利用生成对抗网络增强语音缺失音频信号

Deniss Strods都柏林城市
大学计算机学院格拉斯内文，
都柏林9区，爱尔兰
deniss.strods2
@mail.dcu.ie

艾伦·F·斯米顿都柏林城市
大学数据分析洞察中心爱尔兰
都柏林9区格拉斯内文
alan.smeaton@dcu.ie

arXiv:2305.05780v1 [cs.SD] 9 May 2023

摘要—**受损音频中的间隙、丢失和短片段是一个常见问题，尤其是当它们出现在语音中时，更令人烦恼。本文使用机器学习来重新生成音频语音信号中最长达 320ms 的间隙。通过将音频转换为梅尔谱图，并使用图像修复来重新生成间隙，将音频再生转化为图像再生。随后，使用 Parallel-WaveGAN 声码器将完整的梅尔谱图转换回音频，并将其整合到音频流中。我们从公开可用的 LJSpeech 数据集中选取了 1300 个时长在 1 至 10 秒之间的语音音频片段作为样本。结果表明，在配备 GPU 的系统上，使用生成对抗网络可以近乎实时地再生音频间隙。正如预期的那样，音频中的间隙越小，填补间隙后的质量越好。在 240ms 的间隙下，表现最佳的模型的平均意见分（MOS）平均值为 3.737，评分范围为 1（最差）至 5（最好），这一结果足以让人感知为接近不间断的人类语音。**

索引词—**音频缺失、梅尔频谱图、图像修复、生成对抗网络**

一、引言

语音音频通过网络（包括蜂窝网络）传输时，可能会出现中断、间隙和短片段数据损坏。本文探讨如何利用生成对抗网络（GANs）这一种机器学习方法，通过实时填补这些间隙来提升语音音频的质量。虽然已有基于主成分分析等方法来提升语音音频质量、清理音频信号的传统机器学习方法，但目前还没有适用于实时填补间隙的有效方法。我们的方法是将有缺失的音频转换为梅尔谱图，从而将音频再生转化为图像修复，类似于 [22] 中介绍的工作。我们研究了会造成音频间隙的数据传输丢包情况，间隙长度从40毫秒至320毫秒不等，并模拟了最多丢失8个数据包的一系列网络丢包情况。

下一节将综述涵盖生成对抗网络（GAN）应用及其变体架构，以及噪声环境中的语音增强的相关研究。随后，我们将介绍实验设置，接着展示实验结果，并给出结论。

本研究部分得到爱尔兰科学基金会（SFI）资助，资助编号为 SFI/12/RC/2289 P2。

979-8-3503-4057-0/23/$31.00 ©2023年 电气与电子工程师学会

II. 相关工作

A. 含噪音频中的语音增强

语音增强旨在改善因噪声而退化的语音信号在感知和听感方面的质量。这项任务应用于移动通信、助听器和鲁棒语音识别[12], [16], [17]中。即使语音质量已经达到理解说话者内容所需的最低水平，语音增强仍然十分必要，因为它能够减轻听者的疲劳感。由于音频中语音的保真度较低，听讲话音的审美享受也会受到影响。简单地提高语音信号的保真度，还可能提升语音转文本算法的性能 [14]。

语音信号中的噪声可能来自嘈杂的通信信道，也可能是语音信号源自嘈杂的环境。在网络电话传输中，网络丢包会导致传输出现间隙，并降低语音信号的感知质量和美感。目前，对于实时通信中音频信号间隙的再生问题，仍没有理想的解决方案。

语音增强 有多种方法，包括使用主成分分析、基于统计模型的算法、谱减法和维纳滤波。近期已有研究使用生成对抗网络[12], [16], [17],处理语音增强问题，不过这些研究的训练数据分别包含 462,880 条话语或 224,000 个句子，远多于本文所用的数据。处理音频和/或语音增强的生成对抗网络通常使用梅尔频谱图，即将音频信号表示为图像，如后文图1所示。与较高频率相比，梅尔频谱图能更好地捕捉人类对较低频率声音的感知；而频谱图则是信号频率组成随时间变化的可视化表示。随后可以调整其中的特征，以改善再生成语音音频的听感或质量。

