论文阅读分享

DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification
Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

肖叶

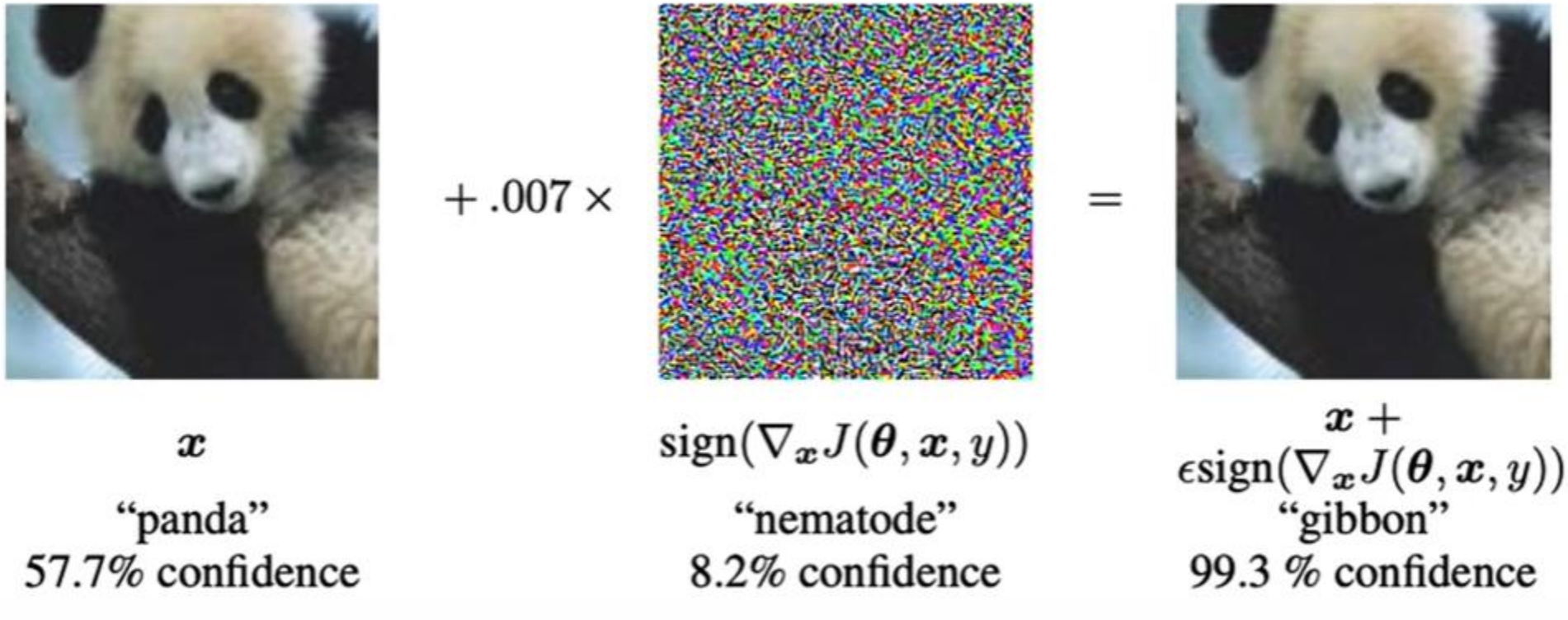
2025.02.13

DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

Qing Wang[†], *Jixun Yao*[†], *Zhaokai Sun*[†], *Pengcheng Guo*[†], *Lei Xie*[†*], *John H.L. Hansen*[§],
[†]Audio, Speech and Language Processing Group (ASLP@NPU), School of Computer Science,
Northwestern Polytechnical University, Xian, China
[§]Center for Robust Speech Systems (CRSS), The University of Texas at Dallas, USA

https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

Adversarial Attack



在语音中，对抗攻击（Adversarial Attacks）通过在任意语音中引入精心设计的扰动来迷惑目标模型。

Adversarial Attack

欺骗攻击：通过模仿目标说话人的音色来欺骗系统。这种攻击主要依赖于对音色的模仿，而不一定利用模型的漏洞。

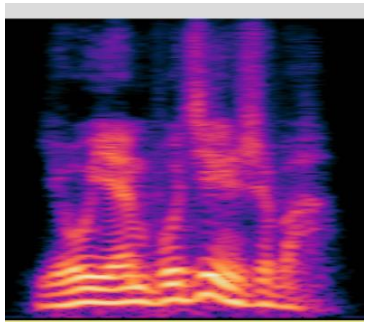
对抗攻击：通过在音频中添加精心设计的微小扰动来欺骗模型。这些扰动是基于**模型的漏洞**生成的。

模型可能被利用的漏洞：

1. 模型对微小扰动的敏感性：SID模型可能对输入音频中的微小扰动过于敏感，这些扰动虽然在人类听觉范围内难以察觉，但足以使模型的预测结果发生改变。
2. 模型对特定特征的过度依赖：如果SID模型过度依赖某些特定的音频特征（如特定频率范围内的信号），攻击者可以通过有针对性地修改这些特征来欺骗模型。

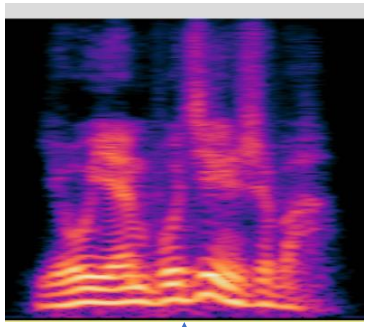
Adversarial Attack

提出了一种在说话者识别中使用的音色保留的对抗性攻击。在语音转换（VC）模型的**不同训练阶段**，通过添加一个**对抗性约束**来生成音色保留的对抗性音频。



正常SID模型

标签 y



目标SID模型

输出

标签 y

比较
Y==y'

对抗样本攻击成功

比较
Y!=y'

对抗性约束调整 VC 模型的输出

VC模型输出

https://arxiv.org/abs/2309.00929 Timbre-reserved Adversarial Attack in Speaker Identification

Adversarial Attack

LCE(·)是制作对抗性样本的损失函数

$$\begin{aligned} &\min L_{CE}(f(M+\delta), y'), \\ &\text{s.t.} \quad \|\delta\| < \epsilon, \end{aligned}$$

