Generative Adversarial Network-based Postfilter for STFT Spectrograms

Takuhiro Kaneko[1], Shinji Takaki[2], Hirokazu Kameoka[1], Junichi Yamagishi[2]

[1]NTT Communication Science Laboratories, NTT Conrporation, Japan

[2]National Institute of Informatics, Japan

INTERSPEECH 2017
Situated Interaction
August 20-24, 2017 | Stockholm, Sweden
www.interspeech2017.org

1 Introduction

Background

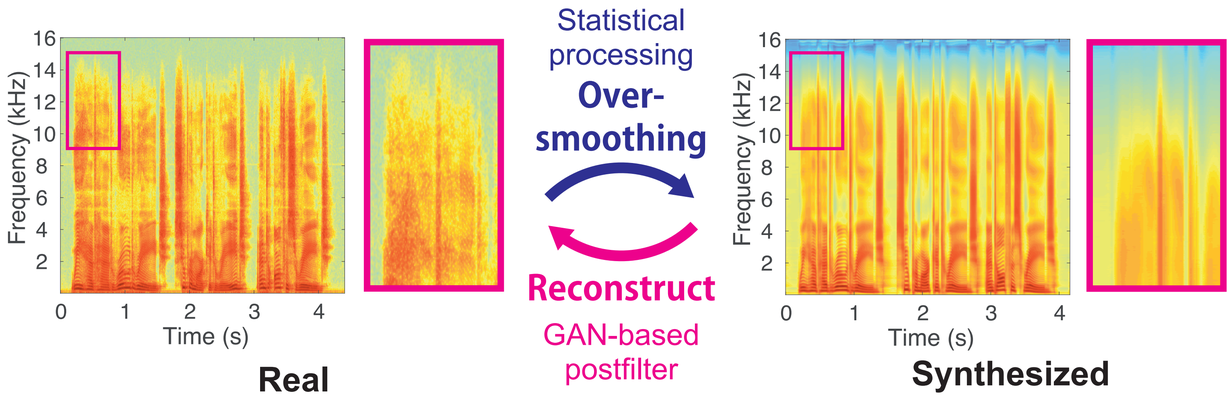
- In speech-processing systems, **STFT spectrograms** have been widely used as key acoustic representations.
- The aim with these systems is to produce spectrograms with quality indistinguishable from real ones, but typically they **lack fine structure** through statistical processing.

Objective

- Overcome these limitations and reconstruct **spectrograms having finer structure**.

Solution

- **Generative adversarial network (GAN)-based postfilter for STFT spectrograms**.



2 GAN-based Postfilter for STFT Spectrograms

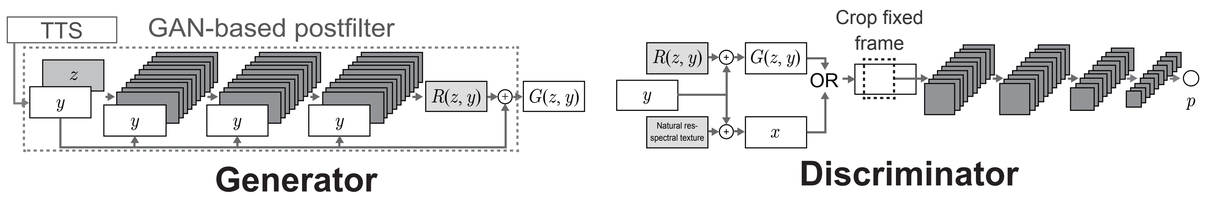
i GAN [Goodfellow+2014]

- The goal is to learn generative distribution $P_G(x)$ matching true data distribution $P_{\text{Data}}(x)$.
- Composed of two networks:
  - **Generator** G: Map noise variable $z \sim P_{\text{Noise}}(z)$ to data space $x = G(z)$.
  - **Discriminator** D: Assign probability p for "real" sample x, 1 - p for generated sample $G(z)$.
- D and G play two-player **minmax** game:

$$\min_G \max_D \mathbb{E}_{x\sim P_{\text{Data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim P_{\text{Noise}}(z)}[\log(1 - D(G(z)))].$$

ii GAN-based Postfilter for Vocoder Parameters [Kaneko+2017]

- The goal is to reconstruct convincing **vocoder parameters** from synthesized ones.



