7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)

October 19-20, 2024

Generative Adversarial Network based Neural Vocoder for Myanmar End-to-End Speech Synthesis

Aye Mya Hlaing, Win Pa Pa

Natural Language Processing Lab.,

University of Computer Studies, Yangon, Myanmar

Contents

- Introduction
- Myanmar End-to-End Speech Synthesis
- Parallel WaveGAN
- HiFi-GAN
- Dataset
- Experimental Setups
- Results

Introduction

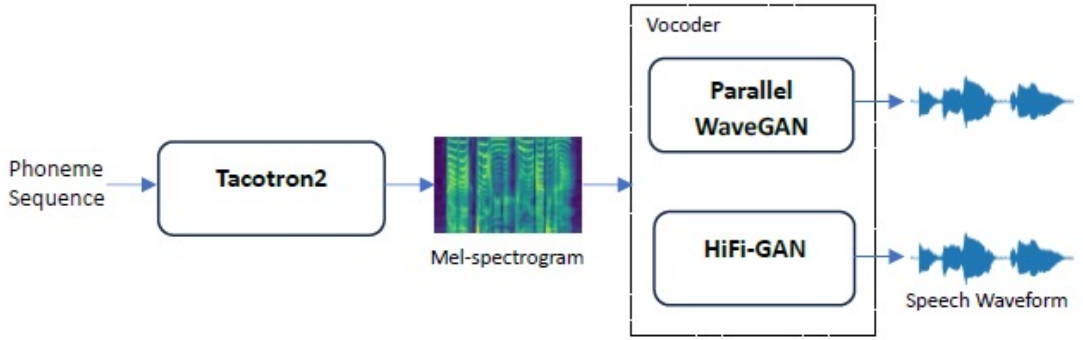
- Text-to-speech (TTS) models focus on synthesizing intelligible and natural sounding speech
- In recent years, end-to-end neural TTS models have emerged to simplify traditional speech synthesis pipeline and their synthesized speeches can be comparable with human recordings.
- The end-to-end neural TTS is typically composed of two main processing models
  - *spectral representation generator* : generates the spectral representation such as mel-spectrograms given the input text or phoneme
  - *vocoder* : converts the speech waveforms from the generated mel-spectrograms

Introduction (Cont'd)

- The separately trained neural vocoders based on Generative Adversarial Network(GAN) have demonstrated remarkable capabilities in generating natural-sounding synthetic speech.
- In this work, two GAN based neural vocoders, Parallel WaveGAN and HiFi-GAN were trained on Myanmar speech dataset
- The ability of each vocoder in ground truth mel-spectrogram inversion, generalization on unseen speakers, and Myanmar end-to-end speech synthesis was examined
- This is the first effort to explore the advance of neural vocoder in Myanmar end-to-end TTS.
- The audio samples are available on

  http://nlpresearch-ucsy.edu.mm/subeval-voc.html

Myanmar End-to-End Speech Synthesis



- Tacotron2 (Shen et al., 2018) model was trained for phoneme to mel-spectrogram generation
- Tacotron2 is a recurrent sequence-to-sequence feature prediction network with attention that maps phoneme embeddings to mel-spectrograms
- Parallel WaveGAN (Yamamoto et al., 2020) and HiFi GAN (Kong et al., 2020) are separately trained on Myanmar speech dataset.
- The generated mel-spectrograms were given into the GAN-based vocoders as the input conditions to synthesize speech waveform.

Parallel WaveGAN

- The Parallel WaveGAN (Yamamoto et al., 2020) is a distillation-free, fast, and small-footprint waveform generation method using GAN.
- The model is non-autoregressive at both training and inferencing.
- The generator is trained by jointly optimizing the multi-resolution short-time Fourier transform (STFT) auxiliary loss L_{aux} and the waveform domain adversarial loss L_{adv}

$$L_G = L_{aux}(G) + \lambda_{adv} L_{adv}(G, D)$$

- The discriminator is trained to correctly classify the generated sample as fake and simultaneously ground truth sample as real

HiFi-GAN

- HiFi-GAN has been composed of one generator and two discriminators containing multi-scale discriminator (MSD) and multi-period discriminator (MPD) (Kong et al., 2020)
- The generator of HiFi-GAN is a fully convolutional neural network with multi-receptive field fusion (MRF) module that can perceives the various length of patterns in parallel

$$L_G = L_{Adv}(G;D) + \lambda_f L_F(G;D) + \lambda_m L_M(G)$$

- In the discriminator part, each sub-discriminator of MPD handles equally spaced samples of input audio and MSD was used to capture consecutive patterns and long-term dependencies.

