arXiv:2602.05443v1 [eess.AS] 5 Feb 2026

WAVE-TRAINER-FIT：采用可训练先验和不动点迭代、基于SSL特征实现高质量语音生成的神经声码器

Hien Ohnaka[1,*]，*Yuma Shirahata*[2]，*Masaya Kawamura*[2]

[1]奈良先端科学技术大学院大学，奈良，日本[2]LY Corporation，东京，日本

摘要

我们提出了WaveTrainerFit，这是一种神经声码器，可根据SSL特征等数据驱动特征生成高质量波形。WaveTrainerFit建立在WaveFit声码器之上，后者融合了扩散模型和生成对抗网络。此外，所提出的方法还进行了以下关键改进：1. 通过引入可训练先验，推理过程从接近目标语音的噪声开始，而不是从高斯噪声开始。2. 通过对可训练先验施加约束，使其匹配语音能量，从而实现参考感知增益调整。这些改进有望降低从数据驱动特征进行波形建模的复杂度，使模型能够以更少的推理步数生成高质量波形。通过实验，我们证明了WaveTrainerFit能够根据数据驱动特征生成高度自然且说话人相似度更高的波形，同时所需的迭代次数少于WaveFit。此外，我们还证明，所提出的方法对于提取SSL特征的网络深度具有较强的鲁棒性。代码和预训练模型可从 https://github.com/line/WaveTrainerFit获取。

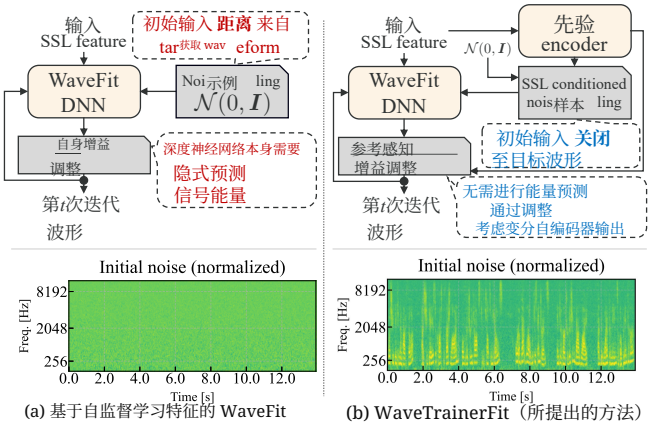
索引词— 神经声码器、自监督学习、变分自编码器、语音合成。

1. 引言

神经声码器是能够根据声学特征生成语音波形的深度神经网络（DNN）。近年来，它们已成为文本转语音 [1, 2], 语音转换 [3], 和语音到语音翻译 [4]等多种任务中不可或缺的组成部分。具体而言，基于梅尔频谱图的神经声码器已得到广泛研究，并且已能够生成与人类语音质量相当的波形 [5, 6]。

近年来，自监督学习（SSL）模型 [7, 8, 9] 改变了传统语音处理任务的实现方式。SSL 模型在文本转语音合成 [10, 11], 和语音修复 [12, 13]等生成应用中展现出优异性能。它们的优势在于能够从大量数据中学习表示语音信号的数据驱动特征大量未标注数据。鉴于这一背景，基于自监督学习特征的鲁棒声码器正变得越来越重要。

WaveFit [5, 12] 是一种固定点迭代声码器，将生成对抗网络（GAN） [6, 14, 15, 16] 与类似扩散模型 [17, 18]的推理相结合。由于结合了这两种生成模型，WaveFit 不仅能够根据梅尔频谱图生成高质量波形，还能根据并非专为波形生成任务 [12, 13, 19, 20] 设计的自监督学习特征生成高质量波形。



(a) 基于自监督学习特征的 WaveFit (b) WaveTrainerFit（所提出的方法）

图 1。各方法的概念示意图和噪声示例。底部图像显示了初始噪声的对数梅尔频谱图。

将梅尔频谱图用作 WaveFit [5], 的输入时，可以运用信号处理知识，从两个方面有效训练扩散模型：1. 初始噪声可以根据频谱包络 [17],从人工设计的先验分布中采样，而不是使用高斯噪声。2. 根据梅尔频谱图的功率显式进行增益调整。然而，当使用自监督学习特征作为输入时 [12], , 由于这些特征并非源自信号处理，因此无法采用上述方法。