生成式深度学习架构（例如生成对抗网络）的存在，使我们能够解决间断语音问题。生成对抗网络能够从任意复杂的数据分布中生成数据，这表明可以训练生成对抗网络，以实时重建缺失的音频。在实际环境中训练此类模型所需的数据可以从说话者的

此前的语音，以及一个经过训练、能够为该说话人重新生成存在缺失片段的音频信号的模型。作为说话人之间协议的一部分，可以交换语音模型，用于增强接收到的语音信号，通过弥补丢包造成的通信缺失来恢复语音。

B. 生成对抗网络（*GAN*）

生成对抗网络（GANs）是一种使用深度学习进行生成式建模的方法，由 Goodfellow 等人于 2014 [7]首次提出。生成模型无需大量标注的训练数据，即可学习数据分布。基于训练数据，生成对抗网络能够生成与其训练集相似的新数据。

GAN架构基于博弈论，其中反向传播信号通过竞争过程得出。两个神经网络，即生成器（G）和判别器（D），相互竞争。 G 通过尝试欺骗 D，使其将生成样本识别为真实样本 [8]，从而学习对数据分布进行建模。这种方法特别有用，因为GAN模型可以通过训练来模拟任意数据分布，因此还有许多尚待发现的实际应用 [1]。

生成对抗网络最初仅用于图像增强任务，例如生成高质量图像。直到大约 2017 年，首个能够生成人脸图像的生成对抗网络才问世。生成对抗网络因此受到关注，如今在需要合成数据生成的领域中都能看到它的应用，包括自然语言处理 [3]、、计算机视觉 [2] 和音频生成 [4]。

对于某些应用，使用原始 GAN架构训练 GAN 较为困难，因为有些生成器无法充分学习训练数据的分布，因此，文献2015 [18]提出了深度卷积生成对抗网络（DCGAN）。在这一架构中，研究人员使用卷积神经网络取代了全连接多层感知机神经网络。文献 [18] 的作者确定了一组适用于 GAN 框架的卷积神经网络。为稳定训练过程，生成器在各层使用 ReLU激活函数，但在最后一层使用 Tanh函数。开发 DCGAN 时确定的一些模型约束，为许多后续的、基于 DCGAN 的 GAN架构奠定了基础。其中包括可加入标签的条件生成对抗网络（cGAN） [15]、用于音频合成的 WaveGAN [4] ，以及同样用于音频、并使用梅尔谱图形式的辅助输入特征的 Parallel WaveGAN [20] 。

对于语音 增强， [16]中介绍了语音增强生成对抗网络（SEGAN）架构。生成器网络用于增强信号，其输入为噪声信号和潜在表示，输出为增强信号。生成器的结构与自动编码器相同。编码过程包含若干步幅卷积层，后接参数化整流线性单元（PReLU），其中滤波器每经过 N 步，结果便进行一次卷积。判别器扮演专家分类器的角色，用于判断分布是真实的还是判别为假后，生成器会调整权重，使其趋向于真实分布。

鉴于生成对抗网络在图像到图像转换和图像修复领域已发展成熟，而这些任务与间隙再生任务类似，[9], [11], [19], [22], 我们提出将音频信号转换为梅尔谱图，并使用图像修复器填补图像中的间隙。随后，可对梅尔频谱图进行修复，再通过 Parallel-WaveGAN 等神经声码器将其转换回音频 [20]。我们提出训练一个模型，在梅尔谱图末尾的固定位置再生间隙。这样可以简化生成对抗网络的处理过程，因为它始终知道图像修复的位置。

III. 实验设置

A. 数据集

我们使用了公共领域的 LJSpeech数据集， [10]其中包含 13,100 段由单一说话者朗读的短音频片段，朗读内容选自 7 本英语非虚构类书籍。这些片段时长为 1 至 10 秒，总时长约为 24 小时，数据集包含 13,821 个不同的单词。在我们的实验中，我们随机选取了 1,300 段片段，其中 1,000 段用于训练，300 段用于测试。我们使用子集，是为了更贴近真实应用场景：对于需要在音频电话和视频会议中进行间隙填补的特定语音，训练数据可能较少。如前所述，相关工作 [12], [16], [17] 使用了 462,880 条话语或 224,000 个句子进行训练。