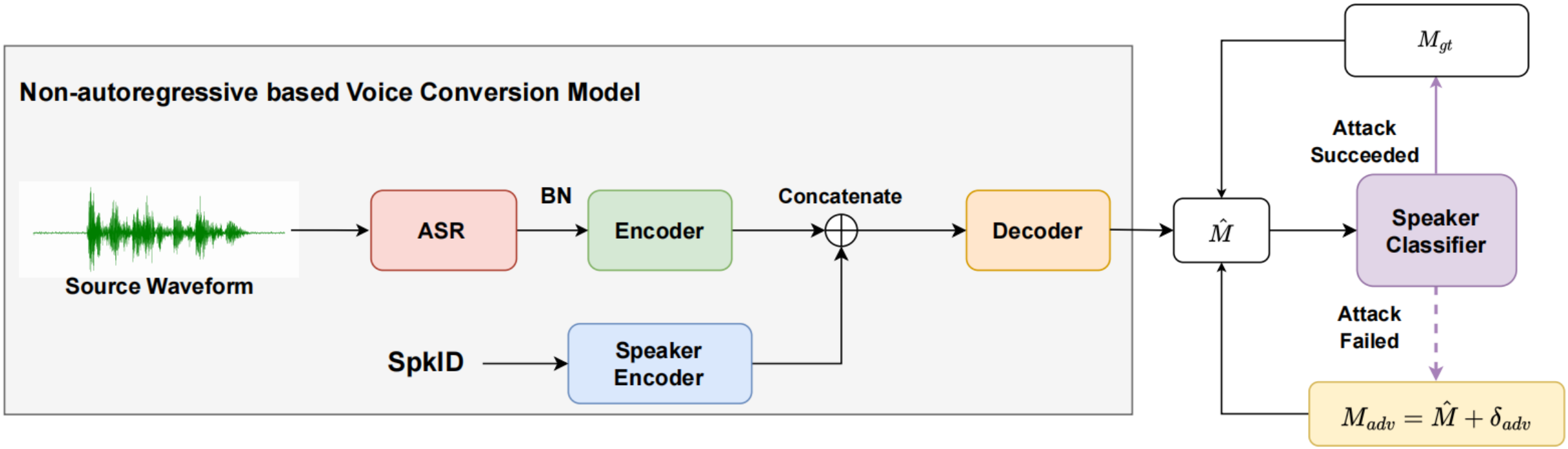
对抗性约束优化，骗过SID模型:

$$\delta \leftarrow \text{clip}_{\epsilon}(\delta - lr \cdot \text{sign}(\nabla_{\delta} L_{CE}(f(M+\delta), y')))$$

https://arxiv.org/abs/2309.00929 Timbre-reserved Adversarial Attack in Speaker Identification

Adversarial Attack

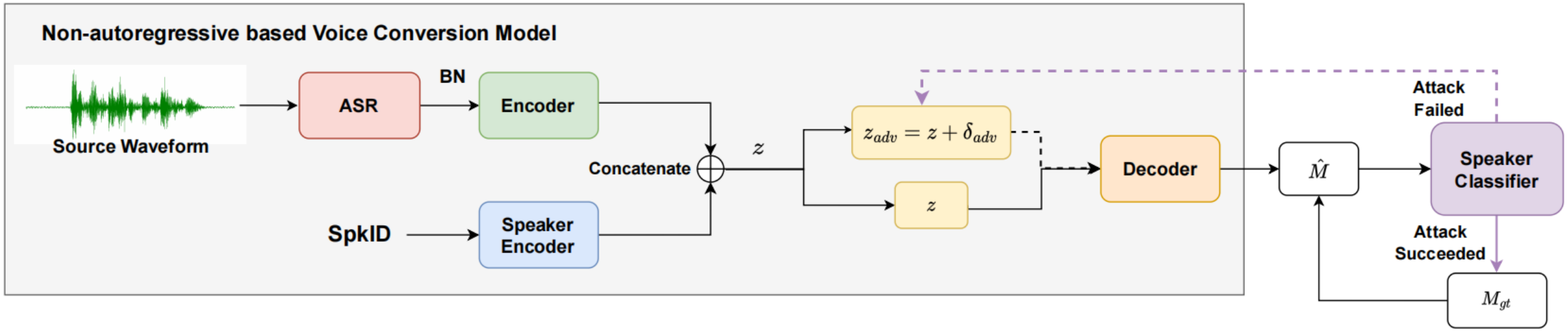
（1）在Mel频谱图上加入对抗性约束



https://arxiv.org/abs/2309.00929 Timbre-reserved Adversarial Attack in Speaker Identification

Adversarial Attack

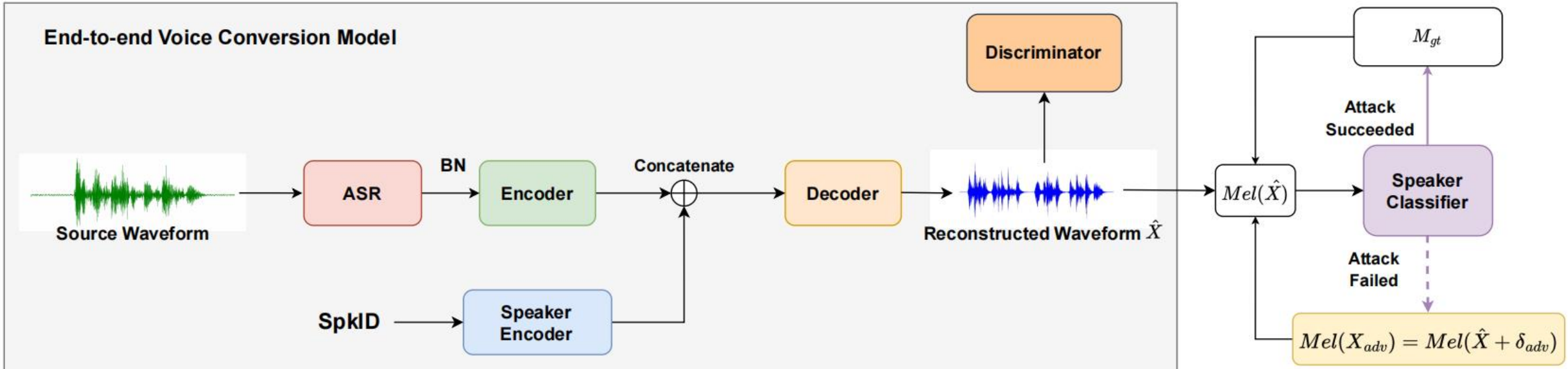
（2）在潜在表示上加入对抗性约束



https://arxiv.org/abs/2309.00929 Timbre-reserved Adversarial Attack in Speaker Identification

