HW 2 - GAN-based Mel-Vocoder

Name: 譚至斌
ID: d12942015

Novelty

- No novelty
  - I just reproduced Hifi-GAN in the homework dataset……

Methodology

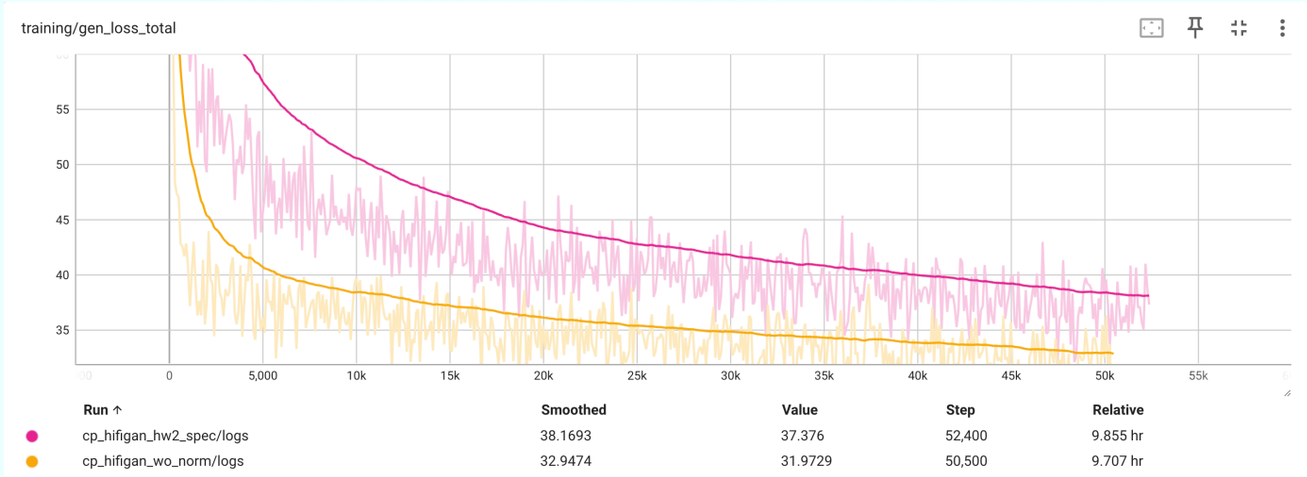
- The methodology can be described by 4 steps:

    1. Resample all the audio to 22050
    2. Replace Hifi-GAN's `melspectrogram` with hw2's `melspectrogram`
    3. Train Hifi-GAN
    4. Inference

Result

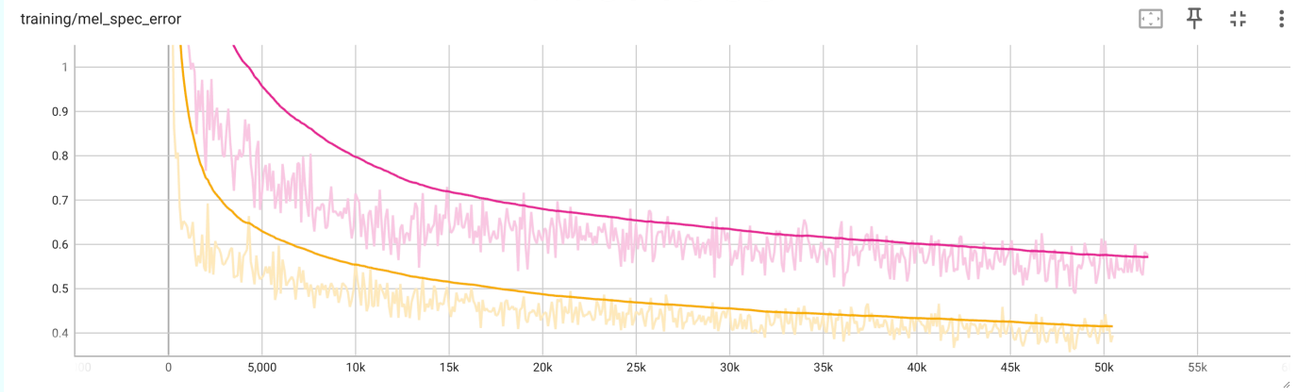
- About 50k steps, the total training loss decreased to **~36** and the mel loss decreased to **~0.58**
- The validation loss of mel decreased to **~0.76**
- Note that we consider the pink line which represents the loss computed by hw2's `melspectrogram`

