Research Projects Summary

Ke Fang, Naver Clova AI Research

2018.7.24

Projects

- Speech style transfer (2018/04/13 – 06/02)
- Continuous audio generation with GAN (2018/06/03 – 07/24)

Speech style transfer

Goal

- Given two audios with different voices of person **A** and **B**, we try to replace **B**’s speech/song with **A**’s voice.

Obama’s speech

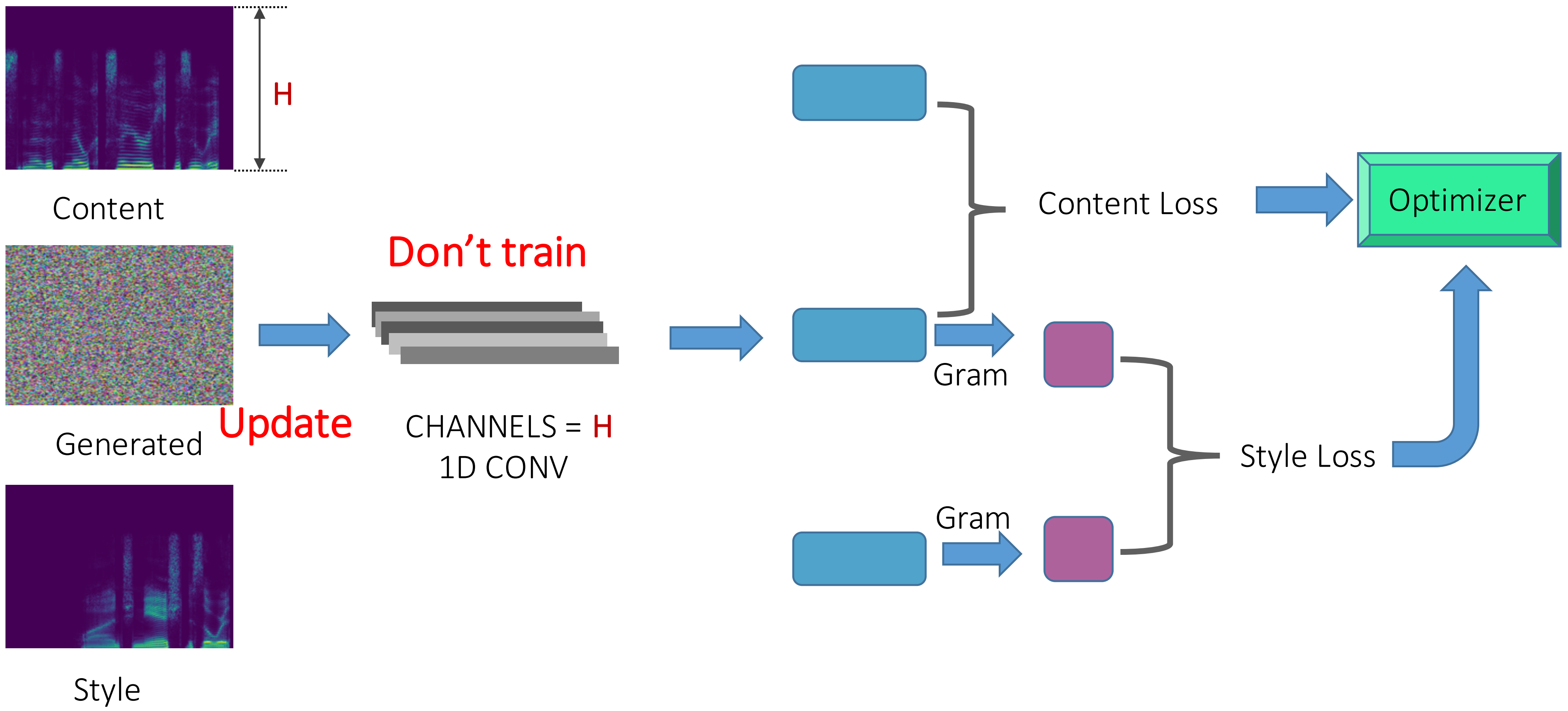
Adele – Rolling in the deep

Singing this song with Obama’s voice

Remarkable related work

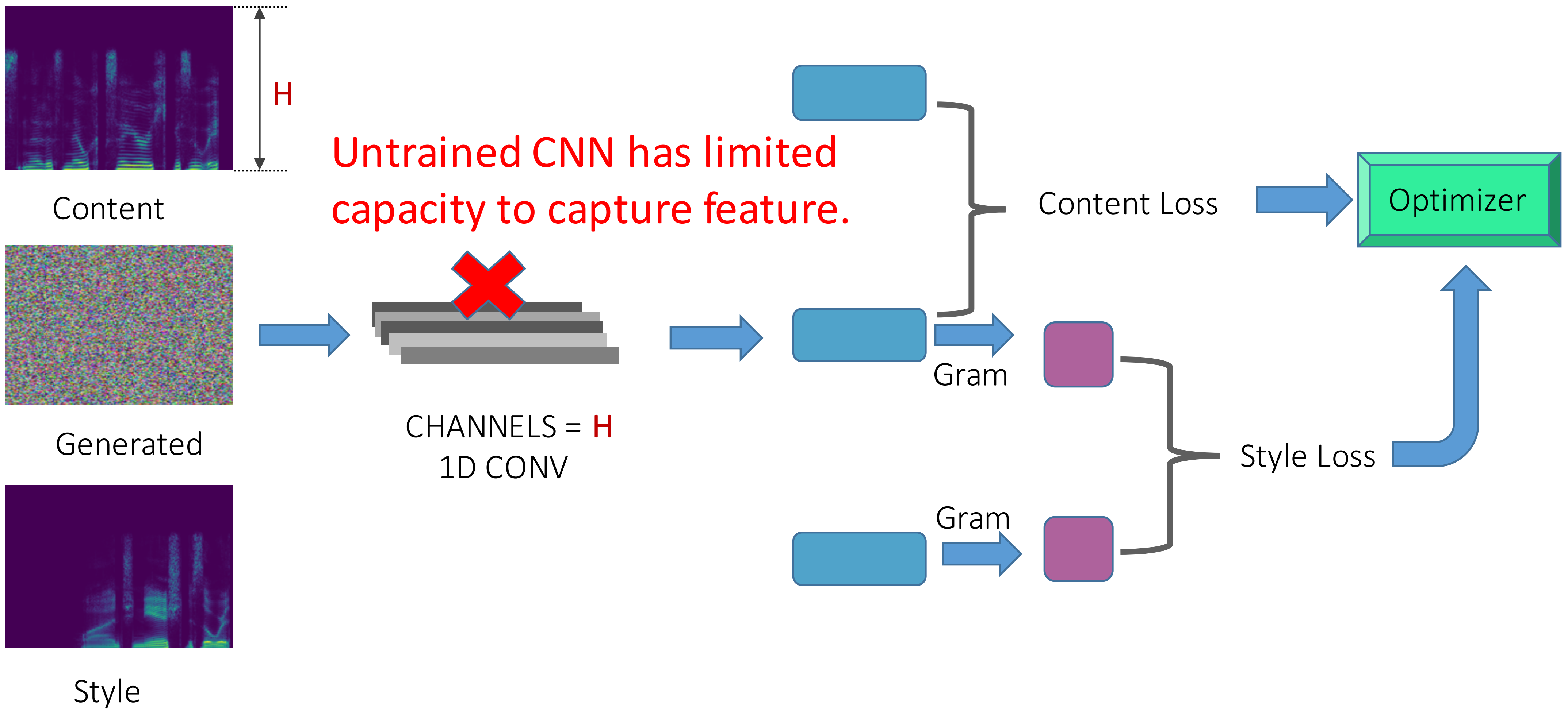
- Shallow untrained CNN style transfer (Dmitry Ulyanov, *2016*)
- Neural Voice Cloning with a Few Samples (**Baidu**, *Feb 2018*)
- A universal music translation network (**Facebook**, *May 2018*)
- Adversarial learning disentangled audio representation (**NTU** Hung-yi Lee group, *Interspeech 2018*)

