leftarrow \boldsymbol{s}_\theta(\boldsymbol{x}_i^\tau, \tau) - \zeta_{\boldsymbol{x}}(\tau) \nabla_{\boldsymbol{x}_i^\tau} \mathcal{L}_{\text{rec}}^f$ ▷ 语音后验得分

end for

$\hat{\boldsymbol{n}}^0 \leftarrow G_\phi^N(\boldsymbol{n}^\tau, \tau)0$

定性:

$$\zeta_{\boldsymbol{x}}(\tau) = \frac{\zeta\sqrt{d}}{\tau \left\| \nabla_{\boldsymbol{x}_{1:K}^\tau} \mathcal{L}_{\text{rec}}^f \right\|_2} \tag{11a}$$

$$\zeta_{\boldsymbol{n}}(\tau) = \frac{\zeta\sqrt{d}}{\tau \left\| \nabla_{\boldsymbol{n}^\tau} \mathcal{L}_{\text{rec}}^f \right\|_2}. \tag{11b}$$

我们在算法 1 中总结了获取语音和噪声分量后验评分估计 $\{\boldsymbol{p}_{\boldsymbol{x}_i^\tau}\}_{i=1}^K, \boldsymbol{p}_{\boldsymbol{n}^\tau}$ 的步骤。最后，(8a)，(8b) 中的 ODEs 通过在 [10] 中提出的随机二阶采样器求解。

5. 实验研究

本节介绍了用于评估所提方法与基线方法的实验研究。我们还描述了数据集并提供了关键实现细节。

数据。为了与我们视觉编码器训练所用的数据集保持一致，使用了仅包含英语的音频-视频数据集 VoxCeleb2 [15] 的干净语音语句作为来源。仅包含英语的视频由 [16] 汇编而成，总计长度为 1,759 小时。全脸视频通过遵循 [17] 中的规定转换为口部区域视频。音频和视频的采样率分别为 16 KHz 和 25 Hz。

Table 1. DAVSS-NM 和基线方法的语音分离结果。最佳结果用粗体表示，次佳结果用下划线表示。

Method	Trained noise	Unupervised	VoxCeleb2 + DNS			VoxCeleb2 + WHAM!		
			SI-SDR ↑	PESQ ↑	ESTOI ↑	SI-SDR ↑	PESQ ↑	ESTOI ↑
Input	-	-	-2.66	1.46	0.38	-2.62	1.52	0.36
FlowAVSE [13]	DNS	✗	7.82	2.15	0.65	7.13	2.18	0.62
VisualVoice [14]	DNS	✗	1.89	1.85	0.52	1.45	1.86	0.48
AV-UDiffSE [8]	DNS	✓	-2.33	1.76	0.44	-2.27	1.81	0.44
DAVSS-NM (Ours)	DNS	✓	<u>5.06</u>	<u>2.06</u>	<u>0.61</u>	<u>4.54</u>	<u>2.04</u>	<u>0.58</u>
FlowAVSE [13]	WHAM!	✗	7.04	2.19	0.62	6.41	2.07	0.61
VisualVoice [14]	WHAM!	✗	0.03	1.67	0.46	0.35	1.78	0.46
AV-UDiffSE [8]	WHAM!	✓	-2.27	1.81	0.44	-2.33	1.76	0.44
DAVSS-NM (Ours)	WHAM!	✓	<u>4.58</u>	<u>2.07</u>	<u>0.58</u>	<u>3.93</u>	<u>1.95</u>	<u>0.58</u>

对于噪声录音，使用了两个数据集。第一个是 WHAM! [18]，包含大约 30 小时 /10 小时 /5 小时的背景噪声用于训练/验证/测试。所有噪声信号均在公园、餐厅和办公楼等城市环境中录制。第二个数据集是 DNS [19]。它由约 6.5 万个属于 150 种噪声类别的片段组成，这些片段来自 AudioSet [20] 和 Freesound[1]。录音的总时长为 181 小时。我们随机选择了 64,000 个噪声信号用于训练、100 个用于验证和 900 个用于评估。

网络架构。F_θ^S 和 G_ϕ^N 基于 NCSN++M [21]，一种轻量级的 U-Net 架构。与 [3] 类似，NCSN++M 主干网络被 STFT 和逆 STFT 包围。所有 STFT 操作均使用 510 个样本的 Hann 窗口和 160 的跳跃长度。这导致了 256 个频率槽。

