Speech waveform synthesis from MFCC sequences with generative adversarial networks

Lauri Juvela[1], **Bajibabu Bollepalli**[1], **Xin Wang**[2], **Hirokazu Kameoka**[3], **Manu Airaksinen**[1], **Junichi Yamagishi**[2], **Paavo Alku**[1]

[1] Department of Signal Processing and Acoustics, Aalto University, Finland
[2] National Institute of Informatics, Japan
[3] NTT Communication Science Laboratories, NTT Corporation, Japan

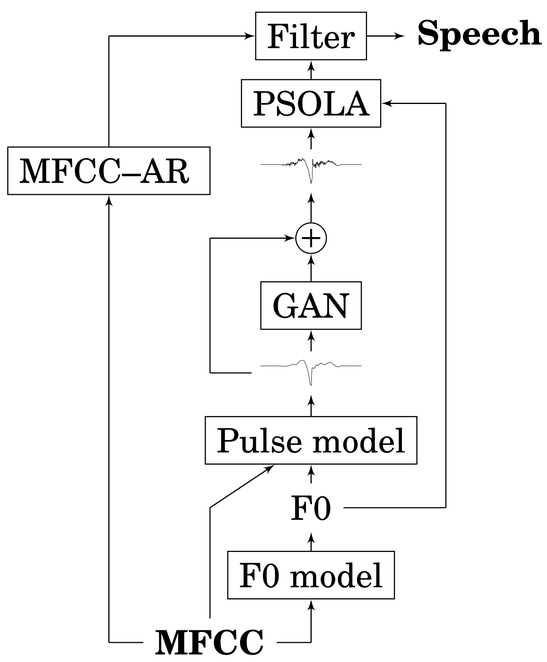
Aalto University
School of Electrical
Engineering

1 Introduction

- High-quality synthesis from filterbank MFCCs only?
- MFCCs contain spectral envelope information
  → use this explicitly and only generate a residual excitation
- Pitch-synchronous waveform generation is easier (than WaveNet)
  → predict F0 from MFCCs

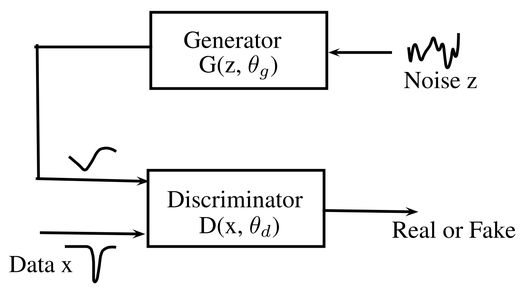
2 Speech synthesis system

- F0 model predicts quantized F0 from MFCCs [1]
- Pulse model predicts average (smooth) excitation pulses, given MFCC and F0
- Residual GAN generates an additive stochastic noise component
- Excitation signal is assembled with PSOLA and filtered with MFCC-AR to produce speech

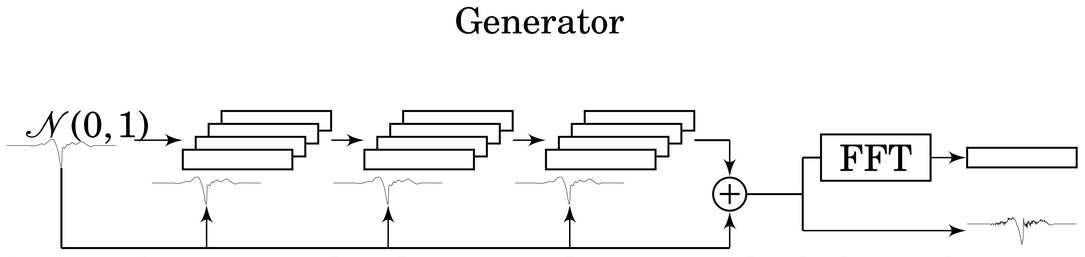


3 Generative adversarial networks

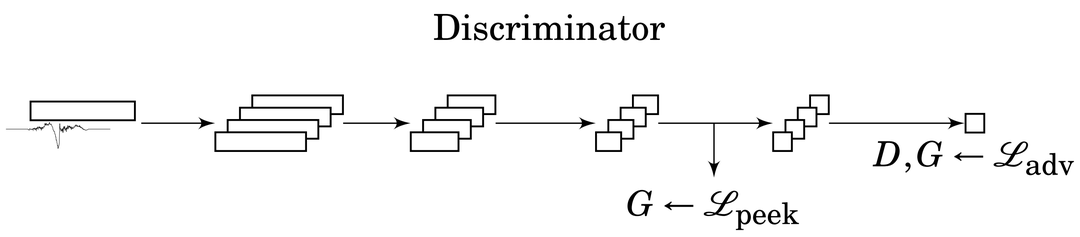
- Generator G attempts to fool Discriminator D by generating realistic samples
- Discriminator sees real and generated data and attempts to separate them



Proposed residual GAN architecture



- A smooth generated pulse is used as a residual channel input to each layer
- Forcing additivity at output makes the generator learn a residual noise component
- FFT layer allows discriminator a simultaneous time-frequency view of the signal