Shallow untrained CNN



https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/

Shallow untrained CNN



https://dmitryulyanov.github.io/audio-texture-synthesis-and-style-transfer/

Neural Voice Cloning

$t_{i,j}$: The jth text of speaker i
s_i: Speaker information of i
W: Trainable parameters
e_{s_i}: Trainable speaker encoder
$a_{i,j}$: The ith audio of speaker i

Trainable

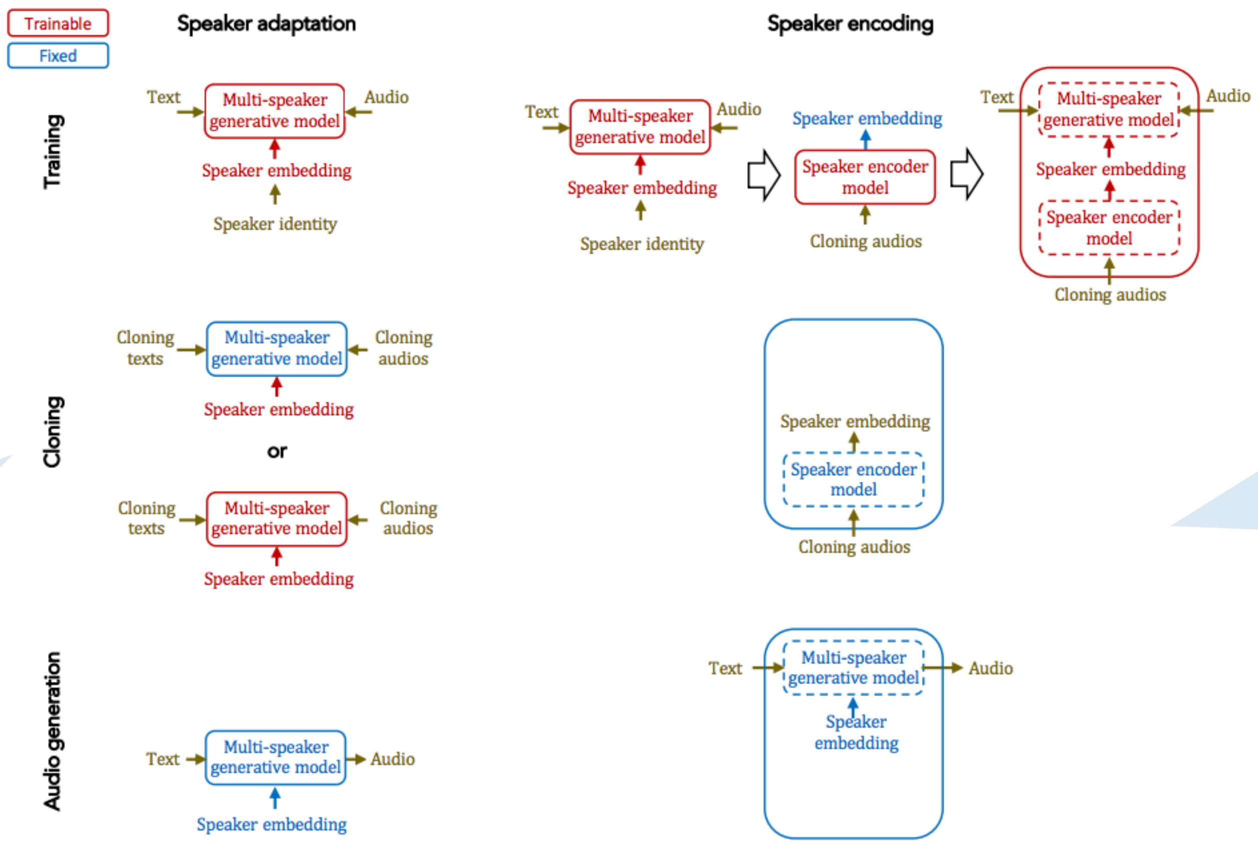
$$\min_{W,\mathbf{e}} \mathbb{E}_{\substack{s_i\sim\mathcal{S},\\ (\mathbf{t}_{i,j},\mathbf{a}_{i,j})\sim\mathcal{T}_{s_i}}} \{L(f(\mathbf{t}_{i,j}, s_i; \underline{\underline{W, \mathbf{e}_{s_i}}}), \mathbf{a}_{i,j})\}$$

1. Train generative model to update W to reconstruct $a_{i,j}$ based on $(t_{i,j}, s_i, e_{s_i})$

2. Fix the generative model, use speaker embedding to clone audios

3. Train speaker encoder e_{s_i} to predict the embedding from sampled cloning audios

4. For unseen speaker:
- Use pre-trained speaker encoder model to get speaker embedding
- Clone audios with text and speaker embedding by generative model