B. 数据预处理

在将原始 22千赫兹 音频片段通过短时傅里叶变换（STFT）转换为梅尔频谱图[5], 之前，先对信号的开头和结尾进行裁剪，以去除静音。随后，对音频执行 STFT，帧长为 1024 个采样点（对应 46 毫秒），帧移为 256 个采样点（11 毫秒）。通过平方函数提升 STFT 峰值，以突出语音音高，然后利用 80 通道特征将其转换到梅尔刻度。梅尔滤波器组的另一个参数——频率被设置为使音频涵盖 80Hz 至 7.6kHz 的范围。

对梅尔谱图进行缩放，使每个通道具有近似恒定的能量，随后进行 log10动态范围压缩。通过减去数据集的全局均值（μ），再除以标准差（σ）对其进行归一化。最后，将数值归一化到 [-1,1] 的范围内。为执行归一化，收集了整个数据集的统计数据。音频片段的长度在时域上统一为 256 帧（对应 2.8 秒）。

为了模拟基于分组的IP中典型的通信故障，我们在梅尔频谱图末尾创建了带有40毫秒至320毫秒音频间隙的梅尔频谱图，因为音频通信具有实时性，必须尽快进行重建。因此，间隙之后的音频信号不可用。40毫秒至320毫秒的间隙可模拟最多连续丢失8个数据包的情况，

假设音频压缩每个数据包包含40毫秒的音频。间隙超过320毫秒时，可能会生成原本未说出的词语，因为快速语音的典型语速最高可达每分钟160个词 [21]（每个词375毫秒），因此这设定了我们进行间隙填补的目标上限。

完整的数据集由源图像（存在缺口的梅尔谱图）和目标图像（真实值）的梅尔谱图对构成。图1展示了一个训练样本对。模型的输入是掩码图像，模型尝试生成与真实值相似的完整图像。

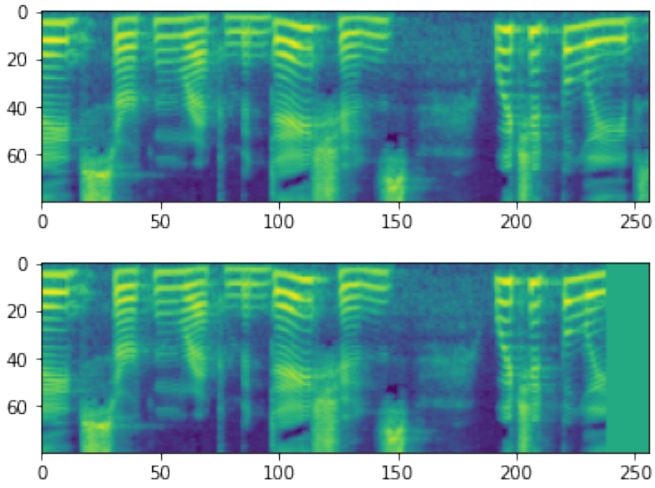


图 1. 用作模型输入的训练对：真实值（上）和带有绿色间隙的 maskedMel（下）。

C. 模型和损失函数

我们图像修复的起点是 Pix2Pix GAN[9]。该模型此前曾用于多个图像到图像转换任务，也曾用于类似的工作 [22]。在那项相关工作中，作者研究了基于同步音频和视频构建联合特征空间，其中视频由音频的频谱图组成。该工作重点是对频谱图进行图像修复，以重新生成嘈杂或受损音频，不过他们的实验针对的是音乐音频，而不是我们在此关注的语音。