Adversarial Attack

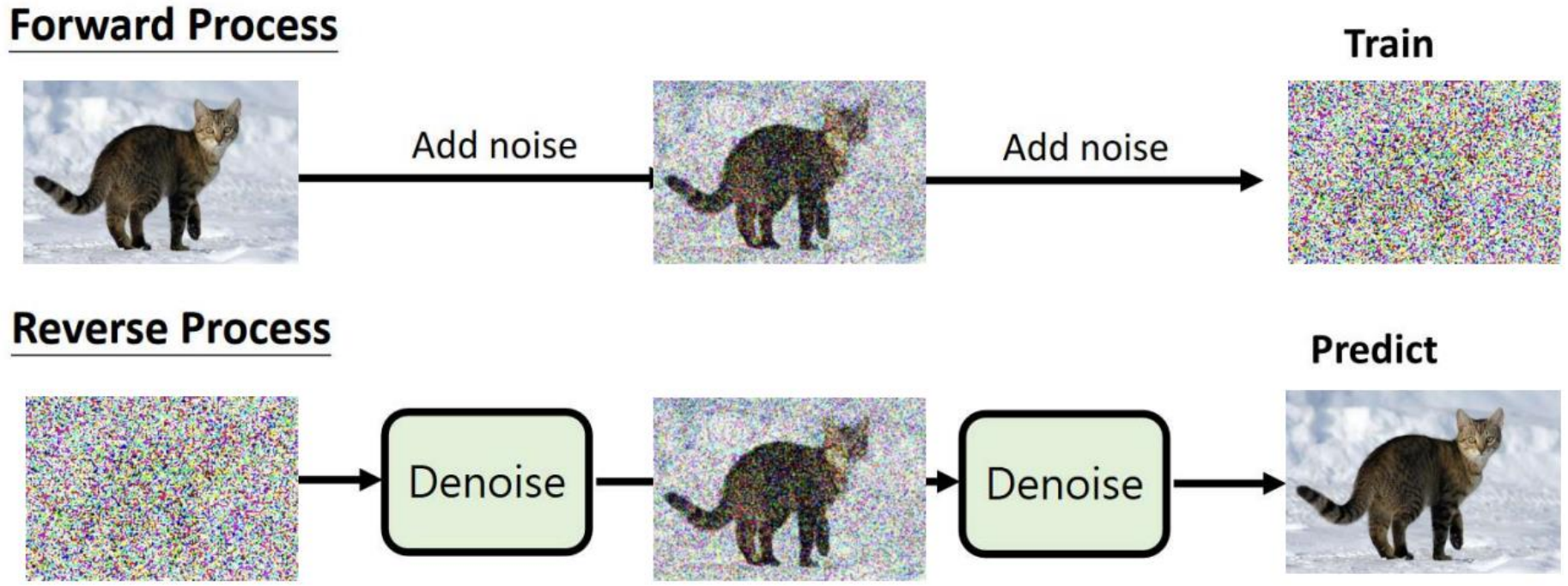
（3）在重构波形上加入对抗性约束



https://arxiv.org/abs/2309.00929 Timbre-reserved Adversarial Attack in Speaker Identification

DiffAttack--Motivation

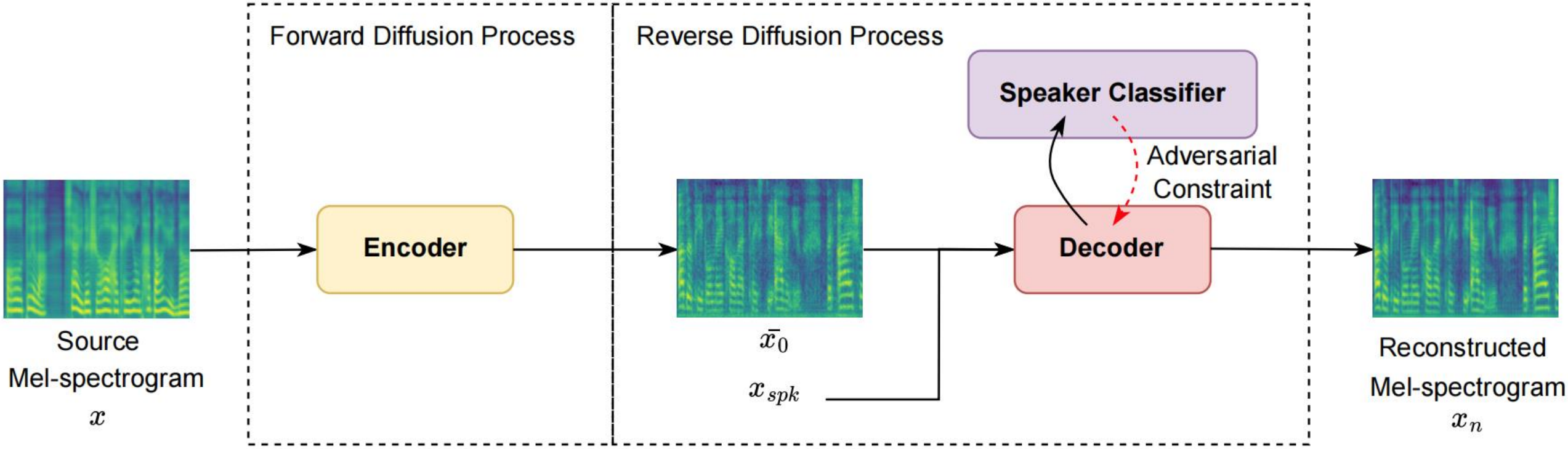
DiffVC是一种基于扩散概率建模的语音转换方法。在DiffVC的框架中，**正向扩散**过程逐渐向数据添加高斯噪声，作为编码器；相反，**逆向**扩散过程则尝试去除这些噪声，作为解码器。
与上述高斯噪声类似，**传统对抗攻击**通常在优化中利用随机采样的高斯噪声。
对抗约束隐式引导逆向扩散过程，使其与目标说话人分布对齐？



https://arxiv.org/abs/2109.13821 Diffusion-Based Voice Conversion with Fast Maximum Likelihood Sampling Scheme
https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