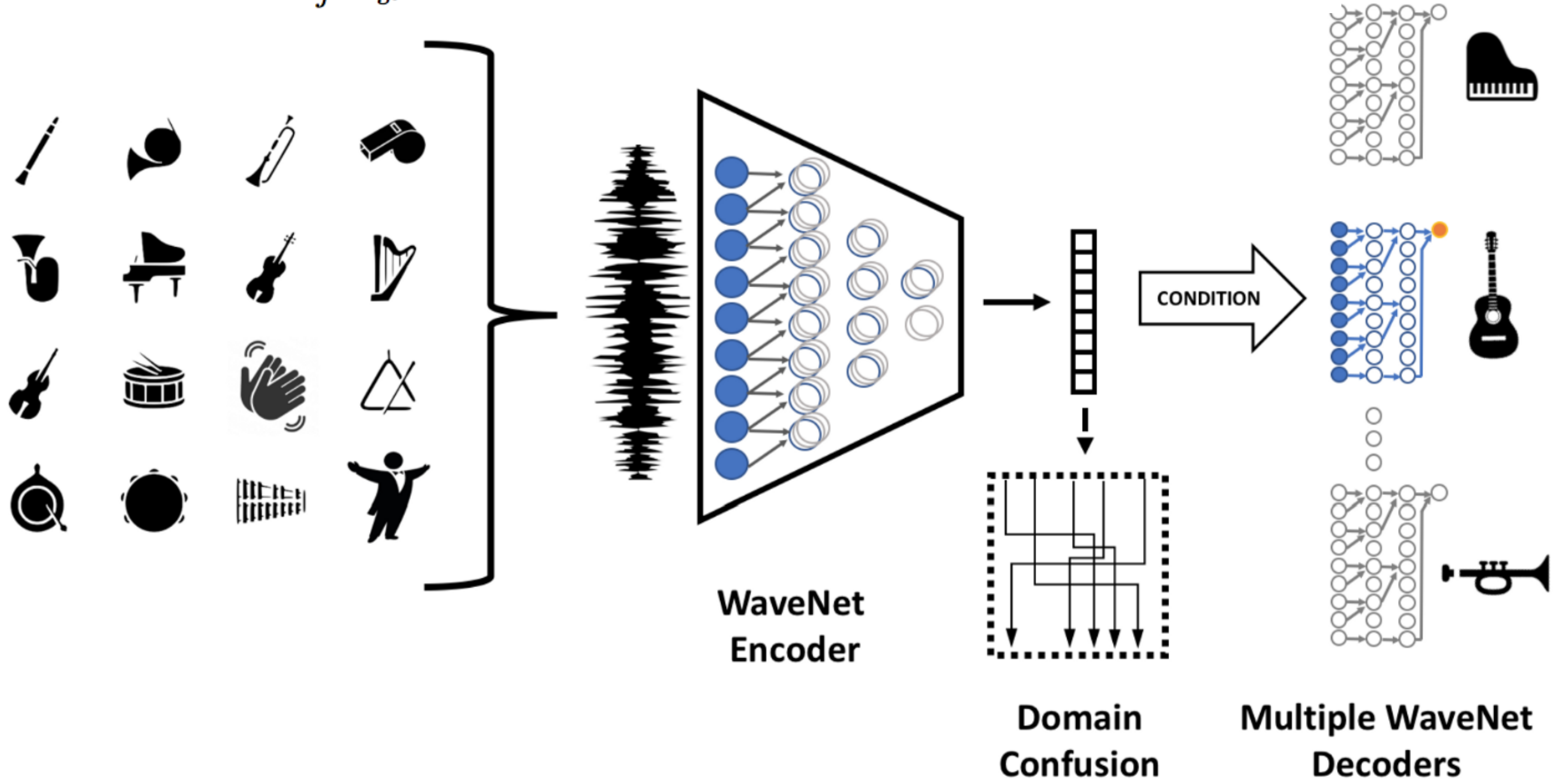


Sercan O. Arik, et al., "Neural Voice Cloning with a Few Samples", *Mar 2018*

Universal music translation

s^j: Input sample from domain j
r: Random seed.
O: Audio augmentation
E: Shared encoder
D^j: Decoder of domain j

$$\sum_j \sum_{s^j} \mathbb{E}_r \mathcal{L}(D^j(E(O(s^j, r))), s^j) - \lambda \mathcal{L}(C(E(O(s^j, r))), j)$$



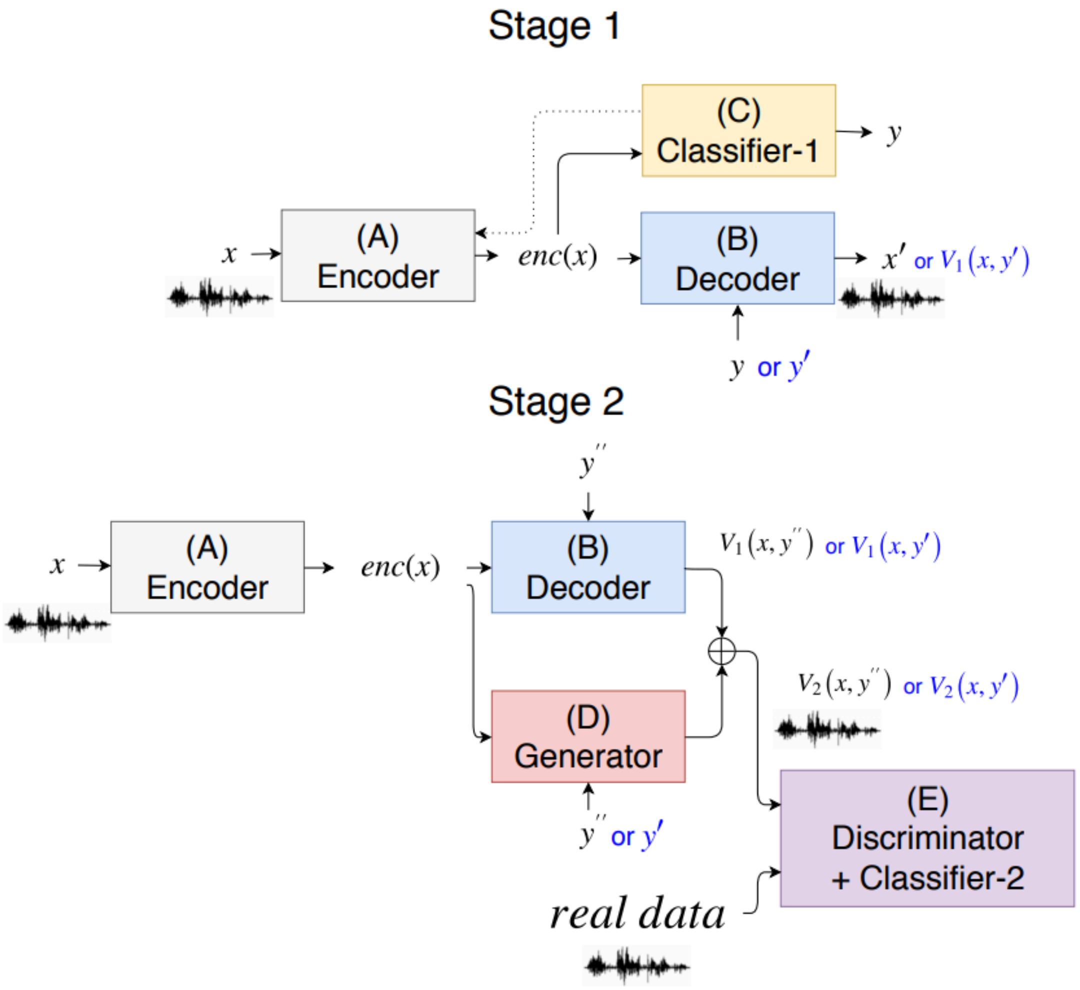
Highlights:
- Shared Encoder.
- A Classifier to discourage Encoder from encoding texture information
- Every instrument has it's own Decoder

Training:
- Classic piano reconstruction with [Encoder, Decoder, Classifier, Instruments Information]

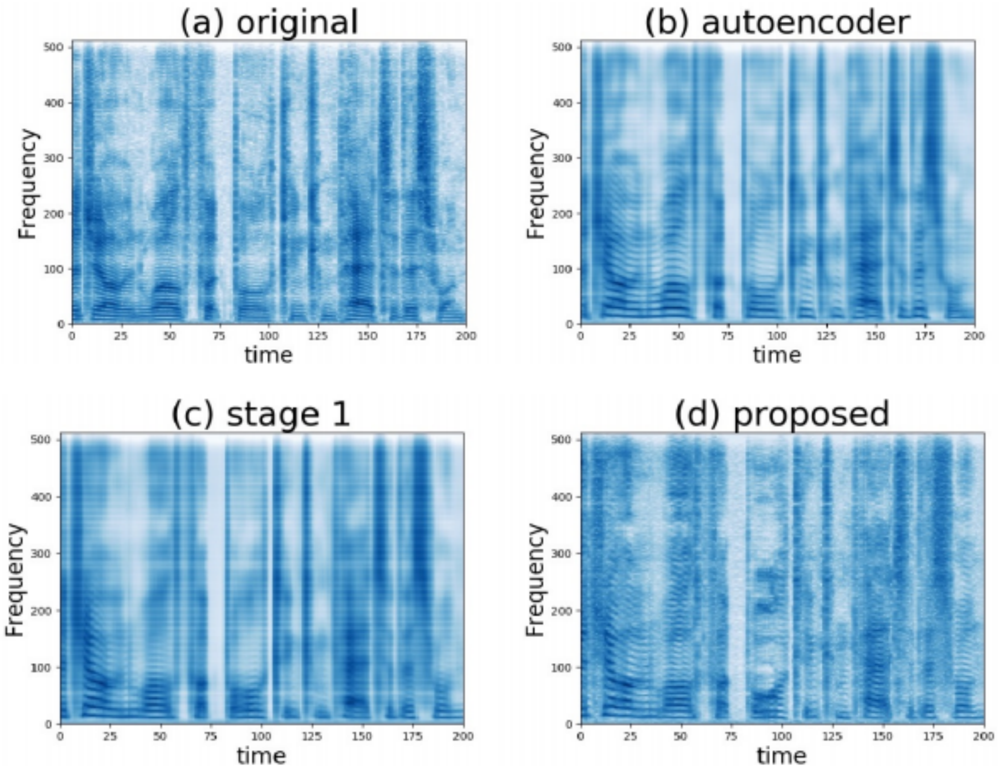
Transfer instrument $A \rightarrow B$:
* Audio a from unseen instrument A pass through shared Encoder, then decoded by B's Decoder

