A Limitations and Societal Impact

Limitations The training time of CLONE is relatively long, especially the 48 kHz CLONE. In addition, the amount of parameters of CLONE is relatively large. CLONE still has room for compression to be more suitable for on-device deployment.

Societal Impact CLONE can generate high-quality speech with better prosody. Thus, CLONE can be applied to any scenario requiring speech synthesis, especially scenarios requiring more prosody and emotional changes, such as video dubbing and speech synthesis for the virtual human. However, CLONE can also be abused to generate fake audios or videos, causing adverse effects such as telemarketing scams.

B Details of Formula

The derivation of the formula for KL divergence in prosody modeling is as follows:

$$\begin{aligned}\mathcal{L}_{kl} &= D_{kl}\left(q_\phi(z \mid x, c) \| p_\theta(z \mid c)\right) \\ &= \mathbb{E}_{q_\phi(z|x,c)}\left[\log q_\phi(z \mid x, c) - \log p_\theta(z \mid c)\right] \\ &= \mathbb{E}_{q_\phi(z|x,c)}[\log q_\phi(z \mid x, c)] - \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(z \mid c)],\end{aligned} \tag{11}$$

as we assume that the approximate posterior distribution of the phoneme-level prosody latent variable z is a normal distribution rather than a standard normal distribution, we have the follows:

$$q_\phi(z \mid x, c) \sim \mathcal{N}(\mu_\phi, \sigma_\phi), \tag{12}$$

and the differential entropy for a univariate normal distribution $p(x) \sim \mathcal{N}(\mu, \sigma)$ is as follows:

$$\mathbb{E}[-\log p(x)] = \log(\sigma * \sqrt{2\pi e}). \tag{13}$$

Thus, we have:

$$\mathbb{E}_{q_\phi(z|x,c)}[\log q_\phi(z \mid x, c)] = -\log(\sigma_\phi * \sqrt{2\pi e}). \tag{14}$$

Besides, $\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(z \mid c)]$ does not have a closed-form solution. So we compute $\log p_\theta(z \mid c)$ for each sampled z and then average them. For each sampling, we have:

$$\begin{aligned}\mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(z \mid c)] &= \mathbb{E}_{q_\phi(z|x,c)}\left[\log\left(\mathcal{N}\left(f_\theta(z); \mathbf{0}, \boldsymbol{I}\right) \cdot \left|\det \frac{\partial f_\theta(z)}{\partial z}\right|\right)\right] \\ &= -\frac{\log(2\pi)}{2} - \frac{f_\theta(z)^2}{2} + \log\left(\left|\det \frac{\partial f_\theta(z)}{\partial z}\right|\right),\end{aligned} \tag{15}$$

where $z \sim q_\phi(z \mid x, c)$. Thus, we have:

$$\begin{aligned}\mathcal{L}_{kl} &= \mathbb{E}_{q_\phi(z|x,c)}[\log q_\phi(z \mid x, c)] - \mathbb{E}_{q_\phi(z|x,c)}[\log p_\theta(z \mid c)] \\ &= -\log(\sigma_\phi * \sqrt{2\pi e}) + \frac{\log(2\pi)}{2} + \frac{f_\theta(z)^2}{2} - \log\left(\left|\det \frac{\partial f_\theta(z)}{\partial z}\right|\right) \\ &= -\log(\sigma_\phi) - \frac{1}{2} + \frac{f_\theta(z)^2}{2} - \log\left(\left|\det \frac{\partial f_\theta(z)}{\partial z}\right|\right).\end{aligned} \tag{16}$$

The training loss of the prosody predictor is the KL divergence between two normal distributions which are the distribution of prosody predictor $s_\psi(z \mid t) \sim \mathcal{N}(\mu_{pp}, \sigma_{pp})$ and the distribution of posterior encoder $q_\phi(z \mid x, c) \sim \mathcal{N}(\mu_\phi, \sigma_\phi)$, respectively. As KL divergence between two normal distributions has closed-form solution, we have:

$$
\begin{aligned}
\mathcal{L}_{pp} &= D_{kl}(\mathcal{N}(\mu_{pp}, \sigma_{pp}), \mathcal{N}(\mu_{\phi}, \sigma_{\phi})) \\
&= \frac{1}{2}\log(2\pi\sigma_{\phi}^2) + \frac{\sigma_{pp}^2 + (\mu_{pp} - \mu_{\phi})^2}{2\sigma_{\phi}^2} - \frac{1}{2}\left(1 + \log(2\pi\sigma_{pp}^2)\right) \\
&= \log\frac{\sigma_{\phi}}{\sigma_{pp}} + \frac{\sigma_{pp}^2 + (\mu_{pp} - \mu_{\phi})^2}{2\sigma_{\phi}^2} - \frac{1}{2}.
\end{aligned}
\tag{17}
$$

C Hyperparameter and Model Configuration of CLONE

We list the hyperparameters of each module of CLONE as shown in Table 4.

Table 4: The hyperparameter and model configurations of CLONE.

Module	Parameter
Speaker Embedding Size	256
Text Encoder	
Phoneme Embedding Size	192
Feed-Forward Transformer Layers	6
Feed-Forward Transformer Hidden Channels	192
Feed-Forward Transformer Conv1D Kernel Size	3
Feed-Forward Transformer Conv1D Filter Size	768
Feed-Forward Transformer Attention Heads	2
Prosody Predictor	
Dilated CNN Layers	4
Dilated CNN Hidden Channels	192
Dilated CNN Kernel Size	5
Dilated CNN Dilation Rate	1
Duration Predictor	
Conv1D Kernel Size	3
Conv1D Filter Size	256
Posterior Encoder	
Dilated CNN Layers	8
Dilated CNN Hidden Channels	192
Dilated CNN Kernel Size	5
Dilated CNN Dilation Rate	1
Acoustic Encoder	
Dilated CNN Layers	8
Dilated CNN Hidden Channels	192
Dilated CNN Kernel Size	5
Dilated CNN Dilation Rate	1
Posterior Wave Encoder	
Dilated CNN Hidden Layers	8
Dilated CNN Hidden Channels	192
Dilated CNN Kernel Size	5
Dilated CNN Dilation Rate	1
Flow	
Flows	4
Residual Coupling Layers	4
Residual Coupling Layer Hidden Size	192
Residual Coupling Layer Kernel Size	5
Residual Coupling Layer Dilation Rate	1
Multi-Band Discriminator	
24 kHz Model Band Number	2
48 kHz Model Band Number	4
λ of Loss Function	
$\lambda_{[1\rightarrow 6]}$	$[45.0, 1.0, 10.0, 0.1, 1.0, 1.0]$