DiffAttack--System Overview

两个基本组件：**编码器和解码器**，分别作为正向和逆向扩散过程。编码器执行正向扩散过程，逐步向数据中引入高斯噪声；解码器则执行逆向扩散操作以消除这些噪声。在训练阶段，**编码器的作用是生成与说话人无关的表示，而解码器的任务是在对抗约束的引导下重建目标Mel频谱图。**

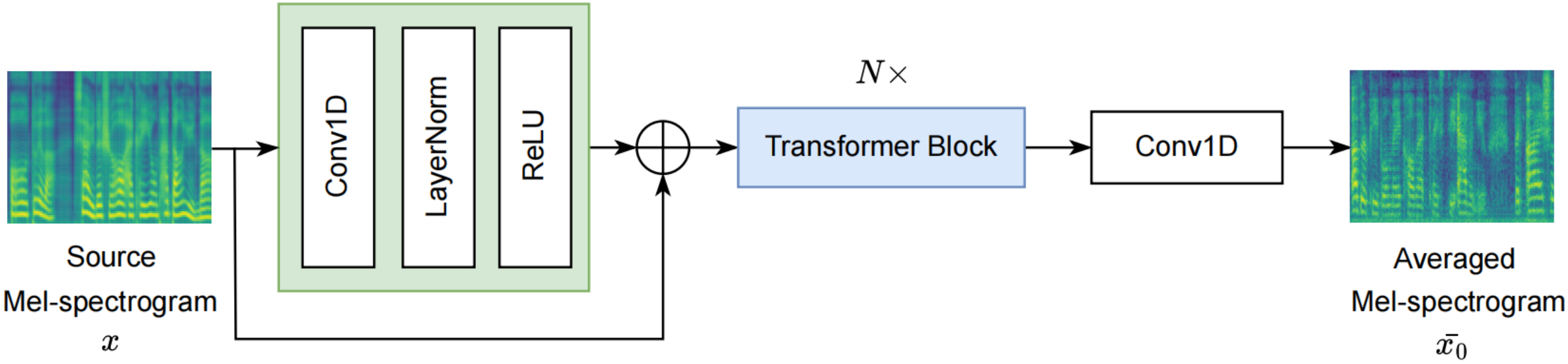


https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

DiffAttack--Adversarial Diffusion Process

在**正向扩散过程**中，**平均Mel频谱图**作为**编码器**的目标。主要的训练目标是最小化编码器输出与真实平均Mel频谱图之间的均方误差（MSE），其公式如下：

$$\mathcal{L}_{enc} = ||Enc(x) - \bar{x_0}||_2$$



https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

DiffAttack--Adversarial Diffusion Process

为什么使用平均梅尔频谱图作为目标?

可以引导编码器学习如何去除输入语音中的说话人特定特征，从而生成一个更接近通用语音特征的表示。

- **说话人特征**：每个说话人的语音都有其独特的音色特征，这些特征主要体现在梅尔频谱图的特定频率成分和能量分布上。例如，某些频率范围的能量峰值、频谱的形状等，这些特征是特定说话人的标志。
- **通用特征**：语音信号中也存在一些与说话人无关的通用特征，例如语音的基本频率成分、共振峰的分布等。这些特征是所有语音信号共有的，与具体的说话人无关。

在平均过程中，特定说话人的独特频率成分（如特定的共振峰、能量峰值等）会被多个说话人的数据“平滑”掉。因为每个说话人的这些特征是不同的，当它们被平均时，这些特定的特征会被削弱，甚至消失。

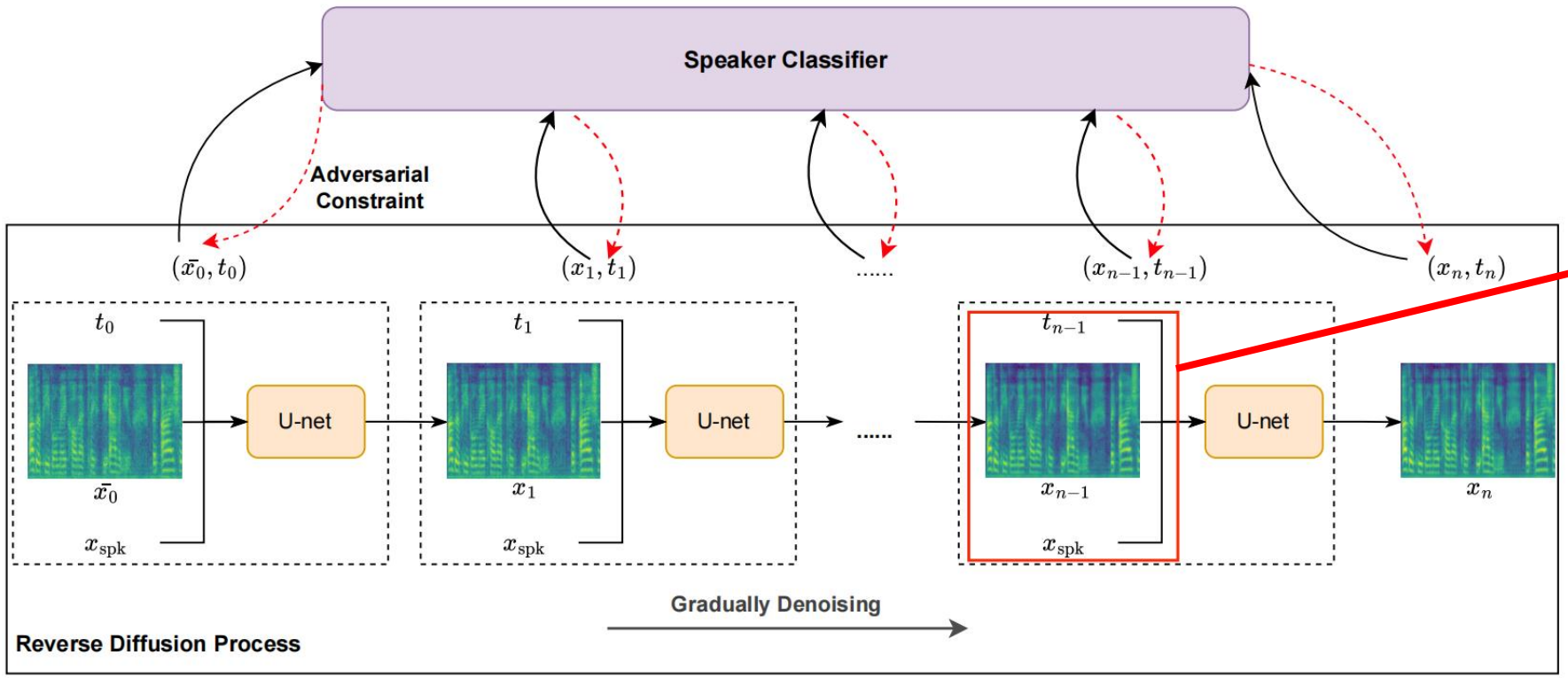
如何平均?