Noam Mor, et al., "A Universal Music Translation Network", *May 2018*

Disentangle audio representation



- **Same with FAIR's work**(but eariler.)
- Adversarial training for texture.



Ju-chieh Chou, et al., "Multi-target Voice Conversion without Parallel Data by Adversarially Learning Disentangled Audio Representations", *Interspeech 2018*

Unmentioned related works

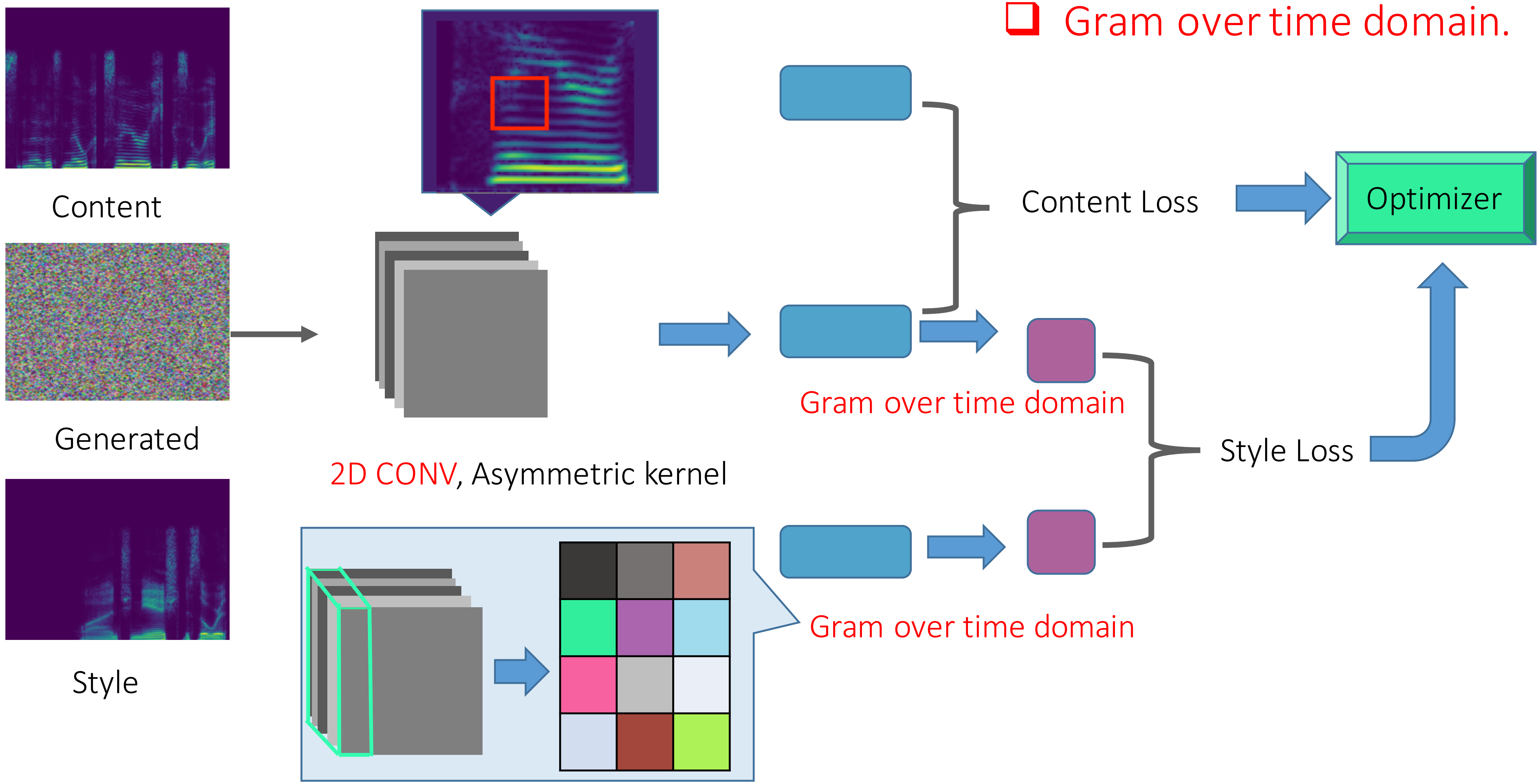
- **StarGAN** voice conversion (**NTT** Lab, *Jun 2018*)
- On Using Backpropagation for Speech Texture Generation and Voice Conversion(**Google**, *Dec 2017*)
- **VQ-VAE** for style transfer(**Google**, *NIPS 2017*)
- Deep voice conversion: Speaking like Kate Winslet(github@andabi)

My work of speech style transfer

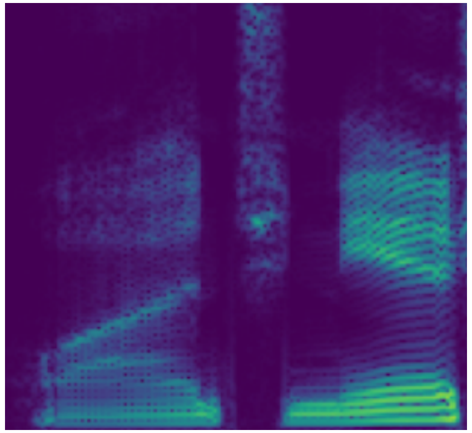
Shallow untrained CNN

Major differences:

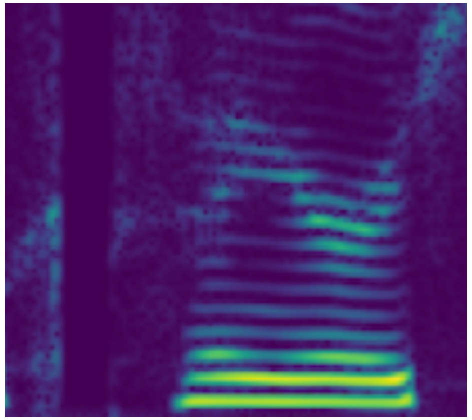
- 2D CONV rather than 1D.
- Gram over time domain.