为了建立本研究的基准模型，采用了图 2 所示的标准 U-Net 架构五层生成器，并使用 L1 像素级损失，输入尺寸为 256x256。作为 Pix2Pix 架构的一部分，使用 Patch-GAN 判别器计算对抗损失。该判别器会生成标量对抗损失，并通过均方误差（MSE）比较图像中的小图像块。这些图像块构成一个网格，并生成 0 到 1 之间的评分，其中每个图像块都会被分类为真实或伪造。

对标准 U-Net 架构进行了调整。我们调整了卷积神经网络层的步幅配置，以适应大小分别为 125x128 和 256x80 像素的不同输入维度，从而更好地匹配我们的数据集特征。我们还采用了不同版本的损失函数，引入了近期图像修复研究 [11], [19] 提出的更先进的损失准则。我们替换了 L1 逐像素损失。

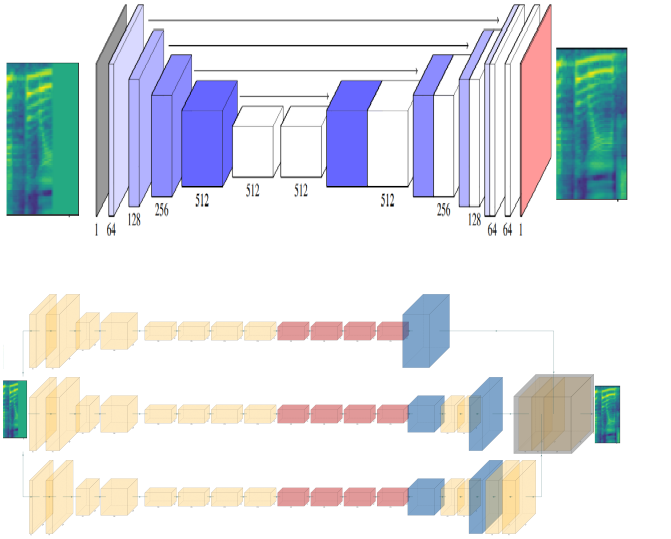


图 2. U-Net 网络（上），跳跃连接以相同颜色的色块表示；GMCCN 网络（下）。

采用 VGG19 特征匹配损失，利用 VGG19 卷积神经网络的特征提取层比较生成图像与真实值图像，并根据比较所得的均方误差更新网络中的梯度。VGG19 已在 ImageNet 上进行了预训练，并将 VGG19 特征匹配损失添加到图像修复片段中。

生成式多列卷积神经网络（GMCCN） [19] 采用了与基于 U-Net 的生成器相同的设置。如图 2 所示，GMCCN 并行使用 3 个网络，并在最后一层将它们的输出进行拼接。我们使用 Patch-GAN 判别器计算对抗损失，并使用 VGG19 特征匹配损失训练 GMCCN 生成器。实验过程中发现，该网络容易出现梯度爆炸问题，因此我们通过引入批归一化层对 GMCCN 进行了改进。

我们流程的最后阶段是使用 Parallel-WaveGAN 声码器将梅尔频谱图转换为波形。通常，Parallel-WaveGAN 声码器用于 Tactron 2文本转语音（TTS）流程 [20]，其中文本被转换为梅尔谱图，声码器生成音频；它也可用于任何梅尔谱图到音频的转换。用于训练 Parallel WaveNet 的方法不使用任何蒸馏过程，因此生成的模型较小，整体处理速度更快。此处使用了一个在 LJSpeech数据集上预训练的 Parallel-WaveGAN 模型，以避免成本高昂的训练过程。

我们的实现基于 TensorFlow，并在 NVIDIA GTX 1660 Super GPU 上进行训练。网络采用 Adam 优化器进行训练，学习率设为 $1e-4$，批量大小设为 1。所有数据预处理、转换为梅尔频谱图以及数据集矩阵乘法均通过 TensorFlow API 完成。模型训练期间的性能记录在 TensorBoard 中。该实现

Parallel-WaveGAN 声码器的实现基于 PyTorch，权重取自公开 GitHub 仓库：
https://github.com/kan-bayashi/ParallelWaveGAN