为克服这些局限，我们提出了一种神经声码器，**具有可**训练先验**和**定点**迭**代-（**WaveTrainerFit**），~~以提升基于~~自监督学习特征的波形生成效果。图 1 展示了 WaveTrainerFit 相较于 WaveFit 的关键改进。首先，通过引入基于变分自编码器（VAE） [21]的可训练先验，我们能够采样出接近目标波形的噪声。由于推理可以从接近语音的位置开始，因此有望以更少的迭代次数生成高质量波形。此外，通过对先验施加约束，使其学习语音的能量，我们实现了利用先验能量进行增益调整，从而使声码器不再需要隐式地推断能量。因此，模型能够专注于波形建模中更重要的方面，并有望缓解训练难度。

客观评估表明，与 WaveFit 相比，WaveTrainerFit 只需更少的迭代次数即可实现更高质量的波形生成。此外，即使以包含有限声学信息的深层 SSL特征为条件，所提出的方法也能生成自然的波形。最后，主观评估表明，所提出的方法在说话人相似度方面优于基线。音频样本可在我们的演示页面 [1] 上获取。

*在 LY Corporation 兼职期间完成的工作。

[1] https://i17oonaka-h.github.io/projects/research_topics/wave_trainer_fit/

2. 相关工作

2.1. WaveFit：采用不动点迭代的神经声码器 ation

WaveFit [5, 12] 是一种迭代式非自回归神经声码器，将扩散模型的迭代处理与基于生成对抗网络的损失相结合。虽然 WaveFit 有同时接收梅尔频谱图和 SSL特征作为输入的变体，但本文侧重于后者。该方法通过应用去噪映射过程 [12]，根据高斯噪声 $\boldsymbol{y}_T \in \mathbb{R}^D \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ 和 SSL特征 $\boldsymbol{c}$ 生成语音波形 $\boldsymbol{y}_0 \in \mathbb{R}^D$ T：

$$\boldsymbol{y}_{t-1} = \hat{\mathcal{G}}(\boldsymbol{z}_t), \quad \boldsymbol{z}_t = \boldsymbol{y}_t - \mathcal{F}_\theta(\boldsymbol{y}_t, \boldsymbol{c}, t), \tag{1}$$

$$\hat{\mathcal{G}}(\boldsymbol{z}_t) = \beta_{\text{scale}} \cdot \boldsymbol{z}_t / \max(\text{abs}(\boldsymbol{z}_t)), \tag{2}$$

其中，D 表示时域采样点的数量，$\mathcal{F}_\theta$ 表示经过训练以估计噪声分量的深度神经网络，β_{scale} 表示缩放因子，abs(·)返回输入向量各元素的绝对值。$\hat{\mathcal{G}}(\boldsymbol{z}_t)$是一个自增益调整算子，用于对预测波形的功率进行归一化。损失函数 $\mathcal{L}^{\text{WF}}$ 表示为以下公式：

$$\mathcal{L}^{\text{WF}} = \frac{1}{T}\sum_{t=0}^{T-1} \mathcal{L}_{\text{G}}^{\text{gan}}(\boldsymbol{x}_0, \boldsymbol{y}_t) + \mathcal{L}_{\text{D}}^{\text{gan}}(\boldsymbol{x}_0, \boldsymbol{y}_t) + \lambda_{\text{S}}\mathcal{L}^{\text{S}}(\boldsymbol{x}_0, \boldsymbol{y}_t). \tag{3}$$

其中，$\boldsymbol{x}_0$ 是目标波形，$\mathcal{L}^{\text{gan}}{}_{\text{G}}$ 和 $\mathcal{L}^{\text{gan}}{}_{\text{D}}$ 分别是 [5]中生成器和判别器的损失函数，λ_{S} 是权重参数，$\mathcal{L}^{\text{S}}$ 是多分辨率短时傅里叶变换损失 [14, 15]。

WaveFit已成功应用于从SSL特征生成语音的声码器，并用于语音修复 [12, 13] 和高质量数据集创建 [19, 20]等任务。然而，由于WaveFit最初是为梅尔频谱图（一种基于信号处理的声学特征）设计的声码器，因此在使用数据驱动的自监督学习特征作为输入时会作出以下两项折衷：