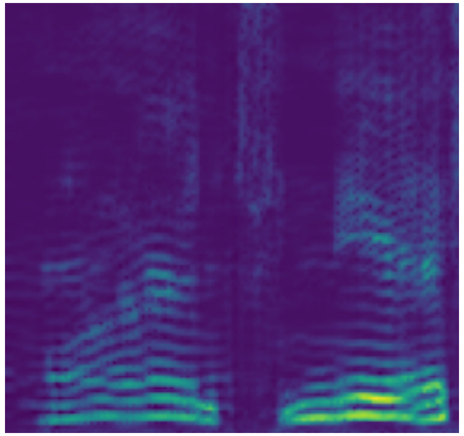
Shallow untrained CNN: Result



Content



Style



Generated

Major problems:

- ❑ L2 loss for content distance.
- ❑ The capacity of untrained CNN.

Does "gram-over-time-domain" work?

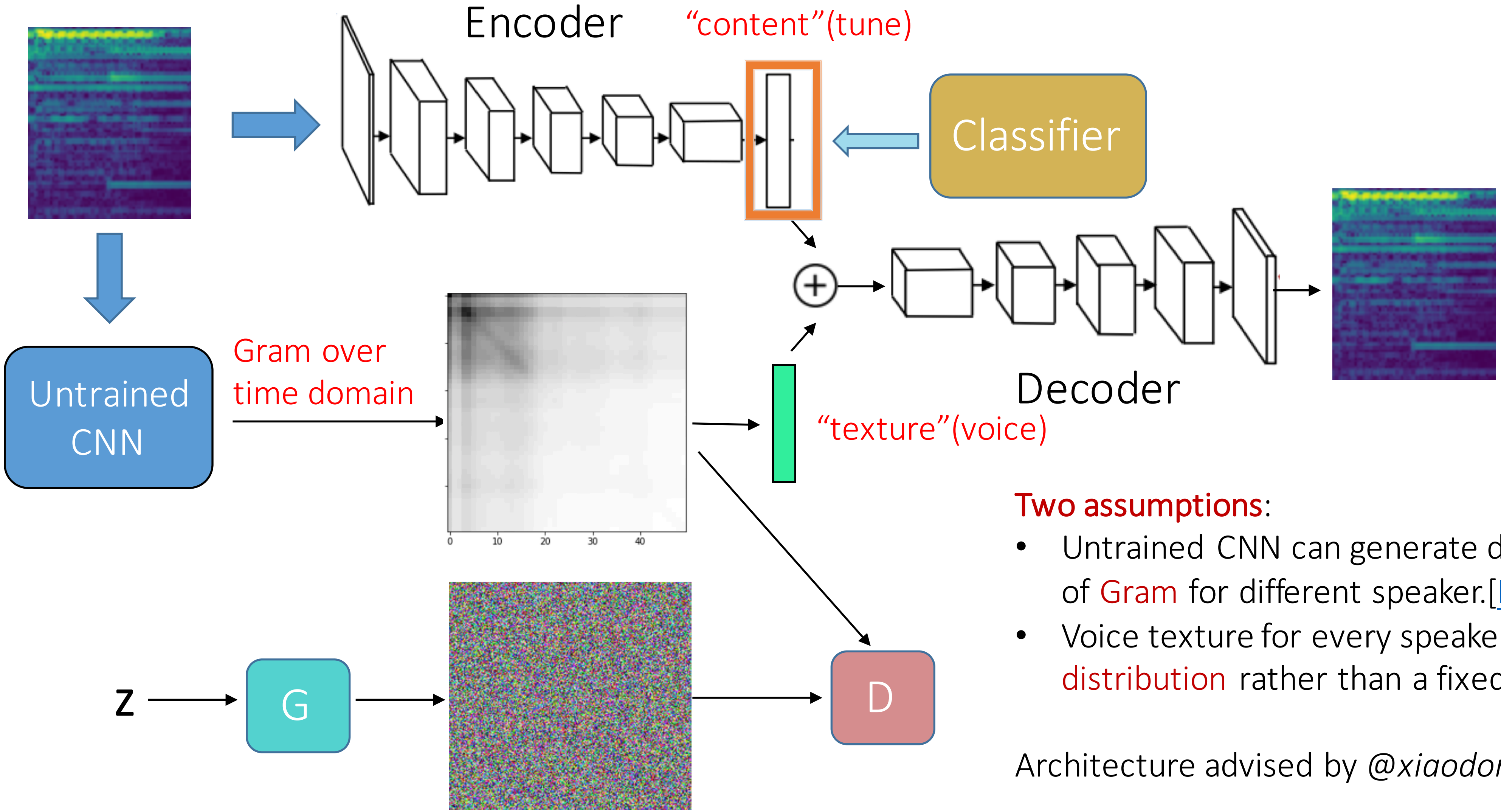
Speaker identification task of speakers in the VCTK dataset using gram-over-time-domain feature with untrained CNN:

