> NIPS 2022，高丽大学

1. 提出 HierSpeech，采用自监督语音表征的、基于 hierarchical VAE 的高质量端到端 TTS
2. 从文本中直接推理 linguistic 和 acoustic 属性有一定的挑战，尤其是 linguistic 属性，会导致合成语音有发音错误和过度平滑问题
3. 采用自监督语音表征作为 linguistic representations，用 hierarchical conditional VAE 连接表征 latent representations 中的 linguistic 能力来分层地（hierarchically）学习每个属性
4. 相比于 SOTA TTS 系统，HierSpeech 在 VCTK 数据集上的 CMOS 为 +0.303，合成语音的 phoneme error rate 从 9.16% 降低到 5.78%
5. 进一步扩展到 HierSpeech-U，一个无文本的 TTS 系统，可以利用自监督语音表征来适应新的说话人

> 确实是一个很有意思的思路，不直接从文本中学习 acoustic feature，而是中间再经过一层 自监督的特征，且这个自监督的特征是来自于语音自监督模型从语音中提取的，和 acoustic feature 算是某种并列的特征，但是层级地建模。

Introduction

1. 之前的 TTS 存在两个限制：
    i. 语音由多种属性组成（发音、节奏、语调、音色），之前的模型大多数一次性从文本序列中合成 acoustic features，导致 one-to-many mapping 问题更加严重
    ii. 两阶段的 pipeline 中，TTS 系统的每个部分都需要独立训练，导致音频质量下降
2. 单阶段端到端 TTS 模型，直接从文本中生成原始波形，成功地减少了两阶段 pipeline 的限制，但是仍然存在发音错误和过度平滑问题，因为在合成语音的过程中，它们仍然一次性从文本序列中生成所有 acoustic attributes，因此，缺少一些属性的细节，尤其是 linguistic 信息
3. 本文采用自监督语音表征作为额外的 linguistic representations
4. 提出 HierSpeech，基于 hierarchical conditional VAE 的高质量端到端 TTS，采用自监督语音表征来丰富 latent representations 中的 linguistic 信息，从 linguistic representations 到 acoustic representations 分层地学习每个属性
5. 基于预训练的 HierSpeech，还提出了新的自适应 TTS 框架 HierSpeech-U，可以在无文本的语音数据下适应到新说话人，利用自监督语音表征来提取 linguistic representations，从无文本语音数据中学习 acoustic representations
6. 贡献如下：
    i. 提出 HierSpeech，采用自监督语音表征的、基于 hierarchical VAE 的高质量端到端 TTS
    ii. 调研了自监督语音表征在 TTS 系统中的应用，进行了超过 30,000 GPU 小时的实验
    iii. 扩展 HierSpeech 到 HierSpeech-U，一个无文本的 TTS 系统，可以利用自监督语音表征来适应新的说话人

HierSpeech

Speech representations

acoustic representations：Mel-spectrogram，由 waveform 通过 STFT 转换而来，包含多种 linguistic 和 style 信息，从文本中合成这个丰富的特征会导致 one-to-many mapping 问题更加严重，且难以从 spectrogram 中提取出 expressive linguistic 信息；于是采用额外的 linguistic feature 来映射 text 和 acoustic 特征

linguistic representations：如下图，采用自监督的 speech representations 作为额外的 linguistic feature。本文采用 XLS-R（XLS-R- Self-supervised Cross-lingual Speech Representation Learning at Scale 笔记） 的第 12 层作为 linguistic representations，且做了大量实验来调研这些 representations 在 TTS 系统中的应用。

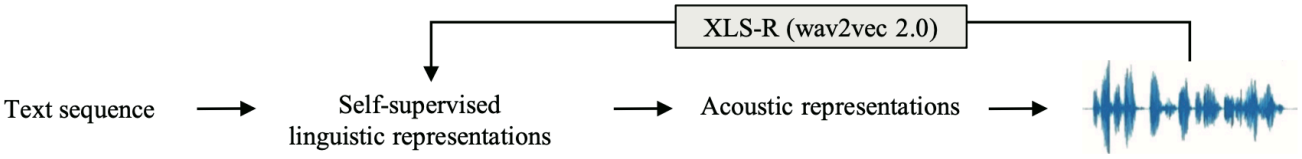

Figure 1: Hierarchical text-to-speech pipeline.