1. 噪声采样。 在扩散模型的初始噪声采样中，无法使用基于梅尔频谱图中信号处理知识的手工设计噪声采样方法 [17, 22]。因此，它只能采用简单的高斯噪声。

2. 增益调整。 在增益调整中，无法利用从梅尔频谱图获得的功率。因此，必须像式（2）那样由模型学习。

2.2. RestoreGrad：具有可训练先验的扩散模型

RestoreGrad [23] 是一种方法，即使难以采用手工设计的噪声采样方法，也能为扩散模型获得信息丰富的先验。为此，RestoreGrad 将扩散模型与由变分自编码器建模的可训练先验相结合。具体而言，RestoreGrad 经过训练，以最小化后验分布 $\mathcal{N}$之间的 Kullback-Leibler（KL）散度

（$\mathbf{0}, \boldsymbol{\Sigma}_{\text{post}}$）和先验分布 $\mathcal{N}$（$\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}}$）：

$$\mathcal{L}^{\text{PM}}(\boldsymbol{\Sigma}_{\text{post}}, \boldsymbol{\Sigma}_{\text{prior}}) = \log \frac{|\boldsymbol{\Sigma}_{\text{prior}}|}{|\boldsymbol{\Sigma}_{\text{post}}|} + \text{tr}(\boldsymbol{\Sigma}_{\text{prior}}^{-1}\boldsymbol{\Sigma}_{\text{post}}). \tag{4}$$

$\boldsymbol{\Sigma}_{\text{prior}}$ 和 $\boldsymbol{\Sigma}_{\text{post}}$ 分别是由先验编码器 $\mathcal{V}_{\text{prior}}(\boldsymbol{c})$和后验编码器 $\mathcal{V}_{\text{post}}(\boldsymbol{c}, \boldsymbol{x}_0)$生成的协方差矩阵。$\boldsymbol{x}_0$ 是目标波形，$\boldsymbol{c}$ 是条件特征。为了帮助后验编码器学习信息表征，引入了额外的损失项 $\mathcal{L}^{\text{LR}}(\boldsymbol{x}_0, \boldsymbol{\Sigma}_{\text{post}})$：

$$\mathcal{L}^{\text{LR}}(\boldsymbol{x}_0, \boldsymbol{\Sigma}_{\text{post}}) = \log|\boldsymbol{\Sigma}_{\text{post}}| + \bar{\alpha}_T \boldsymbol{x}_0^T \boldsymbol{\Sigma}_{\text{post}}^{-1} \boldsymbol{x}_0, \tag{5}$$

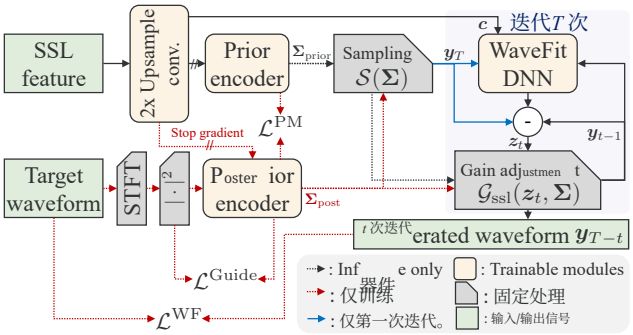


图 2。所提出模型的概览。训练期间，基于目标波形和 SSL特征得到的后验编码器用于噪声采样和增益调整。推理期间，基于 SSL特征得到的先验编码器用于相同的过程。实线箭头表示训练和推理期间均会使用。

其中，$\bar{\alpha}_T$ 是基于方差调度的权重。第一项是正则化项，用于防止编码器输出数值膨胀和训练崩溃。第二项则引导 $\boldsymbol{\Sigma}_{\text{post}}$ 的功率与目标波形 $\boldsymbol{x}_0$相近。

在扩散模型的推理步数中，模型可以从先验 $\mathcal{N}$（$\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}}$）中采样初始噪声。由于先验包含与目标波形相关的信息，因此与从高斯噪声采样这一朴素方法相比，预期它能够促进波形生成。

在实验中，与手工设计的方法 [22] 以及从标准正态分布采样相比，[23], RestoreGrad 在语音增强和图像复原方面表现更优。

3. 所提出的方法

3.1. 动机

我们推测，通过引入第 2.2 节所述 RestoreGrad 中的可训练先验，可以解决第 2.1 节所述 WaveFit 的噪声采样问题。此外，在可训练先验上施加新的约束，使其匹配目标语音的能量，即可实现参考感知增益调整，从而解决增益调整问题。这些改进有望降低从数据驱动特征进行波形建模的难度，使得仅需更少的推理步数即可实现高质量波形生成。