Speakers	Train/Test	Accuracy
30	270/180	45.6%
4	240/160	92.5%

Implementation: https://oss.navercorp.com/ke-fang/Random-CNN-Practice

Samples: https://soundcloud.com/mazzzystar/sets/speech-conversion-sample

Representation disentangling model

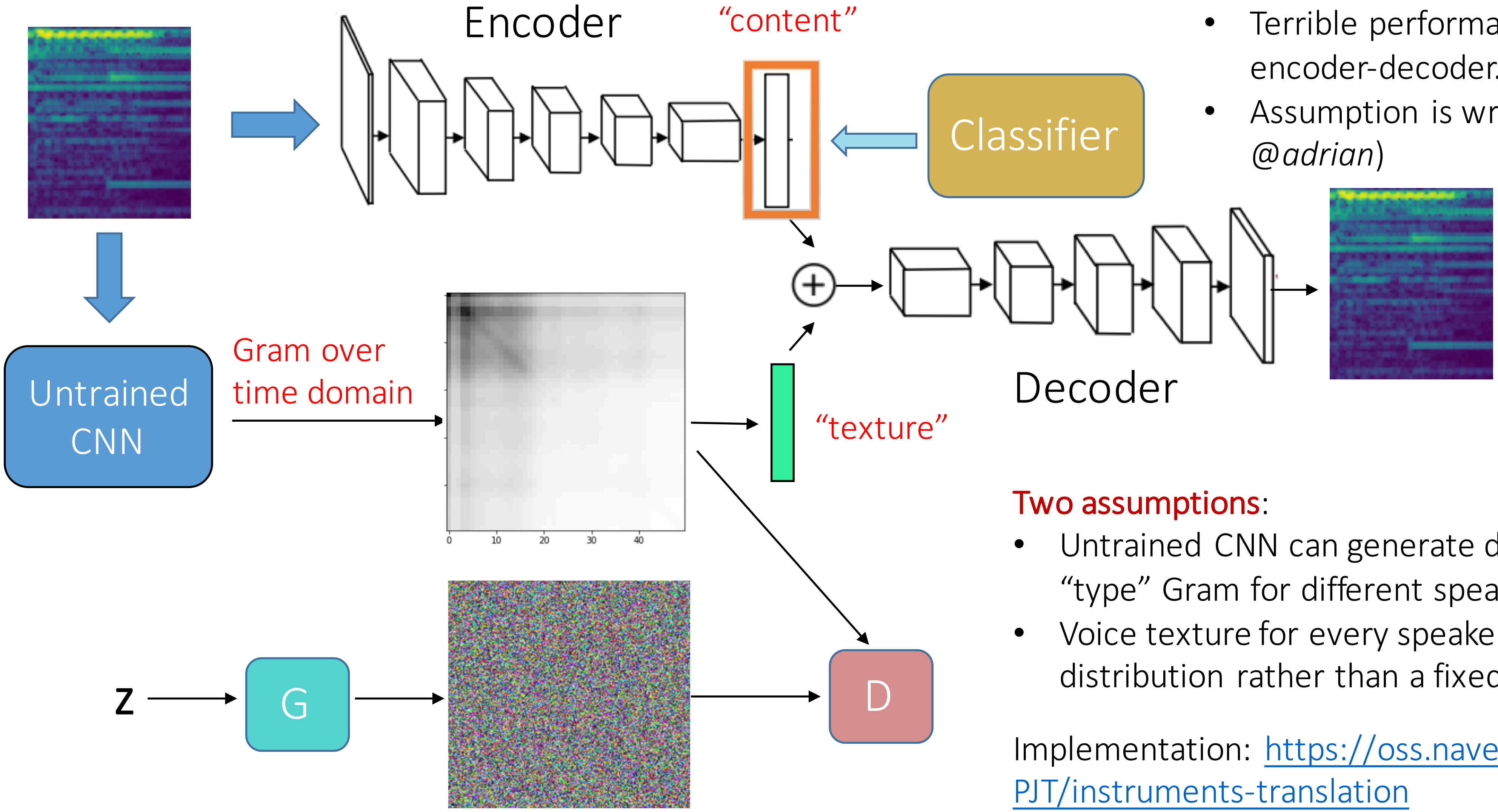


Two assumptions:

- Untrained CNN can generate different “type” of Gram for different speaker.[Inspiration]
- Voice texture for every speaker is a distribution rather than a fixed vector.

Architecture advised by *@xiaodong-gu*

Representation disentangling model



Quit because:

- Terrible performance of self-designed encoder-decoder.
- Assumption is wrong. (suggestion from *@adrian*)

Two assumptions:

- Untrained CNN can generate different "type" Gram for different speaker.
- Voice texture for every speaker is a distribution rather than a fixed vector.

Implementation: https://oss.navercorp.com/ClovaAI-PJT/instruments-translation

Continuous audio generation with GAN

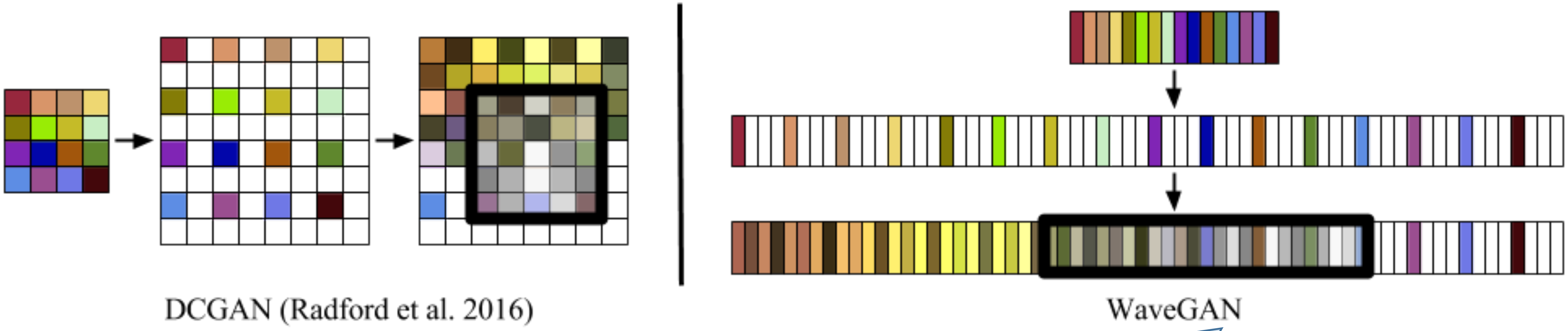
Goal

- Generate long-term dependency(music, speech) audio with GAN
- First we aim at long-term structure raw music audio generation
- If feasible, then try out speech generation.

Remarkable related work

- **WaveGAN** for synthesizing raw audio with GAN (Chris Donahue*, ICLR 2018 Workshop*)
- **C-RNN-GAN** for continuous recurrent neural networks with adversarial training (Olof Mogren, *NIPS 2016 Workshop*)
- A hierarchical latent vector model for learning long-term structure in music (**Google**, *ICML 2018*)

WaveGAN



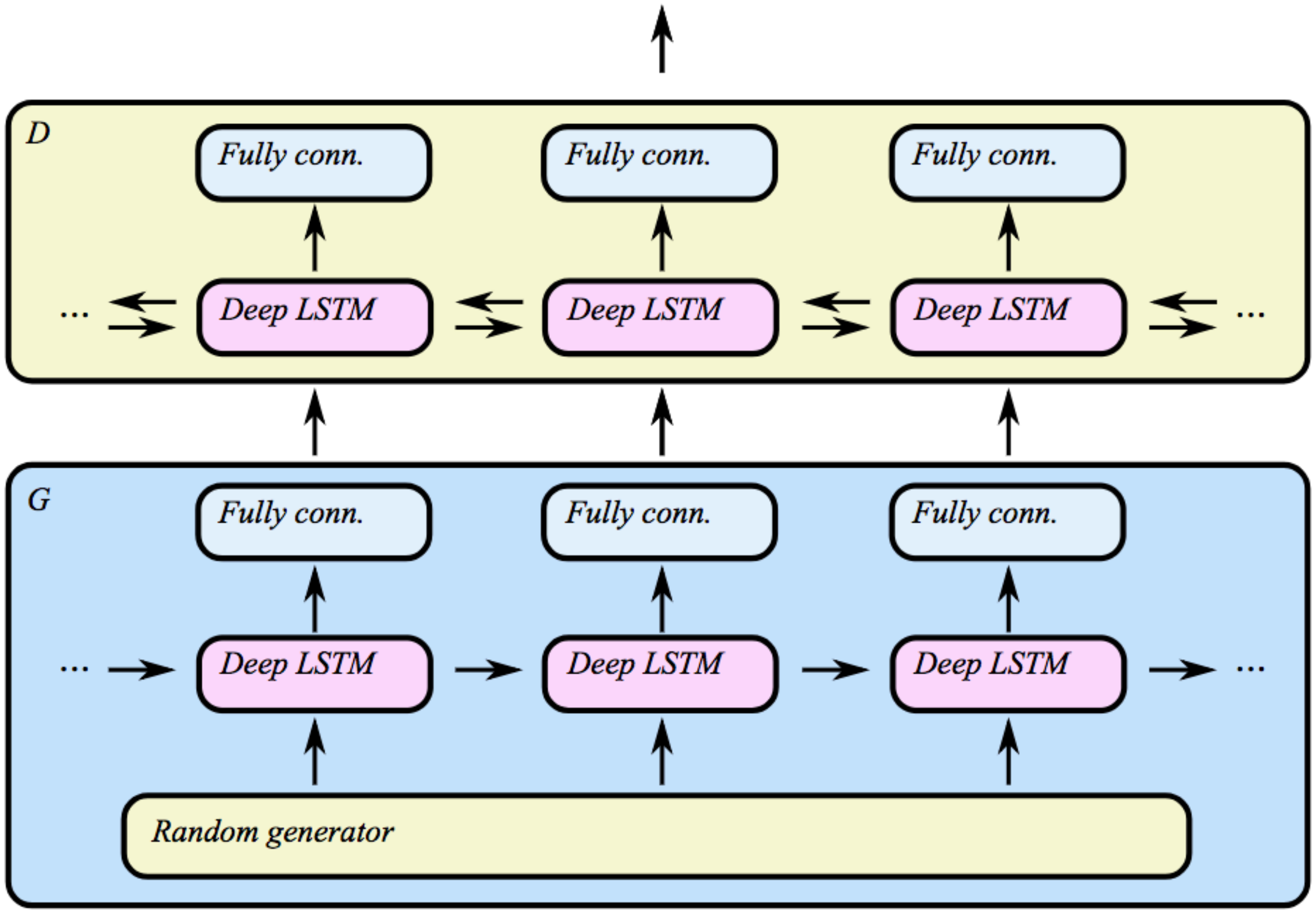
Implementation: https://oss.navercorp.com/CLAIR/WaveGAN-pytorch
Samples: http://wavegan-v1.s3-website-us-east-1.amazonaws.com/

Core technic:

- Use raw wav files, longer 1D filters of length 25 instead of 2D filters of size 5x5.
- Phase shuffle to prevent discriminator learn a trivial solution to rejects generated samples.