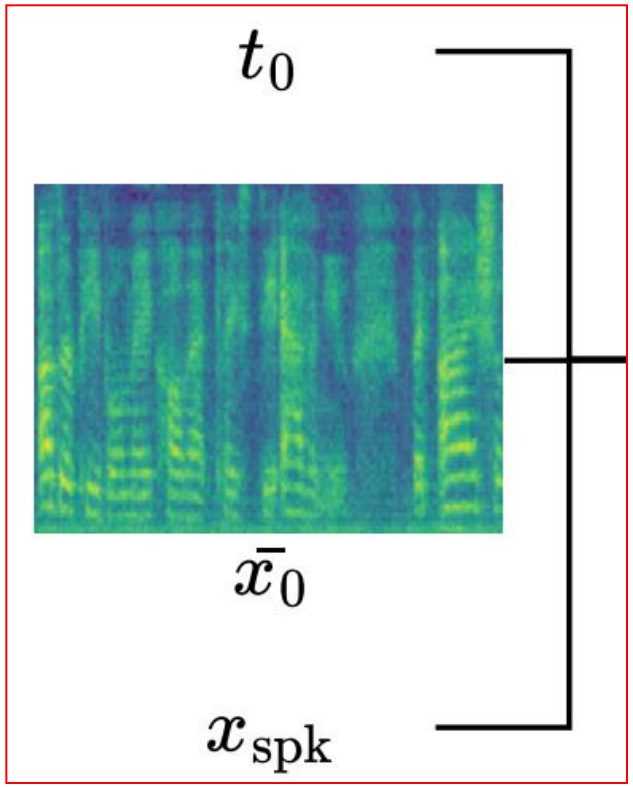
1.先通过蒙特利尔强制对齐器（Montreal Forced Aligner）将语音帧与音素对齐，每个音素对应一段语音帧，才能准确地提取每个音素的特征。
2.对于每个音素对应的语音帧，计算这些帧的Mel频谱的平均值。这样得到的平均Mel频谱特征，反映了该音素在Mel频谱上的平均特性。

https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

DiffAttack--Adversarial Diffusion Process

在**逆向扩散**过程中，在**每个时间步，将说话人嵌入和平均Mel频谱图拼接**后输入解码器。
通过添加一个说话者分类器来进行对抗性约束。预测的梅尔光谱图由说话者分类器进行分类，并确定它是否是目标说话者来决定是否添加对抗性约束。每个时间步t的平均mel谱图x_t是说话人分类器的输入，如果说话人分类器的预测是目标说话人，则模型仅使用原始损失进行优化。





https://arxiv.org/abs/2501.05127 DiffAttack: Diffusion-based Timbre-reserved Adversarial Attack in Speaker Identification

Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

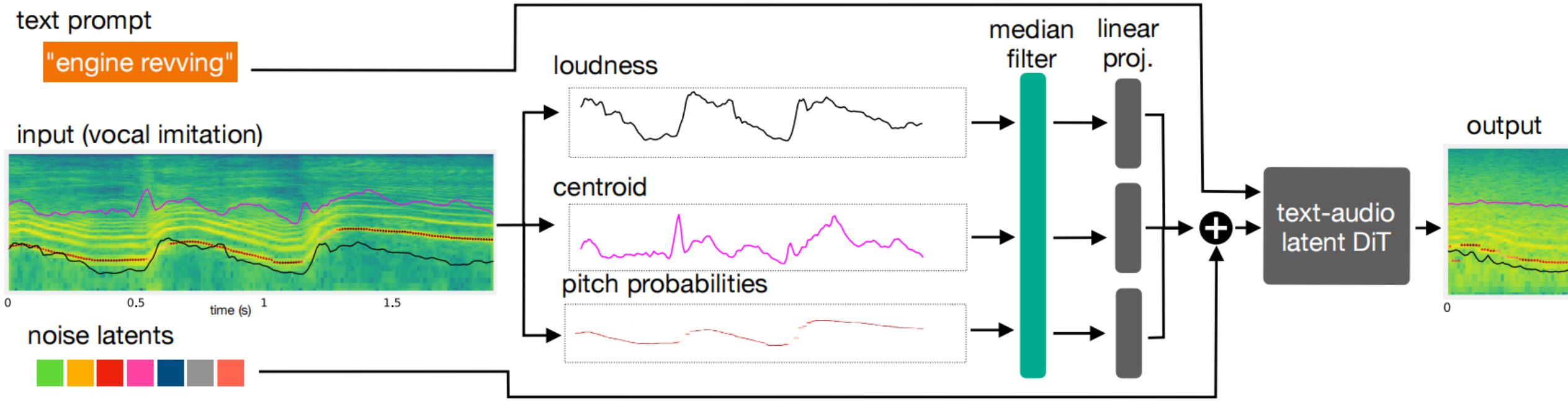
Hugo Flores García[io], Oriol Nieto[i], Justin Salamon[i], Bryan Pardo[o], Prem Seetharaman[i]
[i]*Adobe Research*, [o]*Northwestern University*
hugofloresgarcia@u.northwestern.edu

