OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

Parallel Computing and I/O

Multi-Speaker Speech Synthesis with Diverse Prosody Control using Generative Adversarial Network

- Author: Kaveen Matta Kumaresh
- Type: Master's Thesis
- Date: 2022-02-12
- Reviewers: Jun.-Prof. Dr.-Ing. Ingo Siegert, Jun.-Prof. Dr. Michael Kuhn
- Supervisors: Nazli Deniz Cagatay

Speech synthesis is a domain oriented towards generating human-like speech using machines and algorithms. Research towarc it has existed since the 1980s, in the name of text-to-speech systems. However, the generated speeches had remained monotonic, until recently. With the advancement of machine learning and specifically using deep neural networks, impressive speech quality has been achieved. Typical text-to-speech systems based on machine learning approaches today use a 3-block architecture. One particular block in this 3-stage text-to-speech architecture is the vocoder block. Several different types of vocoders have been experimented with, a recent contender being the Generative Adversarial Network (GAN). In this thesis, we explore the 3-stage pipeline of - acoustic block, vocoder block and the speaker encoder block. We utilize a pre-built acoustic mo - Tacotron2, as our feature extractor. We specifically look into the state-of-art vocoders that use the concept of GANs and choos one of them - MelGAN, as our synthesizer. We start with experimenting the influence of different parameters of MelGAN on the synthesis quality. We further apply our own modifications, namely - introducing the binary cross entropy (BCE) loss in the discriminator, adapting the Wasserstein loss and applying gradient penalty during the learning process, aiming at improving the chosen GAN. Along with this, we also perform batch normalization during model training, to improve the speed of training. Consecutively, we try to include generator input noise parameter during the GAN based vocoder model training and inference. This noise parameter coupled with style tokens of our chosen acoustic model allows us to generate diverse prosody speech. We perform all of the above mentioned contributions by basing the model training on a multi-speaker speech dataset, which allows t system to adapt on different voices. This thesis is performed in collaboration with Bragi.