对于 G_ϕ^N，我们将残差层的数量增加到 2。对于 F_θ^S，NCSN++M 主干通过视觉特征得到了增强。唇部视频帧使用在视觉语音识别上微调的 BRAVEn [17] 进行编码。每个帧都被编码为一个特征向量，在 $\mathbb{R}^{1024}$ 中。视觉模态在分辨率为 64、32 和瓶颈处与 NCSN++M 结合，首先匹配视觉分辨率到音频分辨率，然后通过 FiLM 层 [22] 调制音频特征。总的来说，我们的 F_θ^S 和 G_ϕ^N 分别有 1.295 亿和 0.397 亿个可训练参数。

基线。使用了三个基线模型，我们从头开始训练这些模型。第一个是 VisualVoice [14]，这是一种监督的音频-视频判别模型，鼓励分离出的语音与其对应的视频之间说话人身份的一致性。另一个基线是 FlowAVSE [13]，这是一个基于流匹配的监督式音频-视频生成模型。这些监督算法使用了 VoxCeleb2+WHAM! 和 VoxCeleb2+DNS 的语音噪声组合进行了训练。我们在两个说话人的混合信号上展示了我们的方法。在训练过程中，signal-to-interference ratio (SIR) 和signal-to-noise ratio (SNR) 符合后续描述的评估协议。

最后一个基线 AV-UDiffSE [8] 是一种基于逆问题方法的无监督生成方法。由于它最初是为单扬声器语音增强设计的，我们对其进行了修改以适应我们的场景，通过将 [8] 中的推导适配到多扬声器的情况。此外，我们将基于音频的基础架构替换为我们所使用的音视频架构。

训练设置。为了训练 F_θ^S, G_ϕ^N，采样了 4 秒的音频片段，结果得到了 $d = 64,000$ 和 400 个 STFT 帧。对于 F_θ^S，这相当于 100 个视频帧。我们使用离散化的扩散噪声调度 $\sigma(\tau)$ 来自 [10] 与 $\rho = 10$。在训练过程中，使用了 $T_{\max} = 10, T_{\min} = 1e-5$。训练的批量大小 F_θ^S, G_ϕ^N 对所有实验都是 16，并且我们使用 Adam 优化器，学习率为 1e-4。

后验采样。在推理阶段，使用来自 [10] 的二阶采样器。我们选择了 $T_{\max} = 4, T_{\min} = 1e-5, S_{\text{churn}} = 30, \zeta =$

[1] https://freesound.org/

0.5，并将 ODE(8a), (8b) 离散化为 400 个扩散时间步长。对于 WHAM! 噪声，使用无分类器引导权重 0.5，而对于 DNS 使用 0.8。与以前使用暖启动的方法 [3,5,7] 不同，我们的初始扩散状态 $\boldsymbol{x}_i^{T_{\max}}, \boldsymbol{n}^{T_{\max}}$ 设定为零均值。我们在实现 AV-UDiffSE 时也采用了零均值的初始状态，并将 EM 迭代次数设定为 9。

评估协议。评估集包含两个说话者的混合信号和背景噪声信号。SIR 的值为 -5/0/5 dB，而 SNR 则是从范围 [-3,3] 中随机采样的，该范围是相对于低能量说话者定义的。对于每个 SIR 值，生成了 500 个混合信号。我们考察了两种场景，匹配和不匹配。在匹配设置中，方法是在其训练的数据集上的语音-噪声数据上进行测试的。在不匹配的情况下，一个在 WHAM! 噪声上训练的模型则会在 DNS 噪声上进行测试，反之亦然。这种跨域测试使我们能够评估这些模型的泛化能力。我们使用了以分贝为单位测量的perceptual evaluation of speech quality (PESQ) 分数、extended short-term objective intelligibility (ESTOI) 和scale-invariant signal-to-distortion rati (SI-SDR) 来评估性能。

讨论。表 1 显示了 DAVSS-NM 和基线方法的性能。很明显，监督下的 FlowAVSE 表现最佳。对于所有的语音-噪声组合，DAVSS-NM 排名第二，并且是表现最好的无监督方法。它提高了 AV-DiffSE 的指标，从而显著缩小了无监督和监督范式之间的差距。DAVSS-NM 甚至超过了基于监督的方法 VisualVoice。我们推测 VisualVoice 没有获得更好的结果是因为它是基于 STFT 掩码的，在我们的具有挑战性的场景下受到了不利影响。