https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Sketch2Sound--overview

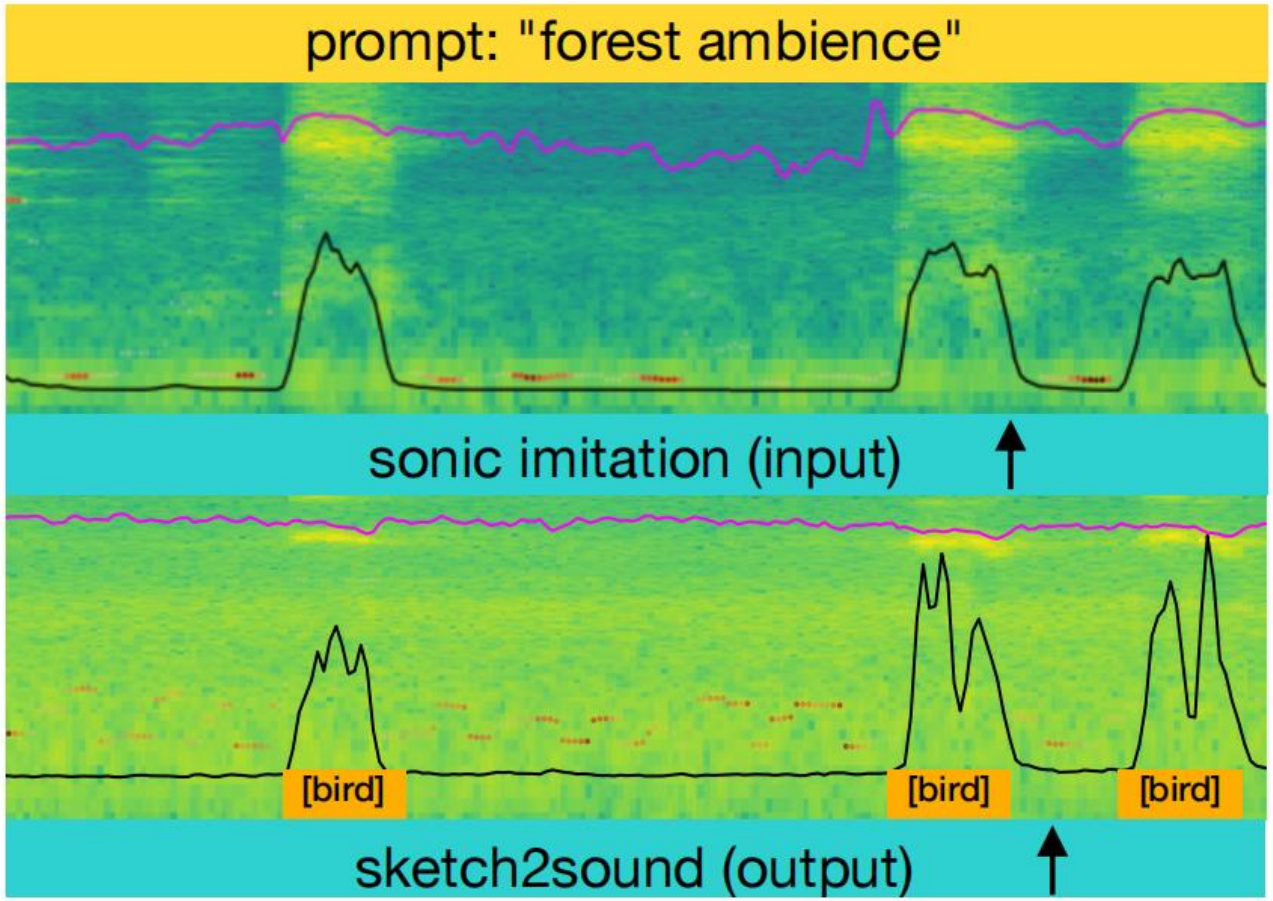
Sketch2Sound 是一种将模仿声音转换为新声音的技术。它从用户输入的模仿声音中提取三个关键信号：**响度**（音量）、**频谱质心**（声音的亮度）和**音高概率**（声音的高低变化）。这些信号经过编码后，融入基于 DiT 的文本到声音生成系统中。

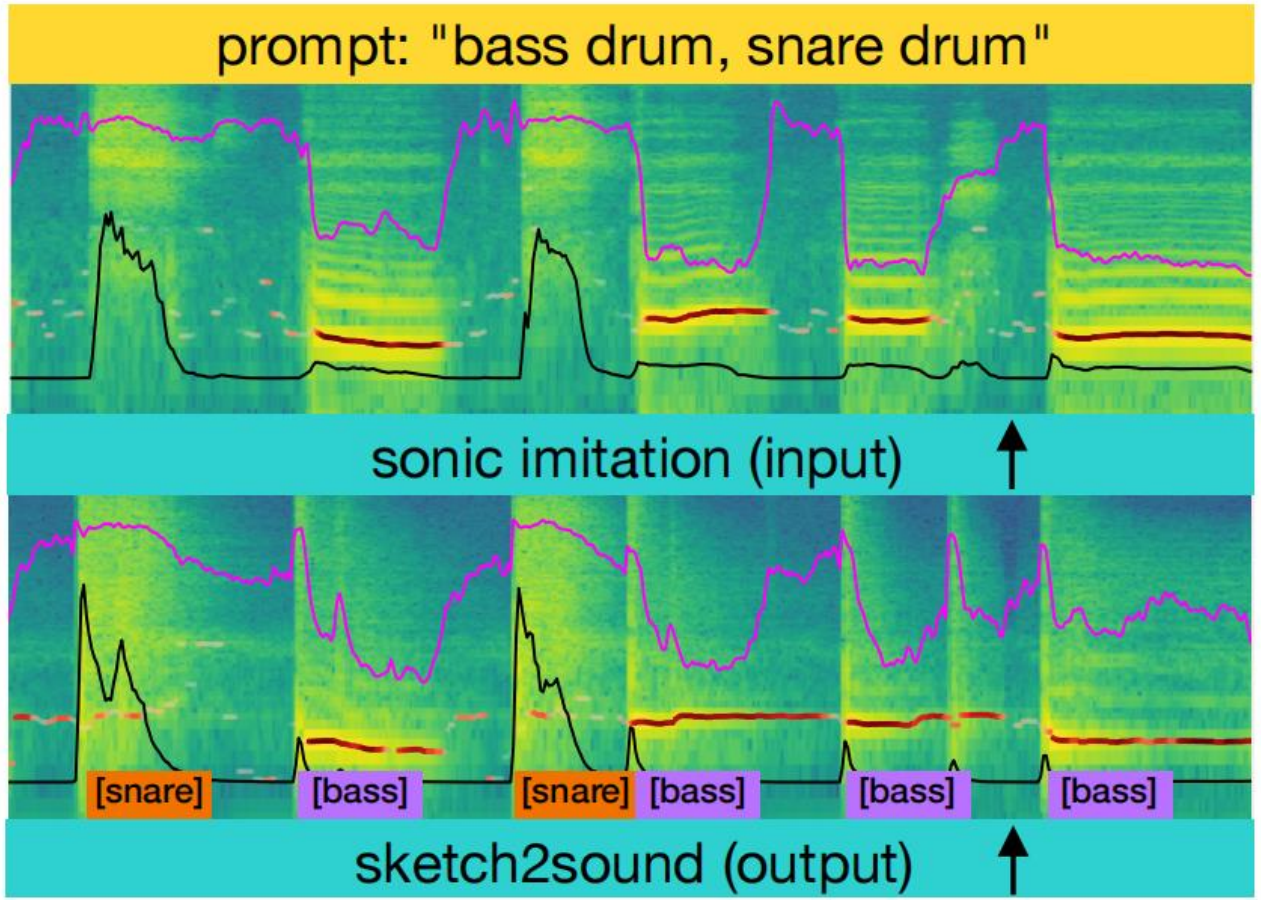
- **响度**：表示声音的强弱，用于控制声音的能量和响亮程度。
- **光谱质心**：表示声音的亮度或音色，使其更明亮或更沉闷。
- **音高概率**：表示声音的旋律特征。