3.2. 模型概述

图 2 展示了所提出方法的概览。该方法引入了一个先验编码器和一个后验编码器，以实现如 RestoreGrad 中的可训练初始噪声采样。对于条件输入 c，我们采用转置二维卷积层将 SSL特征上采样两倍，具体方法如 [12, 13] 所示。$\boldsymbol{c}$ 用作后验编码器、先验编码器和 WaveFit 深度神经网络的输入。在训练期间，初始噪声从后验分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{post}})$ 中采样，该分布以 SSL特征和目标波形为条件。在推理期间，初始噪声从先验分布 $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\text{prior}})$ 中采样，该分布仅以 SSL特征为条件。

3.3. 时频域噪声采样

尽管 RestoreGrad 在波形时域中引入了可训练先验，但所提出的方法将其引入时频域

以缩短序列长度并降低建模复杂度。具体而言，从后验编码器或先验编码器获得的方差特征 $\mathbf{\Sigma}$ 被调整为形状为 $\mathbb{R}^{F\times K}$。其中， F是频率箱大小， K 是帧数。然后，使用公式 (6) 对时域初始噪声 $\boldsymbol{y}_T = \mathcal{S}(\mathbf{\Sigma}) \in \mathbb{R}^D$ 进行采样：

$$\mathcal{S}(\mathbf{\Sigma}) = \mathrm{iSTFT}(\mathcal{R}(\boldsymbol{N}) \odot \mathbf{\Sigma} + i\mathcal{I}(\boldsymbol{N}) \odot \mathbf{\Sigma}), \tag{6}$$

$$\boldsymbol{N} = \mathrm{STFT}(\epsilon) \in \mathbb{C}^{F\times K},\ \epsilon \in \mathbb{R}^D \sim \mathcal{N}(0, \boldsymbol{I}). \tag{7}$$

当假设均值和伪协方差均为零时，该过程对应于复数值变分自编码器中的重参数化技巧 [24]。 $\mathcal{R}(\cdot)$和 $\mathcal{I}(\cdot)$分别是仅提取实部和虚部的算子。

3.4. 损失函数和增益调整

所提出方法的损失函数如下：

$$\mathcal{L}^{\mathrm{TrainerFit}} = \mathcal{L}^{\mathrm{WF}} + \lambda_{\mathrm{PM}}\mathcal{L}^{\mathrm{PM}} + \mathcal{L}^{\mathrm{Guide}}. \tag{8}$$

这里，第一项与公式（3）中的损失函数相同，负责扩散和生成对抗网络的训练。第二项基于公式（4）中的损失项，但扩展到了时频域。该损失函数最小化先验编码器输出与后验编码器输出之间的KL散度。λ_{PM} 是权重参数。最后，第三项为后验编码器输出 $\mathbf{\Sigma}_{\mathrm{post}}$提供软引导：

$$\mathcal{L}^{\mathrm{Guide}} = \left|\mathcal{E}(\mathbf{\Sigma}_{\mathrm{post}}) - \mathcal{E}(|\boldsymbol{X}_0|^2)\right| + \frac{\lambda_{\mathrm{Guide}}}{FK}\sum_{f=0}^{F-1}\sum_{k=0}^{K-1}\frac{\mathbf{\Sigma}_{\mathrm{post}}[f,k]}{|\boldsymbol{X}_0|^2[f,k]}. \tag{9}$$

这里， λ_{Guide} 是权重参数， $|\boldsymbol{X}_0|^2 \in \mathbb{R}^{F\times K}$ 是目标波形的功率谱图，而 f, k 分别表示频率和时间索引， $\mathcal{E}(\cdot)$分别是逐元素求和操作。第一项旨在使后验编码器输出的能量与目标语音的能量相匹配。这促使 $\mathbf{\Sigma}_{\mathrm{post}}$ 的能量接近目标波形的能量。因此，我们可以将参考感知增益调整算子定义如下：

$$\mathcal{G}_{\mathrm{ssl}}(\boldsymbol{z}_t, \mathbf{\Sigma}) = \sqrt{(\mathcal{E}(\mathbf{\Sigma})/(\mathcal{E}(|\boldsymbol{Z}_t|^2) + s))}\boldsymbol{z}_t. \tag{10}$$