Discriminative loss + Mel loss



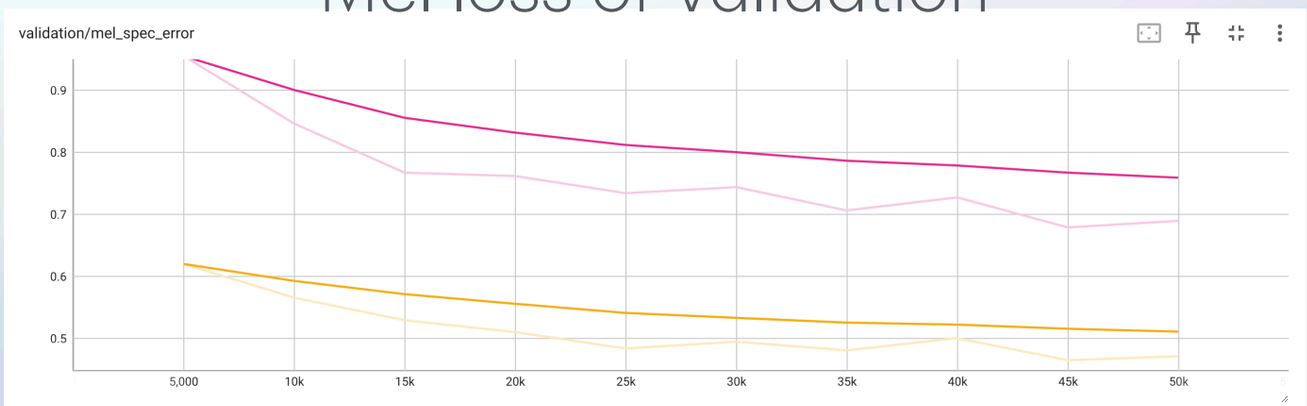
Run ↑	Smoothed	Value	Step	Relative
cp_hifigan_hw2_spec/logs	38.1693	37.376	52,400	9.855 hr
cp_hifigan_wo_norm/logs	32.9474	31.9729	50,500	9.707 hr

Mel loss



Run ↑	Smoothed	Value	Step	Relative
cp_hifigan_hw2_spec/logs	0.5724	0.5756	52,400	9.855 hr
cp_hifigan_wo_norm/logs	0.4159	0.3955	50,500	9.707 hr

Mel loss of validation



Run ↑	Smoothed	Value	Step	Relative
cp_hifigan_hw2_spec/logs	0.7594	0.6899	50,000	8.724 hr
cp_hifigan_wo_norm/logs	0.5116	0.4719	50,000	8.888 hr

Result

- Evaluation on the validation set:
  - Griffin & Lim algorithm gets best scores

Griffin & Lim

Hifi-GAN (hw2)

Hifi-GAN (original)


```
{
    "M-STFT": 0.3219452682323159,
    "log2f0_mean": 0.06180450039017508 5,
    "FAD": 0.02174284536276261
}{
    "M-STFT": 1.120453163981437 7,
    "log2f0_mean": 0.1761471862812564,
    "FAD": 0.3201834354009509 6
}{
    "M-STFT": 9.274185792325785,
    "log2f0_mean": 6.465007469306043,
    "FAD": 8.585338238296353
}%
```

Findings

- Griffin&Lim algorithm **outperform** on the validation dataset
    - It's so surprising to me while hearing the results generated by this classical algorithm and find them nearly same as the source audios
- I find the audio generated by Hifi-GAN containing obvious artifacts which sounds like **trembling**
    - I guess the reason is Hifi-GAN generate audio by frames and cannot keep the consistence among a group of frames

Findings

- Some interesting findings about implementation
  - Even if all the parameters set to the same values, `Melspectrogram` in `torchaudio` still makes **different** outputs from the hand-craft `melspectrogram` implemented by Hifi-GAN's authors
  - `melspectrogram` implemented by Hifi-GAN's authors always get **smaller** loss values in the training process

Discussion

- This homework is a good experience for me to learning lots of audio processing coding skills like:
    - How to deal with spectrogram transforming
    - How to train a GAN
- Since some mistakes on coding, I have to give up an **~100k-steps** checkpoint
    - However, the generated audios still **worse than those from Griffin&Lim algorithm**

Discussion

- I choose the results generated by Hifi-GAN which is trained by myself rather than Griffin&Lim algorithm
    - It makes me feel more sense of accomplishment……

Appendix

- source code: https://github.com/tanchihpin0517/NTU/tree/master/deepmir/hw2
- model: https://www.dropbox.com/scl/fi/b0kdwrxufq91tp5jm8b6j/g_00050000?rlkey=20bved2g0ijev9zx0yycnlv3u&dl=0