初始实验模型在包含 1,300 个样本的子集上训练了 40个训练轮次，使用固定的学习率 $1e-4$，并在 Adam 优化器中将 beta 设为 0.5。默认间隙大小设为 240 毫秒，相当于 6 个网络数据包。为了在相同设置下客观评估不同模型的性能，间隙大小保持固定；不过，后文将展示使用表现最佳的模型进行可变间隔大小实验的结果。

D. 评估指标

我们从梅尔频谱图的图像角度以及重建 WAV 音频的音频角度进行评估。我们采用三种评估指标。首先，计算重建图像与目标图像之间像素的均方误差（MSE）。其次，测量梅尔频谱图中 VGG19 卷积神经网络特征提取层的均方误差，并比较真实值与生成的数据结构。我们更倾向于使用 VGG19 特征均方误差指标，而非 L1 损失指标，因为前者在视觉上更具描述性，但表 IV 除外，该表完整呈现了评估结果。由于图像比较指标无法明确反映图像修复生成的音频在听感上与真实音频的接近程度，因此，第三种指标采用语音质量感知评估（PESQ） [13] 来衡量生成音频的质量，并针对每个测试模型分别计算该指标。PESQ 是一种广泛使用的标准，用于自动评估电信系统中的语音质量。它以两个音频样本作为输入，并生成一个范围为 1（最差）至 5（最好）的平均意见分（MOS）。

IV. 实验结果

报告的首批结果与数据归一化有关。我们对 U-Net 架构进行的初步测试表明，若不对梅尔频谱图进行归一化，模型便无法学习有效模式，也无法生成有意义的结果。我们采用了 III-B 节中所述的一系列归一化技术。模型的评估结果汇总于表 I。

使用归一化数据进行图像修复的基线方法表明，该算法能够学习梅尔频谱图的结构，并修复缺失部分，平均意见分为 2.348。然而，其性能尚未达到实际应用所需的水平。我们尝试将 U-Net 步幅调整为 1，以使梅尔频谱图的尺寸更接近 256×80 像素；具体做法是在编码器-解码器连接层中进行调整。然而，早期特征图尺寸迅速缩小，导致性能下降，平均意见分降至 2.138。我们发现，只需进行最小限度的结构调整，输入尺寸为 256×128 时的图像修复效果便可与原始 256×256 的效果相当。因此，后续所有 U-Net 模型的输入尺寸均设为 256×128。

借鉴近期的图像修复方法，例如 [11], [19],，我们发现较新的方法采用了更复杂的损失函数，而且调整损失函数可能提升性能。随后，我们改进了损失函数，具体做法是对整幅图像使用 VGG19 特征匹配损失，并加入 L1 损失。修复后的图像更接近真实数据分布，VGG19 特征匹配误差也从 6.056 大幅降至 2.896，平均意见分则从 2.348 提升至 3.657。我们还实现了 [11] 中提出的一种思路：将图像修复区域的额外损失计入整体误差。因此，我们加入了图像修复区域损失，以使算法更专注于该区域。这使 VGG19 损失进一步降至 2.721，平均意见分提升至 3.737。

为了了解梅尔谱图的长度是否会影响对遮蔽片段的预测，我们将梅尔频谱图长度减半，把输入尺寸从 256px（2.8s）缩小至 125px（1.4s），从而降低问题的复杂度和计算成本。训练和测试结果表明，减少输入数据显著降低了模型性能。VGG19 特征匹配分数从 6.056（基线）下降至 9.962，这表明基线算法利用了整个梅尔谱图中的信息。我们还进行了增加数据维度的实验，但这会带来额外的计算成本，因此不在本文讨论范围之内。

除 U-Net 生成器外，我们还使用 GMCCN CNN 架构进行了实验。经过标准的 40 个训练周期后，GMCCN 的表现令人失望，VGG19 特征匹配损失为 3.402，平均意见分也显著降至 2.465。此外，GMCCN 的计算成本更高，因为该架构包含 3 个并行运行的网络，如前文图 2 所示。