此处， s 是一个标量，用于避免除零。第二项是将式（5）中的第二项扩展至二维信号，用以通过柔和地反映目标频谱的功率来引导有效的后验学习。

4. 实验评估

4.1. 实验设置

数据集。我们使用了 LibriTTS-R 语料库 [19]。该语料库包含总计 585 小时、采样率为 24 kHz 的语音，以及来自 2,456 位说话人的文本配对。“train-clean-360”、“dev-clean”和“test-clean”分别用于训练、验证和测试。

模型与训练设置。我们的模型通过将自监督学习特征上采样 $480\times$ 来生成波形。具体如下：首先，如图 2 所示，使用一个二维转置卷积层进行 $2\times$ 上采样。然后，采用一种与 WaveFit [5]类似、上采样倍率为 $\{5, 4, 3, 2, 2\}$ 的架构来执行 $240\times$。模型以 8 的批量大小训练了 $400k$ 个训练步数。

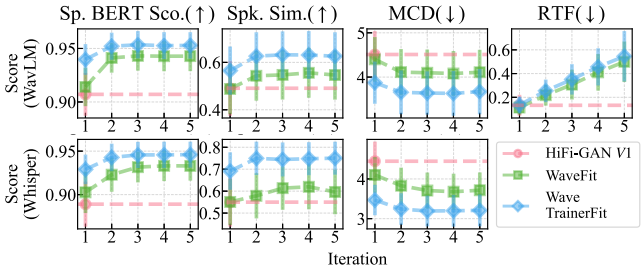


图3。客观指标随迭代次数变化的折线图。实时因子（RTF）是在“Intel(R) Xeon(R) Silver 4316 CPU @ 2.30GHz”上使用随机选取的 200 个样本测得的。

权重参数 λ_{Guide} 和 λ_{PM} 分别设为 0。1 和 10。编码器架构采用实值 DCUnet-10 [25]，并做了两项主要修改。首先，使用一个线性层使自监督学习特征的维度与功率谱图的频率维度对齐。其次，在后验编码器中定义了两个 U-Net [25]：一个用于目标波形的功率谱图，另一个用于自监督学习特征。第一个 U-Net 各模块的中间输出会与第二个 U-Net 相同位置的输出相加。后验编码器和先验编码器的通道数分别为 32 和 45。参数总数分别为 $2.61\ M$和 $2.59\ M$。优化器、调度器和判别器等其他设置与开源实现 [2]相同。

对比方法。为评估所提出方法的有效性，我们设置了两个基线。除 WaveFit [12],外，我们还将 HiFi-GAN [6] 设为基线，以明确 WaveFit 架构本身在从自监督学习特征进行波形生成方面的有效性。为使上采样比例与 240 对齐，我们将上采样率改为 $\{8, 5, 3, 2\}$，并将上采样卷积核尺寸改为 $\{16, 11, 7, 4\}$。生成器的参数总数为 $17.18\ M$。该模型训练了 $1M$ 步，批量大小与所提出的方法相同。

评估指标。我们采用了以下客观指标：SpeechBERTScore [26] 是一种自动语音评估指标，通过自监督学习特征的 BERTScore [27] 衡量生成语音与参考语音之间的语义一致性。计算梅尔倒谱失真（MCD）以评估频谱失真。由 [28] 计算对数基频均方根误差，以评估基频准确度。上述三项指标均使用开源实现 [3]计算。最后，说话人相似度以目标波形与预测波形之间说话人嵌入 [29, 30] 的余弦相似度来衡量。为评估所生成波形的质量，我们进行了两项主观听测。每种方法共450个样本由15名参与者按五分制进行评估。为消除所有评估中波形能量差异的影响，所有波形均与干净语音的能量对齐。

条件特征。我们使用了两种自监督学习特征和一种由 ASR 驱动的特征，作为声码器的条件特征：**1.WavLM-large** [7]是一种通过掩码语言建模目标进行预训练的英语专用自监督学习模型。**2. XLS-R-0.3B** [8] 是一种通过对比学习目标进行预训练的多语言自监督学习模型。**3. Whisper-medium** [31] 是一种多语言语音转文本模型。在此，我们使用 Whisper 的编码器部分作为特征提取器，类似于 [32, 33]。这些特征均来自不同的训练