(1) **Conditional**: Reconstruct natural spectral texture x from synthesized one y and noise z.

$$\min_G \max_D \; \mathbb{E}_{x,y\sim P_{\text{Data}}(x,y)}[\log D(x,y)] + \mathbb{E}_{z\sim P_{\text{Noise}}(z),y\sim P_y(y)}[\log(1 - D(G(z,y),y))].$$

(2) **Residual**: Shorten the entire process of generating the spectral texture.

$$G(z,y) = y + R(z,y).$$

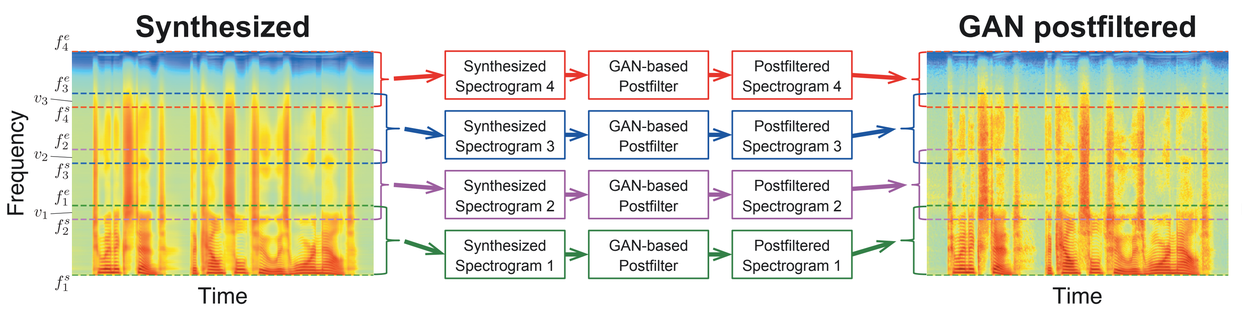
(3) **Convolutional**: Represent spectro-temporal structures with reasonably small parameters.

Fully Convolutional Network for G: Allow input segments to take an arbitrary length.

iii GAN-based Postfilter for STFT Spectrograms

- The goal is to reconstruct convincing **STFT spectrograms** from the synthesized one.
  - **Challenges**: High dimensional, different structures depending on the frequency bands.

  => Take a simple **divide-and-concatenate** strategy.



(1) **Partition**: Divide the spectrograms into N frequency bands with **overlap**.

(2) **Postfiltering**: Reconstruct the individual bands using the GAN-based postfilter trained for each band.

(3) **Concatenation**: Apply a **window function** to each band to smoothly connect and then concatenate them.

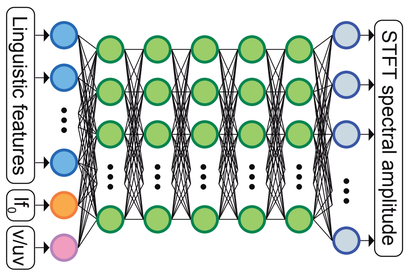
* In the model without overlap, the reconstructed spectrogram tends to have discontinuity between bands, causing a popping sound.

3 Application to Speech Synthesis

STFT-based speech synthesis

[Takaki+2017, **Interspeech 2017 Tue-O-4-1-3**]

- Instead of vocoder features, STFT spectra are directly predicted from text.
- Experimental results show that this method outperforms vocoder-based method.



STFT-based speech synthesis with GAN-based postfilter

(1) **Predict STFT-spectra** directly from text.

(2) **Postfilter** predicted STFT spectrograms: Use **GAN-based postfilter** here!

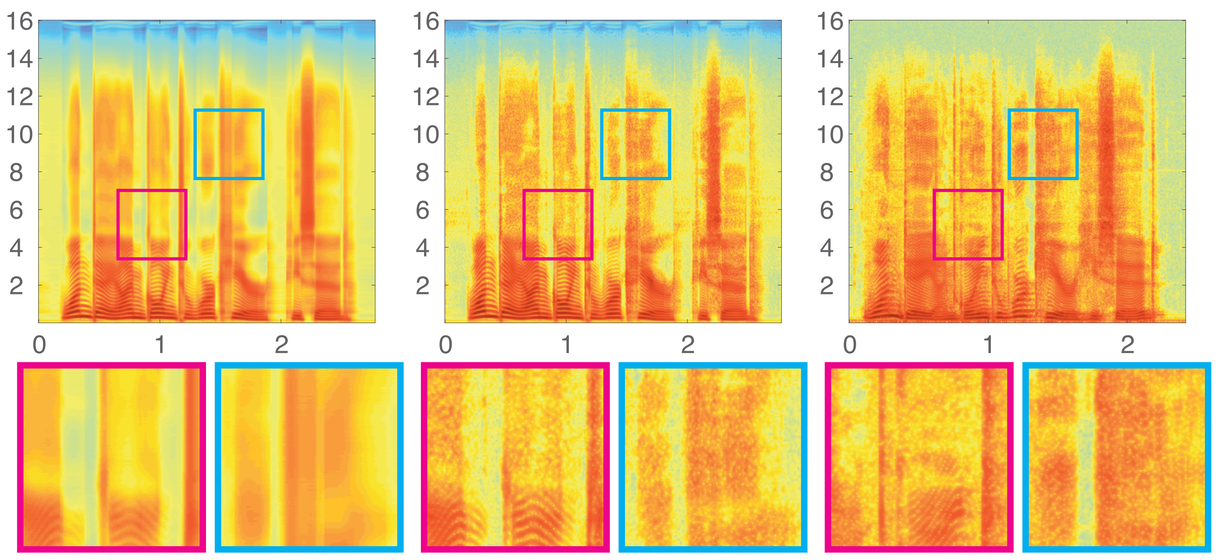
(3) **Generate waveform** using Griffin and Lim phase recovery.