Chris Donahue, et al., "Synthesizing Audio with Generative Adversarial Networks", *ICLR 2018 Workshop*

C-RNN-GAN



Highlight:

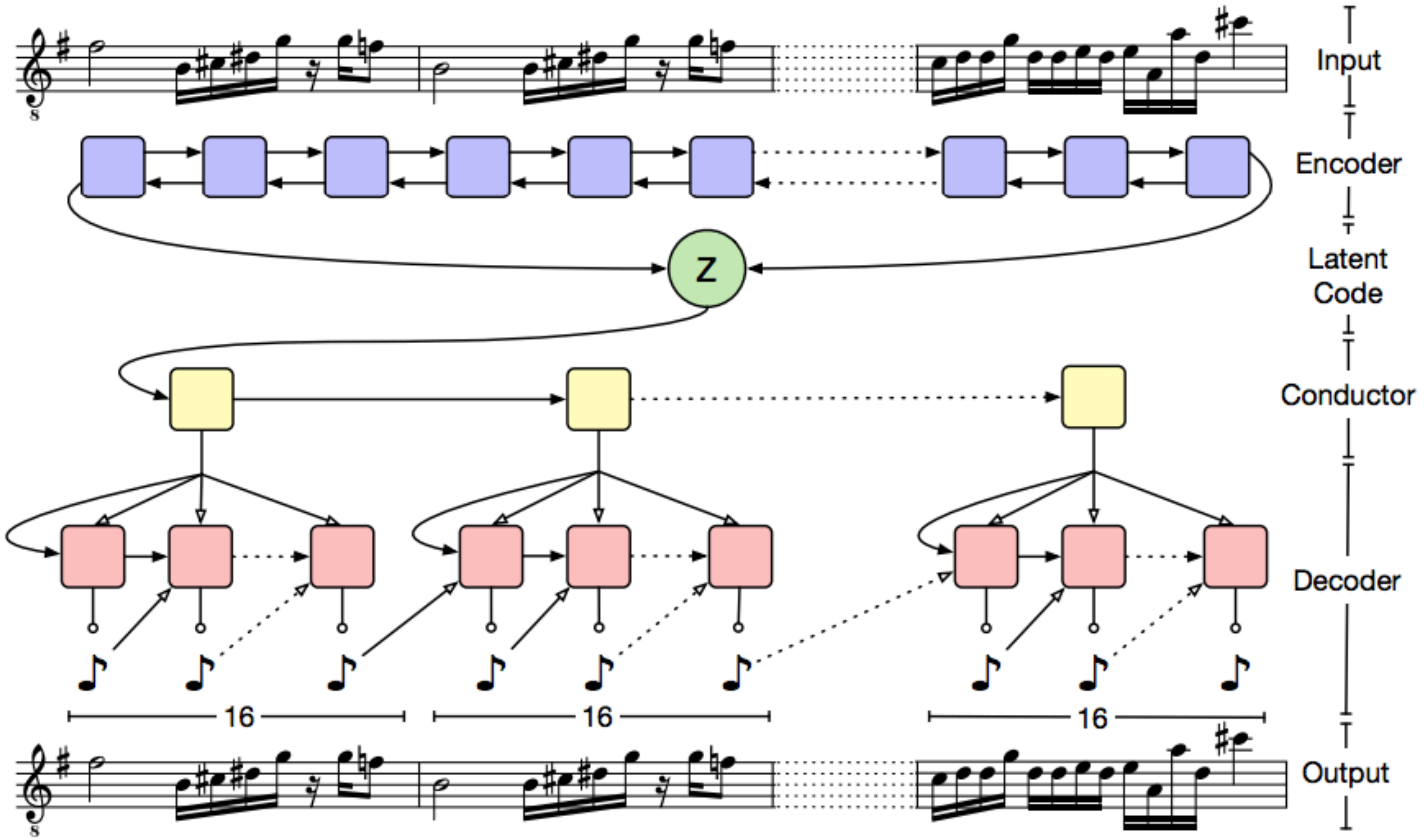
- Combine recurrent network with adversarial training by using **LSTM** for generator and **BiLSTM** for discriminator.

Drawback:

- Structure like LSTM restrict the length network can memorize.(Reason for using MIDI format)

Olof Mogren., "C-RNN-GAN: Continuous recurrent neural networks with adversarial training", *ICLR 2018 Workshop*

Hierarchical VAE for long-term structure music



Highlights:

- BiLSTM Encoder encodes an entire sequence to a single latent vector
- Hierarchical Decoder by a two-layer unidirectional LSTM(Simple RNN resulted in vanishing influence of the latent state.)
- Multi-Stream Modeling with 3 separate distributions over output tokens(drum, bass, and melody), use a separate decoder RNN for each instrument.