[2]https://github.com/yukara-ikemiya/wavefit-pytorch[3] https://github.com/Takaaki-Saeki/DiscreteSpeechMetrics

表1。使用 LibriTTS-R test-clean、第8层自监督学习特征以及 $T = 5$ 时的评估结果。**粗体** 表示相同条件下的最佳方法，下划线表示 WaveFit 与 WaveTrainerFit 之间存在显著差异。

方法		客观指标				主观指标	
自监督学习特征	模型	语音 BERT 得分（↑，%）	梅尔倒谱失真（↓）	对数基频均方根误差（↓）	说话人相似度（↑，%）	N-MOS (↑)	S-MOS (↑)
干净语音		(100.0)	(0.0)	(0.0)	(100.0)	3.92 ± 0.10	–
WavLM	HiFi-GAN $V1$	90.71	4.510	0.1972	49.09	2.39 ± 0.12	2.81 ± 0.12
	WaveFit	94.28	4.109	0.1956	54.67	$\underline{\mathbf{3.76 \pm 0.11}}$	3.02 ± 0.12
	WaveTrainerFit	**95.28**	**3.672**	**0.1810**	**62.61**	3.50 ± 0.11	$\underline{\mathbf{3.38 \pm 0.11}}$
XLS-R	HiFi-GAN $V1$	91.09	4.424	0.1887	51.96	2.54 ± 0.12	2.99 ± 0.12
	WaveFit	94.11	4.196	0.1934	52.78	$\underline{\mathbf{3.79 \pm 0.11}}$	3.04 ± 0.12
	WaveTrainerFit	**94.39**	**4.089**	**0.1762**	**55.54**	3.21 ± 0.12	$\mathbf{3.13 \pm 0.12}$
Whisper 编码器	HiFi-GAN $V1$	88.90	4.446	0.1843	54.98	2.41 ± 0.12	2.96 ± 0.12
	WaveFit	93.30	3.715	0.1695	59.64	3.23 ± 0.12	3.56 ± 0.11
	WaveTrainerFit	**94.60**	**3.208**	**0.1690**	**75.02**	$\underline{\mathbf{3.87 \pm 0.10}}$	$\underline{\mathbf{4.19 \pm 0.09}}$

表2。层级聚焦目标评估结果。此处，我们使用 WavLM 特征作为条件特征。

模型-层	Sp. BERT 分数	MCD	Log-F0 RMSE	Spk. Sim.	UTMOS (↑)
WaveFit-2	95.95	**2.852**	**0.1660**	78.99	4.230
WaveTrainerFit-2	**96.13**	2.860	0.1662	**79.65**	4.203
WaveFit-8	94.28	4.109	0.1956	54.67	4.194
WaveTrainerFit-8	**95.28**	**3.672**	**0.1810**	**62.61**	4.160
WaveFit-24	89.08	5.501	0.3623	35.09	<u>3.681</u>
WaveTrainerFit-24	**92.19**	**5.066**	**0.2301**	**40.76**	**<u>4.198</u>**

目标和预训练数据，因此预计具有不同的特性。

4.2. 实验结果

4.2.1. 客观与主观评估结果

表 1 展示了评估结果。可以看出，无论使用哪种自监督学习特征，所提方法在所有参考感知客观指标上均优于基线。这表明，所提的初始噪声采样和结合可训练先验的增益调整有效提升了客观指标。在 S-MOS 方面，对于所有自监督学习特征，所提方法同样优于基线。原因如下：目标语音的频谱特性通过先验编码器中的公式 (9) 第二项得到学习。因此，可以采样到包含基频信息的初始噪声，如图 1 (b) 所示，由此生成的波形能够充分捕捉说话人的特征。在 N-MOS 方面，所提方法在使用 Whisper编码器时优于基线，但使用 XLS-R 时则表现较差。造成这一结果的原因可能有几种。一种可能是，所提方法对超参数较为敏感。我们的方法基于 RestoreGrad，而 RestoreGrad 表现出对 $\mathcal{L}^{\mathrm{LR}}$ [23] 的权重参数有一定敏感性。受计算成本限制，我们只测试了在 RestoreGrad 中得分最高的权重，但仔细调整这些权重或许能带来更好的结果。另一种可能是 XLS-R 特征本身产生了负面影响。XLS-R 在英语自动语音识别任务中的表现逊于基线，这表明模型 [8] 内存在语言干扰。我们的实验中可能也出现了类似现象，从而导致性能下降。未来一项重要工作是详细分析这些超参数、语言、自监督学习特征及其他因素之间的组合如何影响性能。