Hierarchical variational inference

为了连接 TTS 系统的两个部分，之前的 TTS 模型 VITS（VITS- Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech 笔记）采用了条件变分自编码器，最大化 ELBO，如下：

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z) - \log \frac{q_\phi(z|x)}{p_\theta(z|c)}\right]$$

其中，$p_\theta(z|c)$ 是给定条件 c 的 latent variables z 的先验分布，$p_\theta(x|z)$ 是给定 latent variables z 生成数据 x 的似然函数，$q_\phi(z|x)$ 是近似后验分布。VITS 采用 normalizing flow 和对抗训练来提高先验分布的表达能力。基于 VITS，HierSpeech 采用 hierarchical conditional VAE 通过 speech representations 中的 disentangled latent variables 来连接多层中间表征，且以端到端的方式学习这些 representation。

和最近提出的 hierarchical VAE 采用 top-down path networks 不同，HierSpeech 采用不同的 speech representations 分别近似 acoustic posterior distribution 和 linguistic posterior distribution。如下图：

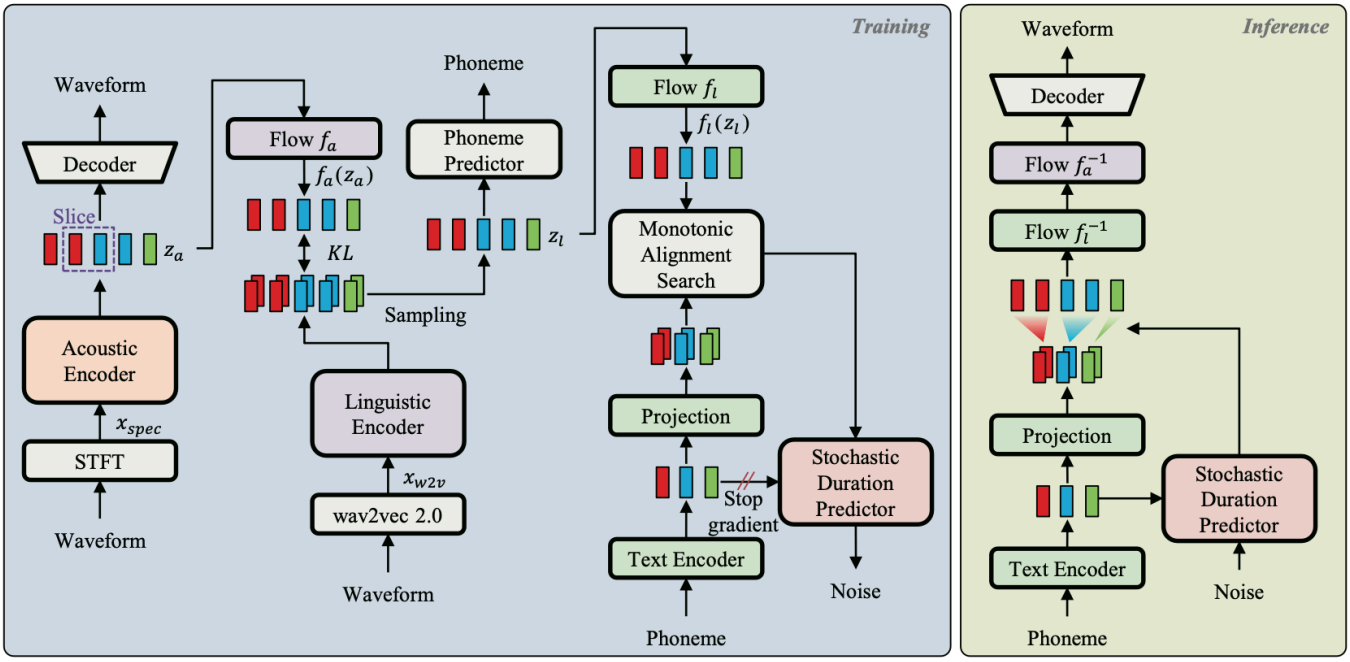

Figure 2: Overall framework of HierSpeech.