为了研究遮蔽间隙大小的重要性，我们基于以下假设开展了实验：该算法需要重新生成从 40ms 到 320ms 的间隙。因此，我们针对不同的间隙大小训练了模型。结果表明，如图 3 所示，减小间隙大小可以减少获得良好性能所需的训练时间。一个有趣的发现是，当在 320ms 间隙模型上生成梅尔频谱图，并在 160ms 间隙模型上生成其他梅尔频谱图时，如果我们取样本的前 160ms，则 320ms 间隙模型梅尔频谱图上的误差相同。这说明，对于相同的间隙窗口，模型的表现相同。后续实验均在间隙为 320ms 的片段上进行。结果还表明，性能呈线性下降，模型在开始时能够以较高置信度重新生成梅尔频谱图；然而，越深入时域，结果的准确性越低，如表 II 所示。

我们还尝试使用可变间隔大小训练模型。在数据处理流水线中，随机间隔选择的范围为40毫秒至320毫秒。模型经过默认的40个训练轮次后，我们发现这些模型的表现明显不如固定间隔模型，且与固定大小模型相比，性能下降了70%。我们使用包含13,000个样本的完整数据集训练模型（将步长从 40×1000 增加到 $40 \times$13,000）。即使如此，该模型的表现仍不如采用

表 I 模型输出指标汇总

指标	VGG19 特征损失	L1 损失	MOS
U-Net 基线 256x128	6.056	0.415	2.348
U-Net 256×80	7.716	0.503	2.138
U-Net 128x128	9.962	0.636	2.029
采用 VGG19损失的 U-Net	2.896	0.249	3.657
采用 VGG19损失和 +分块损失的 U-Net	2.721	0.238	3.737
GMCCN	3.402	0.255	2.465

表 II 不同掩码尺寸的评估结果（以 40 毫秒的倍数计）

	1	2	3	4	5	6	7	8
MOS	4.514	4.321	4.185	4.074	3.938	3.832	3.520	3.214
VGG19损失	0.501	0.989	1.338	1.832	2.314	2.762	3.138	3.738

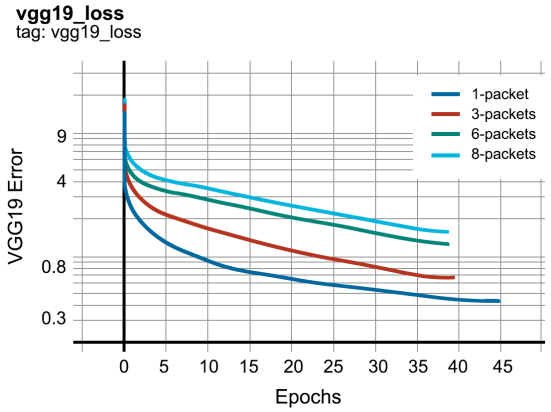


图 3. 不同间隙大小设置下的训练损失

固定间隙大小的模型；其 VGG19 特征损失为6.785，明显高于固定间隙模型的2.721，尽管训练所用的数据集规模大得多。

我们对基于 Parallel-WaveGAN 和 U-Net 的生成器进行了一系列推理速度测试，以确定该模型是否适用于实时场景。基于 U-Net 的模型可在 GPU 上约 50 毫秒内生成修复后的梅尔频谱图，与 [6] 中报告的结果一致。Parallel-WaveGAN 可在 GPU 上以 5 毫秒将梅尔谱图转换为音频，与 [20] 中报告的结果一致。

最后，我们检查了表现最差的修复后梅尔频谱图，以及根据 VGG19 特征损失和平均意见分评定的表现最佳的梅尔频谱图。结果汇总见表 III，图 4 展示了一些代表性示例。表 IV 展示了模型输出对比示例，以及真实值和作为模型输入的梅尔频谱图。

V. 结论

本文提出了一种技术，用于改善因引入可变长度间隙而退化的语音信号；此类间隙在实时音频电话中经常出现，且

表 III 梅尔谱图结果最佳与最差的情况

	MOS	VGG19损失
平均意见分表现最佳	4.792	0.250
基于 VGG19 / 均方误差的最佳表现	4.592	0.238
平均意见分表现最差的	2.735	3.871
VGG19 / 均方误差表现最差	2.935	5.303