https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Sketch2Sound—semantics of control curves





https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Sketch2Sound--motivtion

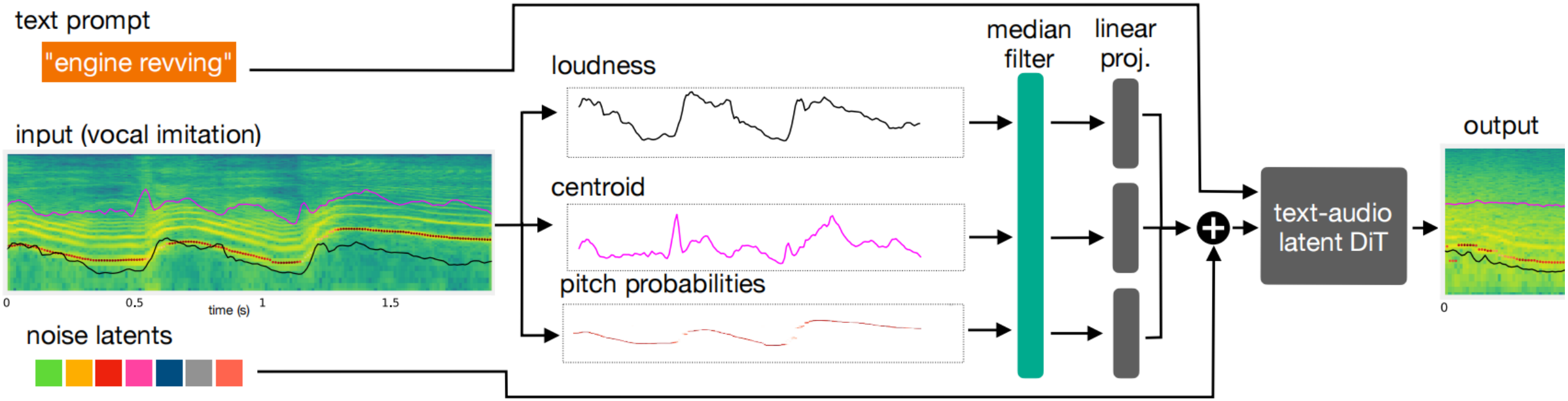
Foley sound：拟声

text-to-sound systems only：

1. 修改生成的声音的时间特征，以便它们能够与编辑时间轴上的视觉效果同步
2. 将声音抽象为文字不够直观，文字描述可能无法捕捉到声音的情感和表现力。

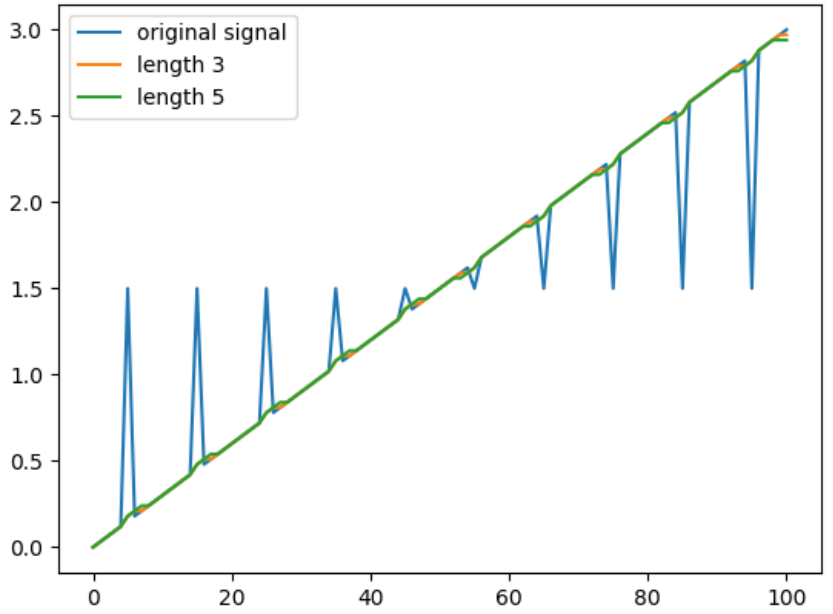
“*The human voice is a gestural sonic instrument*”

