VoiceFixer

TFGAN Vocoder - Training - Frequency Domain losses

- Loss Function: $L_{syn} = L^T + L^F + \lambda_1 L^D$
- Frequency Domain Losses:
  - $L^F = \lambda_2 L^{mel} + \sum_k L_k^f$
  - $L_k^f(\hat{s}, s) = \lambda_3 L_k^{sc}(\hat{s}, s) + \lambda_4 L_k^{mag}(\hat{s}, s)$
- **Mel loss, spectral convergence loss and magnitude loss:**
  - Capture mel domain information: $L^{mel}(\hat{s}, s) = \left\| \, |\hat{S}|_{mel} - |S|_{mel} \, \right\|_2$
  - Loss on linear scale: $L^{sc}(\hat{s}, s) = \frac{\left\| \, |\hat{S}| - |S| \, \right\|_F}{\left\| \, |\hat{S}| \, \right\|_F}$
  - Loss on log scale: $L^{mag}(\hat{s}, s) = \left\| \, log(|\hat{S}|) - log(|S|) \, \right\|_1,$

Table.4 STFT parameter for each k

k	1	2	3	4	5	6	7
win-length	4096	2048	1024	512	256	128	64
hop-length	2048	1024	512	256	128	64	32
fft-size	8192	4096	2048	1024	512	256	128

VoiceFixer

TFGAN Vocoder - Training - Discrminator Losses

- Loss Function: $\mathbb{L}_{syn} = L^T + L^F + \lambda_1 L^D$
- Discriminator Losses:
  - $D(\hat{s}) = D^{T\_sub}(\hat{s}) + D^F(\hat{s}) + \sum_{r=1}^{4} D_r^T(\hat{s})$
  - $L^D = \min_G \max_D (\mathbb{E}_s(log(D(s))) + \mathbb{E}_{\hat{s}}(log(1 - D(\hat{s}))))$.

Table.5 The architecture of time domain discriminator

T-discriminator
Conv1d(1, 128, ks=16), LeakyRelu(0.2)
Conv1d(128, 128, ks=41, stride=4, padding=20, groups=8), LeakyRelu(0.2)
Conv1d(128, 128, ks=41, stride=4, padding=20, groups=16), LeakyRelu(0.2)
Conv1d(128, 128, ks=41, stride=4, padding=20, groups=32), LeakyRelu(0.2)
Conv1d(128, 1, ks=3, stride=1, padding=1), LeakyRelu(0.2)

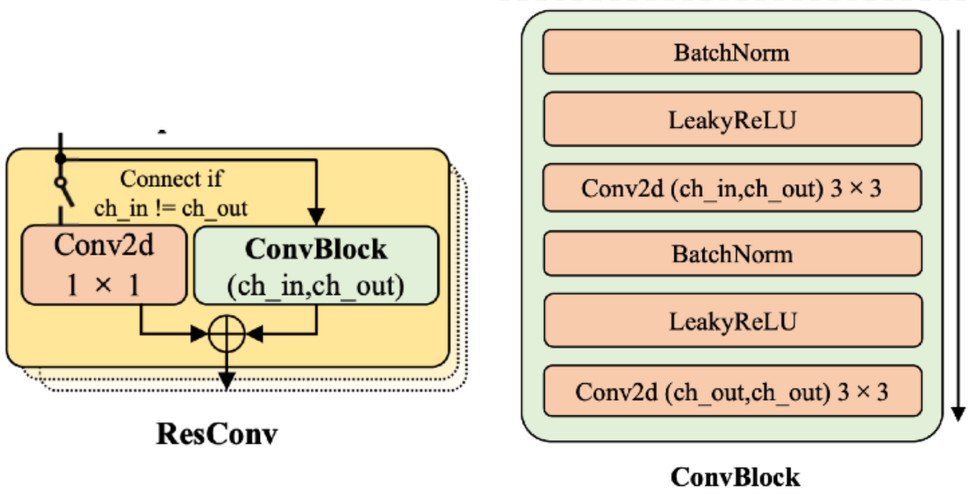

Table.6 The architecture of frequency domain discriminator

F-discriminator
Conv2d(1,32,kernal_size=(3,3))
ResConv(32, 32, stride=1,kernal_size=(3,3))
ResConv(32, 32, stride=1,kernal_size=(3,3))
ResConv(32, 64, stride=2,kernal_size=(3,3))
ResConv(64, 64, stride=1,kernal_size=(3,3))
ResConv(64, 32, stride=2,kernal_size=(3,3))
ResConv(32, 32, stride=1,kernal_size=(3,3))
ResConv(32, 32, stride=2,kernal_size=(3,3))
ResConv(32, 32, stride=1,kernal_size=(3,3))