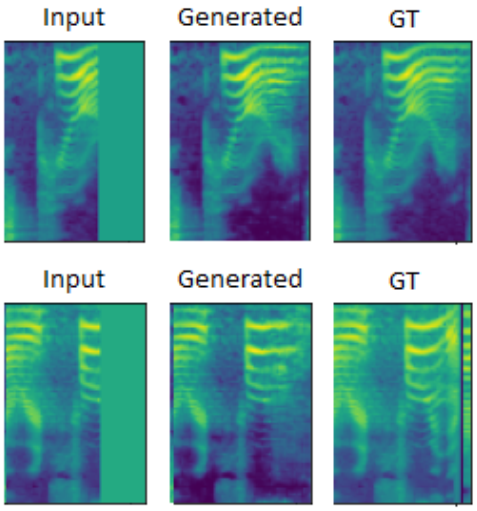
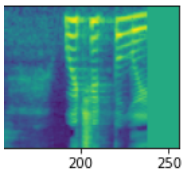
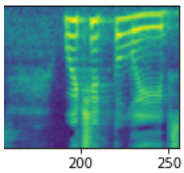
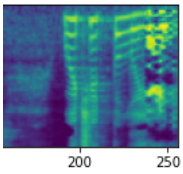
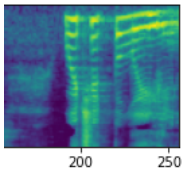
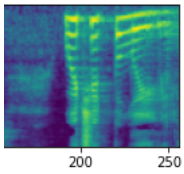
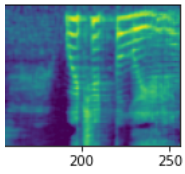


图 4. 根据 VGG19 特征均方误差确定的表现最佳（左）和最差（右）的 Mel 频谱图示例。表现最差的样本中，真实值频谱图里的黑色竖线表示 LJSpeech 数据集样本音频中的一处缺失片段。

视频会议。目前没有适用于此任务的良好解决方案，因此使用生成对抗网络生成缺失的音频间隙，这也使得与以往工作的比较变得困难。我们的主要发现是，基于 U-Net 的生成对抗网络采用基于 VGG19 特征匹配的损失函数 [19]，能够以近乎实时的速度修复音频梅尔频谱图中的间隙。将修复后的梅尔频谱图通过 Parallel-

表 IV 模型输出对比，梅尔频谱图（已调整尺寸以适应版面）

输入	真实值		
			
U-Net 基线	使用 VGG19损失	使用 VGG19损失 + 分块损失	GMCNN
			

WaveGanvocoder [20]转换回音频，并采用损失函数更先进的增强型 U-net 生成器（类似于 [11]，[19]，中的方法）后，我们生成的音频片段在结构上与真实分布相似；对于 320 毫秒的间隙，平均意见分为 3.214；对于 40 毫秒的间隙，平均意见分达到 4.514。在 GPU 上重新生成一个间隙所需的总时间约为 105 毫秒，对于实时通信而言，这一性能可以接受。

对于较大的重建间隙，我们的模型能够在梅尔谱图中几乎精确地重建缺失区域。该模型利用梅尔谱图中的全部信息，因为减小输入梅尔谱图的大小会导致性能大幅下降。我们发现，固定间隙大小模型能够从较小的数据集中学习分布，因为问题的复杂度降低了；而处理可变间隔大小的最有效方法是训练一个能够填补大间隙的模型，并将其用于所有间隙大小。这种方法的性能与在较小间隙大小上训练的模型相近。

我们得出结论：在配备 GPU 的系统上，可以利用我们的修复器-声码器流水线实时修复音频信号中的缺失片段，并且修复结果在人类听感上具有良好的质量。未来的研究应进一步探索是否存在类似于 SD-UNET [6] 的轻量化模型，使其能够在仅配备 CPU 的系统上达到足够好的性能。