4.2.2. 各次迭代的性能与处理速度

图 3 展示了每次推理迭代的客观指标。在推理过程中，所提方法的实时因子（RTF）因先验编码器和采样过程而略有增加；尽管如此，除 RTF 外，该方法在所有迭代中的各项指标上均表现更优。具体而言，在第 1 次迭代时，SpeechBERTScore 的提升幅度最大。这些结果表明，与使用相同迭代次数的基线相比，从接近目标波形的初始噪声开始推理，能够带来更好的波形生成效果。

4.2.3. 不同层的 *SSL* 特征性能

众所周知，SSL 模型浅层的特征包含大量输入中的声学信息，而深层的特征则包含大量语义信息 [34]。为了验证所提方法在以不同属性的特征作为条件时的鲁棒性，我们使用从 WavLM 第2层、第8层和第24层提取的特征进行了评估。在此实验中，我们还加入了无参考UTMOS [35] 作为评估指标，以排除目标波形中的说话人特征偏差。

结果如表2所示。可以看出，除第2层的两项指标外，所提方法在所有参考感知指标上均优于基线。值得注意的是，以声学信息有限的第24层特征为条件时，所提方法不仅在参考感知指标上取得了显著提升，同时也未导致 UTMOS 明显下降。这表明，所提方法能够生成与参考语音更加一致的波形，因此对条件特征的变化表现出更强的鲁棒性。

5. 结论

我们提出了 WaveTrainerFit，用于从自监督学习特征生成高质量波形。通过引入基于VAE的可训练先验，我们实现了更优的初始噪声采样和基于参考的增益调整，解决了传统 WaveFit 在使用自监督学习特征时存在的问题。在评估实验中，我们证明，所提方法能够从缺乏声学信息的特征中生成高质量波形，且所需迭代次数更少。在主观评估中，我们确认，与基线相比，所提方法的说话人相似度MOS始终更优。未来工作包括利用大型数据集构建模型，并将其扩展至多语言模型。

6. 参考文献

[1] Y. Wang、R. J. Skerry-Ryan 等，“Tacotron：迈向端到端语音合成”，载于 Interspeech 论文集，2017年，第 4006–4010 页。[2] Y. Ren、C. Hu 等，“FastSpeech 2：快速、高质量的端到端文本转语音”，载于 ICLR 论文集，2021年。[3] B. Sisman、J. Yamagishi 等，“语音转换及其挑战概述：从统计建模到深度学习”，IEEE ACM 音频、语音与语言处理汇刊，第 29 卷，第 132–157 页，2021年。[4] Y. Jia、M. T. Ramanovich 等，“Translatotron 2：具有声音保留功能的高质量直接语音到语音翻译”，载于 ICML 论文集，第 162 卷，2022年，第 10 120–10 134 页。[5] Y. Koizumi、K. Yatabe 等，“WaveFit：一种基于不动点迭代的迭代式非自回归神经声码器”，载于 IEEE SLT 论文集，2022年，第 884–891 页。[6] J. Kong、J. Kim 等，“HiFi-GAN：用于高效、高保真语音合成的生成对抗网络”，载于 NeurIPS 论文集，2020年。[7] S. Chen、C. Wang 等，“WavLM：面向全栈语音处理的大规模自监督预训练”，IEEE 信号处理精选主题期刊，第 16 卷，第 6 期，第 1505–1518 页，2022年。[8] A. Babu、C. Wang 等，“XLS-R：大规模自监督跨语言语音表征学习”，载于 Interspeech 论文集，2022年，第 2278–2282 页。[9] W. Hsu、B. Bolte 等，“HuBERT：通过对隐藏单元进行掩码预测实现自监督语音表征学习”，IEEE ACM 音频、语音与语言处理汇刊，第 29 卷，第 3451–3460 页，2021年。[10] S. Wang、G. E. Henter 等，“朗读与自发语音文本转语音中的自监督语音表征比较研究”，载于 ICASSP SASB 研讨会论文集，2023年，第 1–5 页。[11] T. Saeki、G. Wang 等，“将多语言语音合成扩展至缺乏转录数据的 100+ 种语言”，载于 IEEE ICASSP 论文集，2024年，第 11 546–11 550 页。[12] Y. Koizumi、H. Zen 等，“Miipher：融合自监督语音与文本表征的鲁棒语音修复模型”，载于 IEEE WASPAA 论文集，2023年，第 1–5 页。[13] S. Karita、Y. Koizumi 等，“Miipher-2：面向百万小时级数据修复的通用语音修复模型”，载于 IEEE WASPAA 论文集（已录用），2025年。[14] R. Yamamoto、E. Song 等，“Parallel WaveGAN：一种基于多分辨率频谱图的快速波形生成模型”，载于 IEEE ICASSP 论文集，2020年，第 6199–6203 页。[15] G. Yang、S. Yang 等，“Multi-band MelGAN：用于高质量文本转语音的快速波形生成”，载于 IEEE SLT 论文集，2021年，第 492–498 页。[16] K. Kumar、R. Kumar 等，“MelGAN：用于条件波形合成的生成对抗网络”，载于 NeurIPS 论文集，2019年，第 14 881–14 892 页。[17] Y. Koizumi、H. Zen 等，“SpecGrad：基于扩散概率模型并采用自适应噪声频谱整形的神经声码器”，载于 Interspeech 论文集，2022年，第 803–807 页。[18] N. Chen、Y. Zhang 等，“WaveGrad：用于波形生成的梯度估计”，载于 ICLR 论文集，2021年。