- Adversarial loss is used to train D and G
- The generator is allowed to peek into discriminator activations and match generated batch with real data batch

AR envelope from MFCC

- The MFCCs $\boldsymbol{C}$ are computed as

$$\boldsymbol{C} = \mathbf{D}\log(\mathbf{M}\boldsymbol{S}),$$

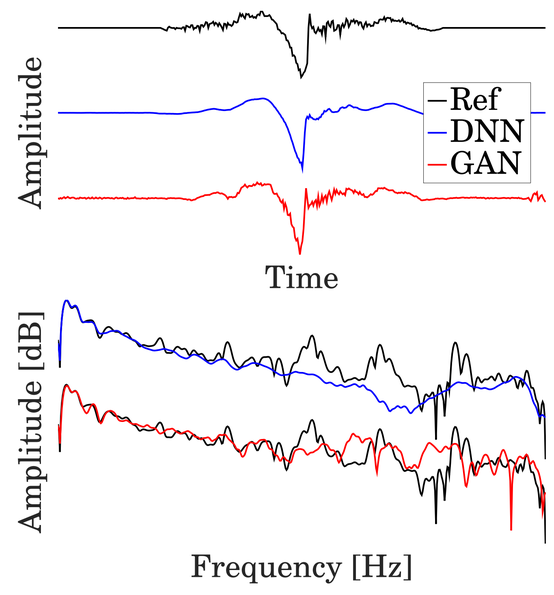
  where $\boldsymbol{S}$ is a FFT magnitude spectrum, $\mathbf{M}$ is a mel-filterbank matrix, and $\mathbf{D}$ is a (truncated) DCT matrix.
- An approximate inversion is given by

$$\hat{\boldsymbol{S}} = \mathbf{M}^{+}\exp(\mathbf{D}^{+}\boldsymbol{C}),$$

  where $\mathbf{M}^{+}$ and $\mathbf{D}^{+}$ denote pseudo-inverses
- AR-envelope (all-pole filter) is calculated by solving the normal equations obtained from $\boldsymbol{r} = \text{IFFT}(\hat{\boldsymbol{S}}^2)$

Generated excitation waveforms

- Speech is inverse filtered with MFCC-AR and target excitation pulses are phase-locked to pitch marks (i.e. GCIs, see [2])
- Pulses generated by deterministic DNN are smooth and lack high frequency content
- GAN is able to generate a realistic additive stochastic component in time and frequency

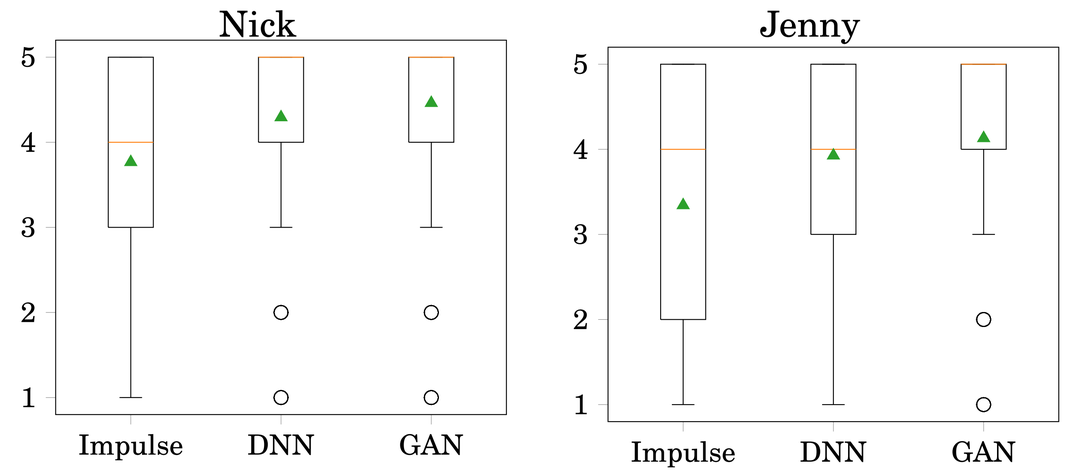


4 Experiments

- Sound samples: `http://tts.org.aalto.fi/mfcc_synthesis/`
- Source code: `https://github.com/ljuvela/ResGAN`

Listening test

- DMOS test to rate degradation compared to natural reference
- Impulse train uses only F0 and MFCC-AR envelope for synthesis
- DNN uses the pulse prediction model only, while GAN also uses the noise model



References

[1] X. Wang, S. Takaki, and J. Yamagishi, "An RNN-based quantized F0 model with multi-tier feedback links for text-to-speech synthesis," in *Proc. Interspeech*, 2017, pp. 1059–1063. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-246

[2] L. Juvela, B. Bollepalli, M. Airaksinen, and P. Alku, "High-pitched excitation generation for glottal vocoding in statistical parametric speech synthesis using a deep neural network," in *Proc. ICASSP*, March 2016, pp. 5120–5124.

This work was supported by the Academy of Finland (proj. no. 256961 and 284671)
E-mail: lauri.juvela@aalto.fi