Waveform generation for text-to-speech synthesis using pitch-synchronous multi-scale generative adversarial networks

Lauri Juvela[1], **Bajibabu Bollepalli**[1], **Junichi Yamagishi**[2,3], **Paavo Alku**[1]
[1]Aalto University, Finland [2]National Institute of Informatics, Japan
[3]The Centre for Speech Technology Research, University of Edinburgh, United Kingdom

Aalto University
School of Electrical Engineering

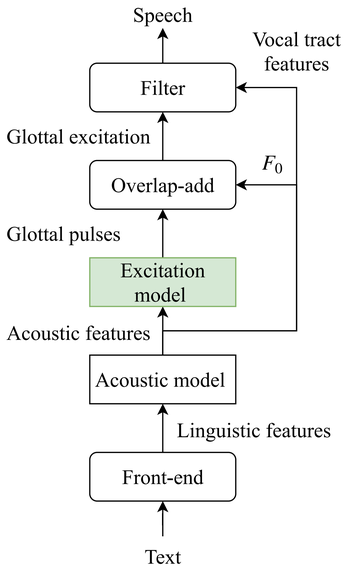
大学共同利用機関法人 情報・システム研究機構
国立情報学研究所
National Institute of Informatics

1 Introduction

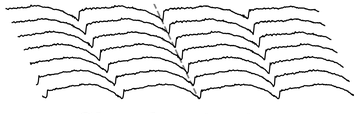
- Neural waveform generators are used in state-of-the-art TTS
- Problem: sample-by-sample synthesis with WaveNet is slow, parallel models are heavy and difficult to train
- Generative models are required to capture the stochastic components in the speech waveform
- GANs are promising, but have suffered from training instability

2 Text-to-speech synthesis system

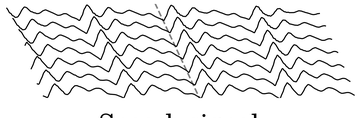
- Some speech processing can make neural vocoding more manageable:
  - Use spectral envelope to remove vocal tract resonances (glottal inverse filtering)
  - Use pitch marking to create phase-locked waveform representations
  - **Use generative neural nets as excitation model**
- Front-end and acoustic model are as in conventional neural net based statistical parametric speech synthesis



Waveform representation



Glottal excitation

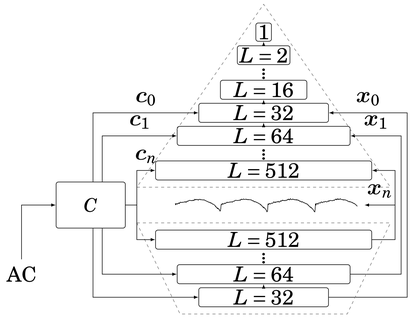


Speech signal

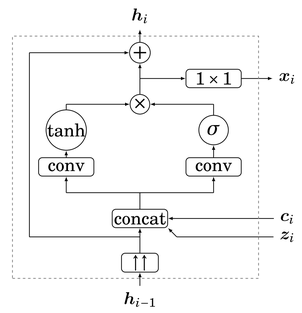
3 GAN waveform model

Multi-scale GAN architecture

Progressive upsampling in Generator



Multi-scale GAN architecture



Generator block

Training criteria

Wasserstein GAN: x is real data and $\hat{x} = G(z, c)$ at all timescales

$$\mathscr{L}_D^{\mathrm{W}} = -\mathbb{E}_{x\sim p_{\mathscr{D}}}[D(x,c)] + \mathbb{E}_{\hat{x}\sim p_G}[D(\hat{x},c)], \tag{1}$$

with gradient penalty for a smooth Discriminator

$$\mathscr{L}_D^{\mathrm{GP}} = \mathbb{E}_{x\sim p_{\mathscr{D}},\hat{x}\sim p_G}\left[(\max\{0, \|\nabla_{\tilde{x}} D(\tilde{x},c)\| - 1\})^2\right], \tag{2}$$

and regularization for large gradients in the data manifold

$$\mathscr{L}_D^{\mathrm{R1}} = \mathbb{E}_{x\sim p_{\mathscr{D}}}\left[\|\nabla_x D(x,c)\|^2\right]. \tag{3}$$

Spectral magnitude matching for the Generator

$$\mathscr{L}_G^{\mathrm{FFT}} = \mathbb{E}\left[(|\mathrm{FFT}(x_n)| - |\mathrm{FFT}(\hat{x}_n)|)^2\right]. \tag{4}$$

Sound samples

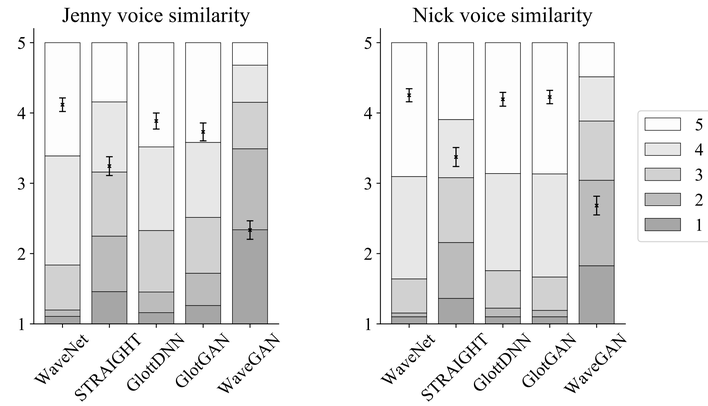
`https://users.aalto.fi/~ljuvela/multiscale-gan/`

Source code (TensorFlow)

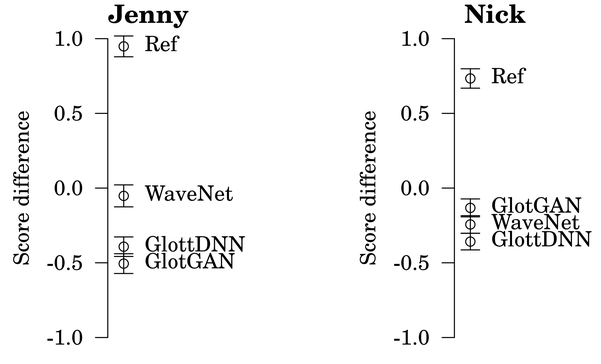
`https://github.com/ljuvela/multiscale-GAN`

4 Listening tests

- TTS systems trained on Jenny (4.7h) and Nick (1.8h) datasets
- Difference mean opinion score (DMOS) test for speaker similarity



- Category comparison rating (CCR) test to evaluate quality



- **The proposed method can match a WaveNet vocoder in TTS quality**

Department of Signal Processing and Acoustics
School of Electrical Engineering
Aalto University, Finland

E-mail: lauri.juvela@aalto.fi

This work was supported by the Academy of Finland (proj. no. 256961 and 284671)