acoustic posterior distribution 和 linguistic posterior distribution 分别由 acoustic encoder ϕ_a（黄色方块） 和 linguistic encoder ϕ_l（紫色方块） 编码。为了 disentangle 每个 latent variable，使用目标语音的线性频谱 x_{spec}（左边黄色方块的输入）作为 acoustic representations z_a，使用 XLS-R 的第 12 层输出 x_{w2v}（中间紫色方块的输入）作为 linguistic representations z_l。HierSpeech 的优化目标如下：

$$\log p_\theta(x|c) \geq \mathbb{E}_{q_\phi(z|x)}\left[\log p_{\theta_d}(x|z_a) - \log \frac{q_{\phi_a}(z_a|x_{spec})}{p_{\theta_a}(z_a|z_l)} - \log \frac{q_{\phi_l}(z_l|x_{w2v})}{p_{\theta_l}(z_l|c)}\right]$$

其中，$z=[z_a, z_l]$，$\theta=[\theta_d, \theta_a, \theta_l]$，$\phi=[\phi_a, \phi_l]$，$q_{\phi_a}(z_a|x_{spec})$ 和 $q_{\phi_l}(z_l|x_{w2v})$ 分别是 acoustic 和 linguistic representation 的近似后验分布，$p_{\theta_l}(z_l|c)$ 是给定条件 c 的 linguistic latent variables z_l 的先验分布，$p_{\theta_a}(z_a|z_l)$ 是 acoustic latent variables z_a 的先验分布，其中 z_l 从 $q_{\phi_l}(z_l|x_{w2v})$ 中采样，$p_{\theta_d}(x|z_a)$ 是给定 latent variables z_a 生成数据 x 的似然函数。对于重构损失，使用 Mel-reconstruction loss $\mathcal{L}_{rec}$，最小化 Mel-spectrogram 之间的 L1 距离。

acoustic encoder ϕ_a 由 non-casual WaveNet residual blocks 组成，本质是gated activation unit + skip connection 组成的 dilated convolutions 层。然后，将输出送入 projection layer，使用重参数技巧从 posterior distribution 中采样得到 acoustic representations z_a。训练时，将 sliced z_a 送入 waveform decoder 重构原始音频 x。本文采用 HiFi-GAN（HiFi-GAN- Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis 笔记） generator G 作为 waveform decoder θ_d，其由 transposed convolution 和 multi-receptive field fusion module 组成。为了进行对抗训练，采用 multi-period discriminator D 来捕获 waveform 的不同周期特征。对抗损失如下：

$$\mathcal{L}_{adv}(D) = \mathbb{E}_{(x,z_a)}\Big[(D(x)-1)^2 + D(G(z_a))^2\Big],$$
$$\mathcal{L}_{adv}(\phi_a,\theta_d) = \mathbb{E}_{(z_a)}\Big[(D(G(z_a))-1)^2\Big]$$