参考文献

[1] Alqahtani, H., Kavakli-Thorne, M., Kumar, G.：生成对抗网络（GAN）的应用：最新综述。《计算工程方法档案》**28**(2)，525–552（2021）[2] Caelles, S., Pumarola, A., Moreno-Noguer, F., Sanfeliu, A., Van Gool, L.：利用时空生成对抗网络实现快速视频对象分割。arXiv 预印本 arXiv:1903.12161（2019）[3] Croce, D., Castellucci, G., Basili, R.：Gan-bert：利用少量标注样本进行稳健文本分类的生成对抗学习。载于：第58届 ACL 年会。第 2114–2119 页（2020）

[4] Donahue, C., McAuley, J., Puckette, M.：使用生成对抗网络合成音频。https://openreview.net/forum?id=r1RwYIJPM（2018），最后访问日期：2023年2月16日[5] Durak, L., Arikan, O.：短时傅里叶变换：两项基本性质及一种最优实现。电气与电子工程师学会信号处理汇刊**51**(5), 1231–1242（2003）[6] Gadosey, P.K., Li, Y., Agyekum, E.A., Zhang, T., Liu, Z., Yamak, P.T., Essaf, F.：SD-UNET：精简 U-Net，用于在计算资源预算有限的平台上进行生物医学图像分割。《诊断学》**10**(2), 110（2020）[7] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.：生成对抗网络。神经信息处理系统进展 **27**（2014）[8] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.：生成对抗网络。《美国计算机协会通讯》**63**(11), 139–144（2020）[9] Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.：利用条件生成对抗网络实现图像到图像转换。载于：电气与电子工程师学会计算机视觉与模式识别会议。第1125–1134页（2017）[10] Ito, K., Johnson, L.：LJ 语音数据集。https://keithito.com/LJ-Speech-Dataset/（2017 11 2023），最后访问日期：-[11] Feb, Jam, J., Kendrick, C., Drouard, V., Walker, K., Hsu, G.S., Yap, M.H.：R-mnet：一种用于图像修复的感知对抗网络。载于：IEEE/CVF 冬季计算机视觉应用会议。第2714–2723页（2021）[12] Kim, H.Y., Yoon, J.W., Cheon, S.J., Kang, W.H., Kim, N.S.：一种基于 GAN 的语音增强多分辨率方法。《应用科学》**11**(2), 721（2021）[13] de Lima, A.A., Freeland, F.P., de Jesus, R.A., Bispo, B.C., Biscainho, L.W., Netto, S.L., Said, A., Kalker, A., Schafer, R., Lee, B., 等：关于声音信号的质量评估。载于：电气与电子工程师学会国际电路与系统研讨会。第416–419页（2008）[14] Loizou, P.C.：语音增强：理论与实践。CRC出版社（2007）[15] Mirza, M., Osindero, S.：条件生成对抗网络。arXiv预印本 arXiv:1411.1784（2014）[16] Pascual, S., Bonafonte, A., Serra, J.：SEGAN：语音增强生成对抗网络。arXiv预印本 arXiv:1703.09452（2017）[17] Phan, H., McLoughlin, I.V., Pham, L., Ch´en, O.Y., Koch, P., De Vos, M., Mertins, A.：改进用于语音增强的 GAN。电气与电子工程师学会信号处理快报**27**, 1700–1704（2020）[18] Radford, A., Metz, L., Chintala, S.：利用深度卷积生成对抗网络进行无监督表征学习。arXiv预印本 arXiv:1511.06434（2015）[19] Wang, Y., Tao, X., Qi, X., Shen, X., Jia, J.：通过生成式多列卷积神经网络实现图像修复。载于：神经信息处理系统进展。第331–340页（2018）

[20] Yamamoto, R., Song, E., Kim, J.M.：Parallel WaveGAN：一种基于生成对抗网络和多分辨率频谱图的快速波形生成模型。载于：国际声学、语音与信号处理会议（ICASSP）。第6199–6203页（2020） [21] Yuan, J., Liberman, M., Cieri, C.：探索对话中语速的综合理解。载于：第九届国际口语处理会议（2006） [22] Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.：视觉增强深度音频修复。载于：电气与电子工程师学会/CVF计算机视觉会议论文集。第283–292页（2019）