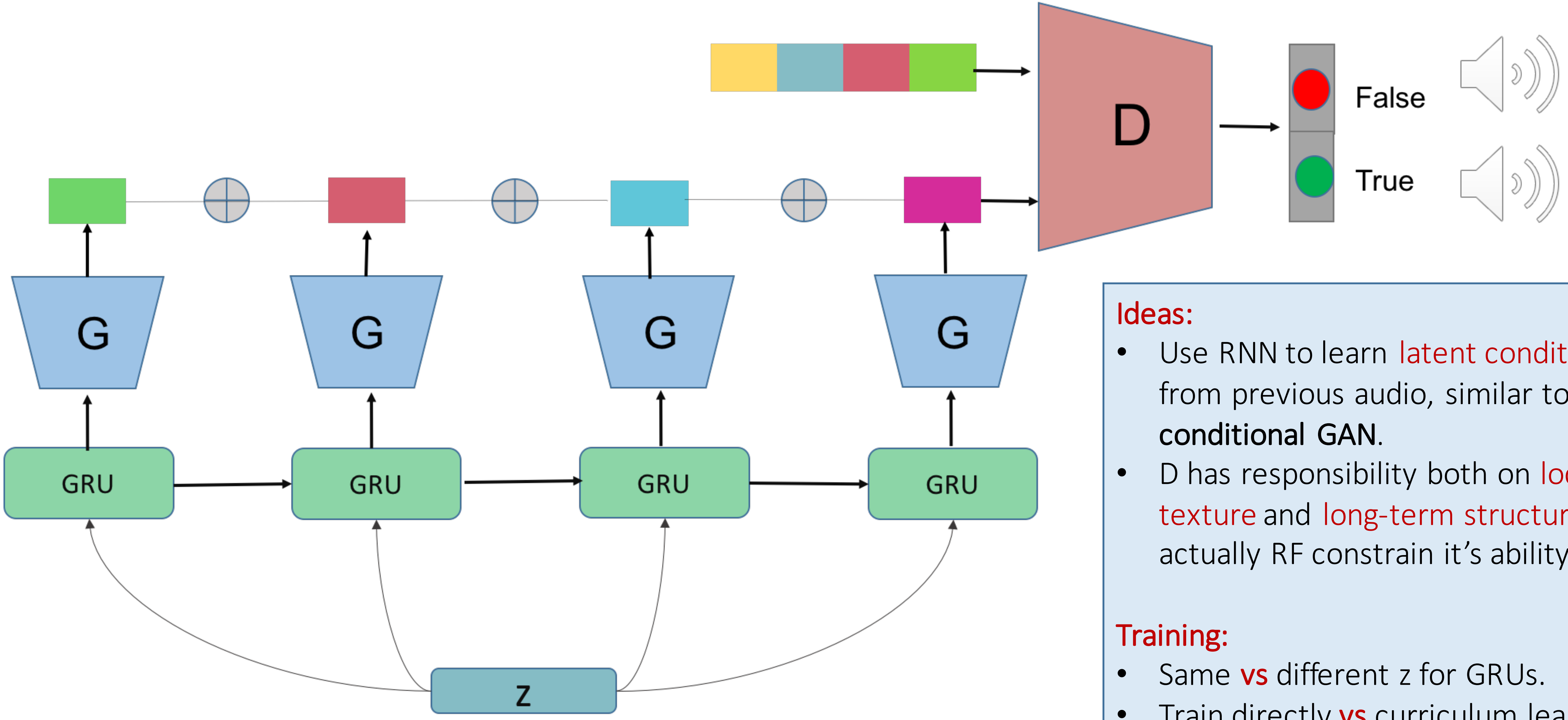
Adam Roberts., "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music", *ICML 2018*

Unmentioned related works

- **AMAE**: Modelling raw audio at scale to generate long-term structure of music. (**DeepMind**, *Jun 2018*)
- **MidiNet** for MIDI generation (Li-Chia Yang , *ISMIR 2017*)
- Semi-Recurrent CNN-based VAE-GAN(Mohammad Akbari, *ICASSP 2018*)
- **MuseGAN**: Multi-track MIDI generation(Hao-Wen Dong, *AAAI 2018*)
- Language Generation with Recurrent Generative Adversarial Networks without Pre-training (Ofir Press, *ICML 2017 Workshop*)
- **SampleRNN** end-to-end audio generation (Soroush Mehri, *ICLR 2017*)

My work on continuous audio generation with GAN

Maybe..MusicGAN?



Ideas:

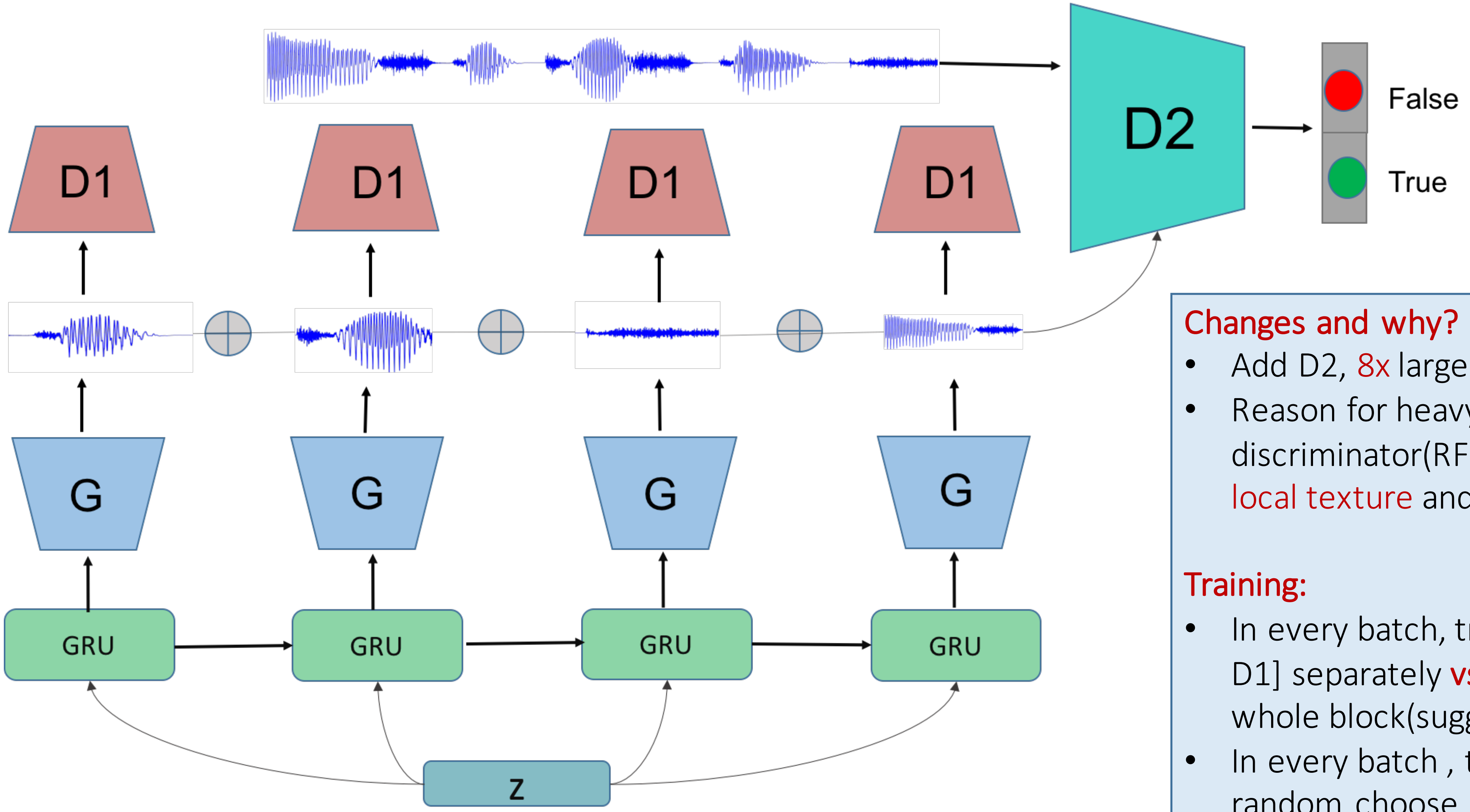
- Use RNN to learn latent condition from previous audio, similar to **conditional GAN**.
- D has responsibility both on local texture and long-term structure(But actually RF constrain it's ability).

Training:

- Same **vs** different z for GRUs.
- Train directly **vs** curriculum learning.

Implementation: https://oss.navercorp.com/ke-fang/rgan

MusicGAN: Version 0.2



Implementation: https://oss.navercorp.com/ke-fang/musicGAN

Changes and why?

- Add D2, 8x larger receptive field than D1.
- Reason for heavy work with only 1 discriminator(RF=0.5s) to deal with both local texture and long-term structure.

Training:

- In every batch, train [RNN, D2] and [G, D1] separately **vs** see [RNN, G] as a whole block(suggestion from @*hseok*).
- In every batch , train every seg for D1 **vs** random choose seg to train.

MusicGAN: Future work

- Improve current architecture for better quality and music sounds feeling.
- Comparison experiments with others work on long-term dependency music.
- Explore of the possibility for speech synthesis.

Q&A