在不匹配的情况下，当在 WHAM!噪声上训练时，除了 VisualVoice 外的所有方法都相对保持了性能。请注意，由于 AV-UDiffSE 公式不需要在噪声数据上进行训练，因此可以从相关条目中复制其结果。令人惊讶的是，当在 DNS 噪声上训练时，所有方法的性能都优于在其 WHAM!对应物上的性能。我们假设这种现象发生是因为 DNS 是一个高度多样化的噪声数据集，其分布很可能覆盖了 WHAM!的分布。

6. 结论

我们提出了一种联合无监督扩散的语音分离和环境噪声去除方法，利用了噪声分布建模。该方法是在逆问题框架下通过耦合噪声先验与音视频语音先验开发出来的。我们的方法 DAVSS-NM 有两个主要优点。首先，与需要在退化模型变化时重新训练的有监督方法不同，DAVSS-NM 不依赖配对的训练数据，使其具有高度适应性。其次，我们的评估显示，该方法显著缩小了无监督技术与最先进的有监督算法之间的性能差距。

7. REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising diffusion probabilistic models,” Advances in neural information processing systems (NeurIPS), vol. 33, pp. 6840–6851, 2020.

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

[3] E. Moliner, J.M. Lemercier, S. Welker, T. Gerkmann, and V. Välimäki, “Buddy: Single-channel blind unsupervised dereverberation with diffusion models,” in International Workshop on Acoustic Signal Enhancement (IWAENC), 2024, pp. 120–124.

[4] E. Moliner, J. Lehtinen, and V. Välimäki, “Solving audio inverse problems with a diffusion model,” in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2023.

[5] E. Moliner, F. Elvander, and V. Välimäki, “Blind audio bandwidth extension: A diffusion-based zero-shot approach,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024.

[6] Z. Xu, X. Fan, Z. Q. Wang, X. Jiang, and R. R. Choudhury, "ArrayDPS: Unsupervised blind speech separation with a diffusion prior," in International Conference on Machine Learning (ICML), 2025.

[7] B. Nortier, M. Sadeghi, and R. Serizel, "Unsupervised speech enhancement with diffusion-based generative models," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 12481–12485.

[8] J.E. Ayilo, M. Sadeghi, R. Serizel, and X. Alameda-Pineda, "Diffusion-based unsupervised audio-visual speech enhancement," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2025.

[9] H. Chung, J. Kim, M.T. Mccann, M.L. Klasky, and J.C. Ye, "Diffusion posterior sampling for general noisy inverse problems," in International Conference on Learning Representations (ICLR), 2023.

[10] T. Karras, M. Aittala, T. Aila, and S. Laine, "Elucidating the design space of diffusion-based generative models," Advances in neural information processing systems (NeurIPS), vol. 35, pp. 26565–26577, 2022.

[11] B. Efron, "Tweedie' s formula and selection bias," Journal of the American Statistical Association, vol. 106, no. 496, pp. 1602–1614, 2011.

[12] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in NeurIPS Workshop on Deep Generative Models and Downstream Applications, 2021.

[13] C. Jung, S. Lee, J. H. Kim, and J.S. Chung, "FlowAVSE: Efficient Audio-Visual Speech Enhancement with Conditional Flow Matching," in Interspeech, 2024, pp. 2210–2214.

[14] R. Gao and K. Grauman, "VisualVoice: Audio-visual speech separation with cross-modal consistency," in IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[15] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep speaker recognition," in INTERSPEECH, 2018.

[16] S. Bowen, H. Wei-Ning, L. Kushal, and M. Abdelrahman, "Learning audio-visual speech representation by masked multimodal cluster prediction," in International Conference on Learning Representations (ICLR), 2022.

[17] A. Haliassos, A. Zinonos, R. Mira, S. Petridis, and M. Pantic, "BRAVEn: Improving self-supervised pre-training for visual and auditory speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2024, pp. 11431–11435.

[18] G. Wichern, J. Antognini, M. Flynn, L.R. Zhu, E. McQuinn, D. Crow, E. Manilow, and J. Le Roux, "WHAM!: Extending speech separation to noisy environments," in Interspeech, 2019, pp. 1368–1372.

[19] H. Dubey, V. Gopal, R. Cutler, S. Matusevych, S. Braun, E.S. Eskimez, M. Thakker, T. Yoshioka, H. Gamper, and R. Aichner, "Deep noise suppression challenge," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[20] J.F. Gemmeke, D.P.W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R.C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, 2017.

[21] J.M. Lemercier, J. Richter, S. Welker, and T. Gerkmann, "StoRM: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2724–2737, 2023.

[22] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "FiLM: Visual reasoning with a general conditioning layer," in AAAI Conference on Artificial Intelligence, 2018.