描述形容+声音模仿



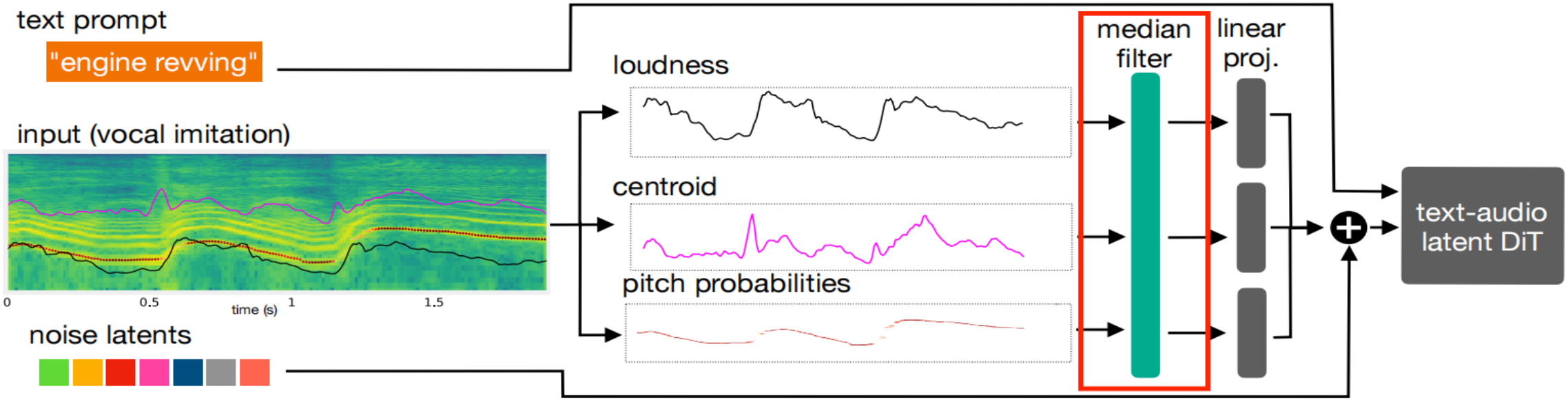
https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Sketch2Sound—median filter



如果一个信号是平缓变化的，那么某一点的输出值可以用这点的某个大小的邻域内的所有值的统计中值来代替。
较大的中值滤波器：会更大幅度地平滑控制信号，减少细节，使得生成的声音更接近一个“草图”或“轮廓”，整体趋势更明显。
较小的中值滤波器：会保留更多的细节，使得生成的声音更精确地跟随输入的声音模仿。

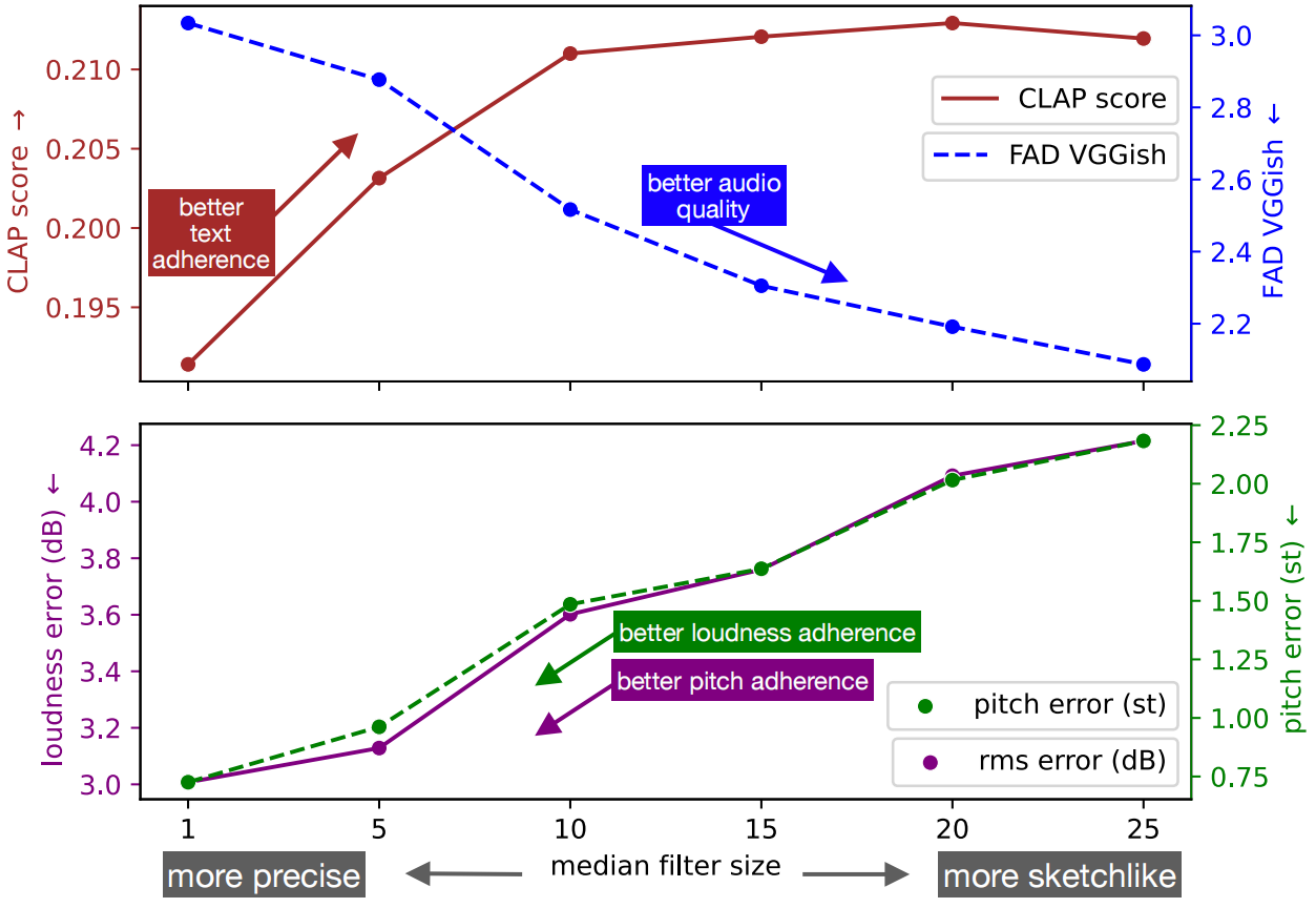
文章里提到，通过在使用控制信号之前应用不同窗口大小的中值滤波器，来改变在训练过程中使用的控制信号的时间细节。



https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations

Sketch2Sound—median filter

如果用户希望生成的声音严格遵循输入的声音模仿，可以使用较小的中值滤波器。这样，生成的声音会更精确地反映输入模仿的每一个细节。然而，如果输入的声音模仿本身不够精确生成的声音可能会带有模仿者的特征，导致音质下降。
如果用户更关注生成声音的音频质量和与文本提示的一致性，可以使用较大的中值滤波器。这样，生成的声音会更自然、更高质量，并且更符合文本描述的语义特征。



https://arxiv.org/abs/2412.08550 Sketch2Sound: Controllable Audio Generation via Time-Varying Signals and Sonic Imitations