4 Experiments

i Experimental Conditions

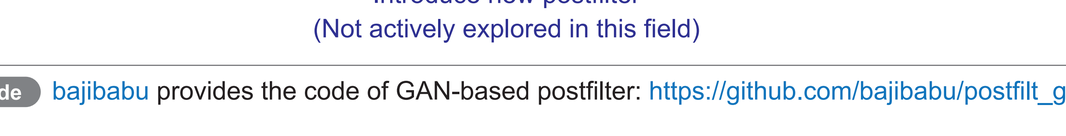
- **Dataset**: Blizzard Challenge 2011
  - **Speaker**: American professional female speaker.
  - **Data**: 12,085 utterances (17 hours), 200 utterances used for evaluation.
  - **Sampling rate**: 32 kHz.
  - **STFT condition**: Frame length: 25 ms. Frame shift: 5 ms. Blackman window function.
- **GAN-based Postfilter**
  - **Partition**: 4 Bands (0-5 kHz, 4-9 kHz, 8-13 kHz, 12-16 kHz). Overlap: 1 kHz.
  - **Concatenation**: Hamming window function.
- **Comparison**
  - **Baseline**: Speech synthesis without GAN-based postfilter (**w/ conventional postfilter**)
  - **Proposed**: Speech synthesis with GAN-based postfilter (**w/ GAN postfilter**)

ii Sample Results



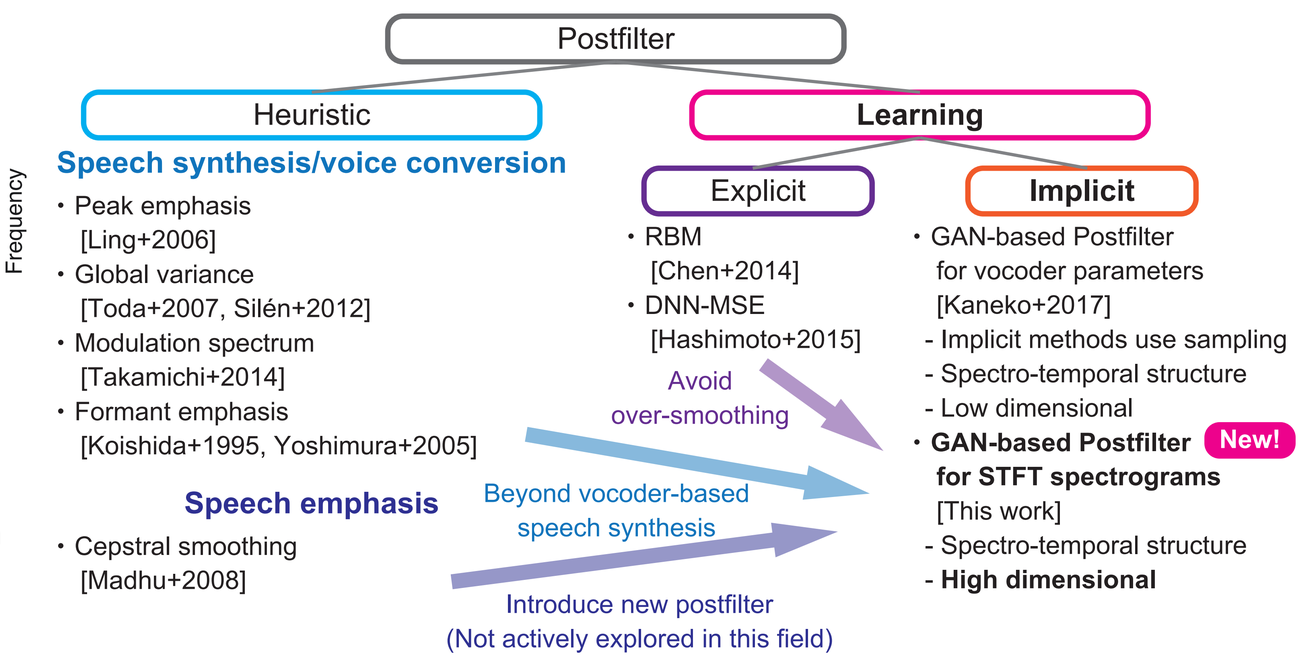
iii Subjective Evaluation

- **AB preference test**
  - **Participants**: 18 native speakers of English.
  - **Sentences**: 8 sentences randomly selected from 200 test sentences.



Difference was statistically significant according to t-test ($p < 0.01$).

5 Related Work



Code bajibabu provides the code of GAN-based postfilter: https://github.com/bajibabu/postfilt_gan