Dataset

- Myanmar phonetically balanced speech corpus (PBC) (Thu et al., 2015) built from Basic Travel Expression Corpus (BTEC) (Kikui et al., 2003)
- 4000 utterances recorded by a native female speaker
- 16 kHz sampling rate of speech data
- 3,800 utterances were utilized for training, 100 utterances each for validation and testing

Experimental Setup of Parallel WaveGAN

- 80-band log-mel spectrograms with band-limited frequency range (80 to 7600 Hz) as the input auxiliary features
- Weight normalization was applied to all convolutional layers of both generator and discriminator
- The model was trained for 200K steps
- The discriminator was fixed for the first 100K steps, and then both the generator and the discriminator were trained

  - https://github.com/kan-bayashi/ParallelWaveGAN

Experimental Setup of HiFi-GAN

- The configuration of HiFi-GAN V1 from the original paper (Kong et al., 2020), was applied to train the model on Myanmar speech dataset.
- 80-band log-mel spectrograms with band-limited frequency range (80 to 7600 Hz) as input conditions
- The model was trained for only 200K steps, the same steps used for training the Parallel WaveGAN model
- This is very small compared to the training steps used in the original paper (2.5M steps)
- Each vocoder model was trained on a Nvidia Tesla K80 GPU

Experimental setup of Tacotron2

- The Tacoron2 model was trained for 125K steps with Adam optimizer and a batch size of 32
- In the training process, the guided attention loss was used to promote a fast and robust attention learning
- ESPnet, an end-to-end speech processing toolkit was used for modelling
  - https://github.com/espnet/espnet
- This model was trained on two Nvidia Tesla K80 GPUs

Results

- Three mean opinion score (MOS) tests for
  - Ground truth mel-spectrogram inversion
  - Generalization to unseen speakers
  - End-to-end Myanmar speech synthesis
- Ten native non-expert speakers participated in all MOS tests.
- Subjects were given the synthesized speeches of two models and ground truth audio
- They had to rate the quality of synthesized speeches on a scale of 1 to 5 where 1 is bad and 5 is excellent.

Results (Cont'd)

- **Ground Truth Mel-spectrogram Inversion**

Model	MOS	RTF
Ground Truth	4.69 ± 0.10	-
Parallel WaveGAN	4.49 ± 0.12	0.015
HiFi-GAN	4.59 ± 0.11	0.011

*RTF is based on the average inference time of 100 utterances on single Nvidia Tesla K80 GPU

Results (Cont'd)

- **Generalization to unseen speakers**
  - 10 utterances of two unseen female speakers were utilized for investigating the ability of our trained models

Model	MOS
Ground Truth	4.68 ± 0.12
Parallel WaveGAN	4.42 ± 0.12
HiFi-GAN	4.48 ± 0.11

Results (Cont'd)

- **End-to-end Myanmar speech synthesis**
  - To verify the effectiveness GAN-based vocoders in Myanmar end-to-end TTS pipeline, each model was integrated to the Tacotron2 model

Model	MOS
Ground Truth	4.68 ± 0.15
Tacotron2 + Parallel WaveGAN	4.33 ± 0.13
Tacotron2 + HiFi-GAN	4.37 ± 0.13

Conclusion

- Both Parallel WaveGAN and HiFi-GAN models achieve high-fidelity speech synthesis with fast inference speeds, showing the ability of generalizing to unseen speakers.
- GAN-based models, even trained on the small dataset with limited training steps, can achieve high quality speech for low-resource languages.
- Future work will focus on the mel-spectogram generator to better capture the prosody of speech and using GAN-based vocoders in various end-to-end speech synthesis settings.

References

- Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. 2018. *Natural tts synthesis by conditioning wavenet on mel spectrogram predictions*. In 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP), pages 4779–4783. IEEE.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. *Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis*. Advances in neural information processing systems, 33:17022– 17033.
- Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. 2020. *Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram*. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6199–6203. IEEE.
- Ye Kyaw Thu, Win Pa Pa, Jinfu Ni, Yoshinori Shiga, Andrew M Finch, Chiori Hori, Hisashi Kawai, and Eiichiro Sumita. 2015. *Hmm based Myanmar text to speech system*. In INTERSPEECH, pages 2237–2241.
- Gen-ichiro Kikui, Eiichiro Sumita, Toshiyuki Takezawa, and Seiichi Yamamoto. 2003. Creating corpora for speech-to-speech translation. In INTERSPEECH, pages 381–384.

Thank you for your attention!