其中，x 是 GT waveform。为了保证训练稳定，还使用了 feature matching loss $\mathcal{L}_{fm}$。

linguistic encoder ϕ_l 和 acoustic encoder ϕ_a 结构相同，但是 linguistic encoder 的输入是从预训练的 XLS-R 中提取的自监督 speech representations x_{w2v}，提取得到的 linguistic representations 为 z_l。为了强化 linguistic 特征，将 z_l 送入一个辅助的 phoneme predictor。最小化 CTC loss $\mathcal{L}_{ctc}$ 来训练 linguistic encoder 和 phoneme predictor。为了去除 projection layer，linguistic encoder 得到的 projected mean 和 variance 通过 θ_a 和 ϕ_l 之间共享权重 直接作为 acoustic prior distribution。为了保持 hierarchy，使用 normalizing flow 来转换 acoustic representations z_a。因此，acoustic prior 和 posterior 之间的 KL 散度最小化如下：

$$\mathcal{L}_{kl1} = \log q_{\phi_a}(z_a|x_{spec}) - \log p_{\theta_a}(z_a|x_{w2v})$$

因为 acoustic prior distribution 是从 linguistic information 中得到的，为了缩小两个分布之间的 gap，使用 normalizing flow 来 disentangle acoustic posterior 中的信息，增加 acoustic prior distribution 的表达能力：

$$p_{\theta_a}(z_a|x_{w2v}) = \mathcal{N}(f_a(z_a);\mu_{\theta_a}(x_{w2v}),\sigma_{\theta_a}(x_{w2v}))|\det(\partial f_a(z_a)/\partial z_a)|,$$
$$z_a \sim q_{\phi_a}(z_a|x_{spec}) = \mathcal{N}(z_a;\mu_{\phi_a}(x_{spec}),\sigma_{\phi_a}(x_{spec}))$$

text encoder θ_l 由 feed-forward Transformer 组成，输入是 phoneme sequence c_{text}，projection layer 用来产生 linguistic prior distribution 的 mean 和 variance。为了将 text 和 speech 的 linguistic representations 对齐，使用 MAS 来搜索满足最大化数据似然的 alignment A：

$$\mathcal{L}_{kl2} = \log q_{\phi_l}(z_l|x_{w2v}) - \log p_{\theta_l}(z_l|c_{text},A)$$

使用 normalizing flow 来增加 linguistic prior distribution 的表达能力：

$$p_{\theta_l}(z_l|c_{text},A) = \mathcal{N}(f_l(z_l);\mu_{\theta_l}(c_{text},A),\sigma_{\theta_l}(c_{text},A))|\det(\partial f_l(z_l)/\partial z_l)|,$$
$$z_l \sim q_{\phi_l}(z_l|x_{w2v}) = \mathcal{N}(z_l;\mu_{\phi_l}(x_{w2v}),\sigma_{\phi_l}(x_{w2v}))$$

为了从给定的 phonemes 采样 duration，采用 flow-based stochastic duration predictor，训练方式是最大似然估计。使用其负变分下界作为 duration loss $\mathcal{L}_{dur}$。对于多说话人设置，将 global speaker

embedding 添加到 acoustic/linguistic encoder 的 residual block，normalizing flow 的 residual block，stochastic duration predictor 和 decoder 中。

HierSpeech 的最终目标如下：

$$\mathcal{L}_{total} = \mathcal{L}_{kl1} + \lambda_{kl2}\mathcal{L}_{kl2} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{ctc}\mathcal{L}_{ctc} + \lambda_{dur}\mathcal{L}_{dur} + \lambda_{adv}\mathcal{L}_{adv}(\phi_a, \theta_d) + \lambda_{fm}\mathcal{L}_{fm}$$

Untranscribed text-to-speech

对于无文本的 TTS 模型 HierSpeech-U，使用 style encoder 来提取 speech 的 style embedding 作为 global conditioning。使用线性频谱作为 style encoder 的输入。在多说话人数据集上预训练后，可以在无文本的情况下适应到新说话人。通过自监督 speech representations，预训练的 linguistic encoder 可以在无文本的情况下从 speech 中提取丰富的 linguistic representations。因此，HierSpeech-U 可以通过仅使用 speech 数据来 fine-tuning acoustic encoder、acoustic prior 的 normalizing flow blocks 和 decoder，来合成具有新说话人风格的语音。

实验和结果（略）