[19] Y. Koizumi、H. Zen 等，《LibriTTS-R：一个修复后的多说话人文本转语音语料库》，载于 Interspeech 论文集，2023年，第 5496–5500 页。[20] M. Ma、Y. Koizumi 等，《FLEURS-R：一个用于生成任务的修复后多语言语音语料库》，载于 Interspeech 论文集，2024年。[21] D. P. Kingma 和 M. Welling，《自编码变分贝叶斯》，载于 ICLR 论文集，2014年。[22] S. Lee、H. Kim 等，《PriorGrad：通过数据依赖自适应先验改进条件去噪扩散模型》，载于 ICLR 论文集，2022年。[23] C. H. Lee、C. Yang 等，《RestoreGrad：使用联合学习的先验和条件去噪扩散模型进行信号恢复》，载于 ICML 论文集，2025年。[24] T. Nakashika，《复值变分自编码器：一种用于直接表示复数频谱的新型深度生成模型》，载于 Interspeech 论文集，2020年，第 2002–2006 页。[25] H. Choi、J. Kim 等，《使用深度复数 U-Net 的相位感知语音增强》，载于 ICLR 论文集，2019年。[26] T. Saeki、S. Maiti 等，《SpeechBERTScore：利用自然语言处理评估指标进行具备参考感知能力的语音生成自动评估》，载于 Interspeech 论文集，2024年。[27] T. Zhang、V. Kishore 等，《BERTScore：使用 BERT 评估文本生成》，载于 ICLR 论文集，2020年。[28] M. Morise，《Harvest：一种从语音信号中估计基频的高性能方法》，载于 Interspeech 论文集，2017年，第 2321–2325 页。[29] B. Desplanques、J. Thienpondt 等，《ECAPA-TDNN：在 TDNN 说话人验证中强调通道注意力、传播与聚合》，载于 Interspeech 论文集，2020年，第 3830–3834 页。[30] M. Ravanelli、T. Parcollet 等，《SpeechBrain：通用语音工具包》，2021年，arXiv:2106.04624。[31] A. Radford、J. W. Kim 等，《通过大规模弱监督实现鲁棒语音识别》，载于 ICML 论文集，第 202 卷，2023年，第 28 492–28 518 页。[32] R. E. Zezario、Y. Chen 等，《关于将 Whisper 用于鲁棒语音评估的研究》，载于 IEEE ICME 论文集，2024年，第 1–6 页。[33] L. Qian 和 G. Skantze，《口语对话中上下文与反馈嵌入的联合学习》，载于 Interspeech 论文集，2024年。[34] A. Pasad、J. Chou 等，《自监督语音表征模型的逐层分析》，载于 IEEE ASRU 论文集，2021年，第 914–921 页。[35] T. Saeki、D. Xin 等，《UTMOS：utokyo-sarulab 的 VoiceMOS 挑战赛 2022 系统》，载于 Interspeech 论文集，2022年，第 4521–4525 页。