Neurally Optimized Decoder for Low Bitrate Speech Codec

H. Y. Kim, J. W. Yoon, W. I. Cho, N. S. Kim,

IEEE Signal Processing Letters, 2021.

Jan. 11. 2022. (TUE)

Youngwon Choi

GIST SAPL

Speech and Audio Processing Lab.

Contents

1. **Introduction**

2. **Neurally Optimized Decoder**
   - A. Speech Coding Overview
   - B. Speech Production Model Based on GAN
   - C. Dequantization Network

3. **Experiments**
   - A. Database of Speech and Codecs
   - B. Model and Training Configuration
   - C. Evaluations

4. **Conclusion**

1. Introduction

Introduction

□ Speech coding algorithms are applied to obtain compact digital representations of high-fidelity speech signals for efficient transmission.

□ Conventional speech codec comprises an encoder that converts the input speech signal into a compact bitstream and a decoder that inversely reconstructs the speech signal from the bitstream.

□ Advanced speech synthesis techniques using deep generative models leads to its wide usage for speech codecs based on neural network such as WaveNet-based neural codecs [1], [2].

[1] C. Garbacea, A. v. d. Oord, Y. Li, F. S. Lim, A. Luebs, O. Vinyals, and T. C. Walters, "Low bit-rate speech coding with VQ-VAE and a WaveNet decoder," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2019, pp. 735–739.
[2] F. S. Lim, W. B. Kleijn, M. Chinen, and J. Skoglund, "Robust low rate speech coding based on cloned networks and wavenet," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2020. pp. 6769–6773.

Introduction

□ Unlike the neural codecs that train both encoder and decoder, neural decoder methods [3], [4] train only the decoder to reconstruct the input speech signal from reconstructed acoustic features by decoding algorithm of the speech codec.

- Without changing the encoder, these approaches are readily applicable approach to existing telecommunication systems.

□ However, the neural decoder requires some information for decoding such as the bit allocation or the dequantization scheme that depends on speech codec.

- Failing to provide a universal solution for many different kinds of speech codecs.
- *"What will it be like if we directly reconstruct the speech from the bitstream using a neural network without any specific knowledge of the speech codec?"*

□ In this letter, authors propose a neutrally optimized decoder which can be flexibly applied to diverse speech codecs.

- Authors view the decoder as a conditional generator model where the bitstream is given as the conditioning variable, and the speech signal is treated as the target output.

[3] W. B. Kleijn, F. S. Lim, A. Luebs, J. Skoglund, F. Stimberg, Q. Wang, and T. C. Walters, "WaveNet based low rate speech coding," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process., 2018, pp. 676–680.
[4] J.-M. Valin and J. Skoglund, "A Real-Time Wideband Neural Vocoder at 1.6kb/s Using LPCNet," in Proc. Interspeech, 2019, pp. 3406–3410.

Introduction

□ The main contributions of this work are summarized as follows:

- Author propose a neurally optimized decoder based on generative model to reconstruct the input speech from the bitstream directly.
- Author propose **a dequantization network** to group and dequantize the related bits without any information of the original speech codec.
- The experimental results show that the proposed neural decoder is generally applicable for many different kinds of speech codecs and provides better subjective and objective quality than the original decoder.

2. Neurally Optimized Decoder

Neurally Optimized Decoder

A. Speech Coding Overview

□ For low bitrate coding, parametric speech coder at rates below 4kb/s compresses the core acoustic features of speech rather than the waveform itself.

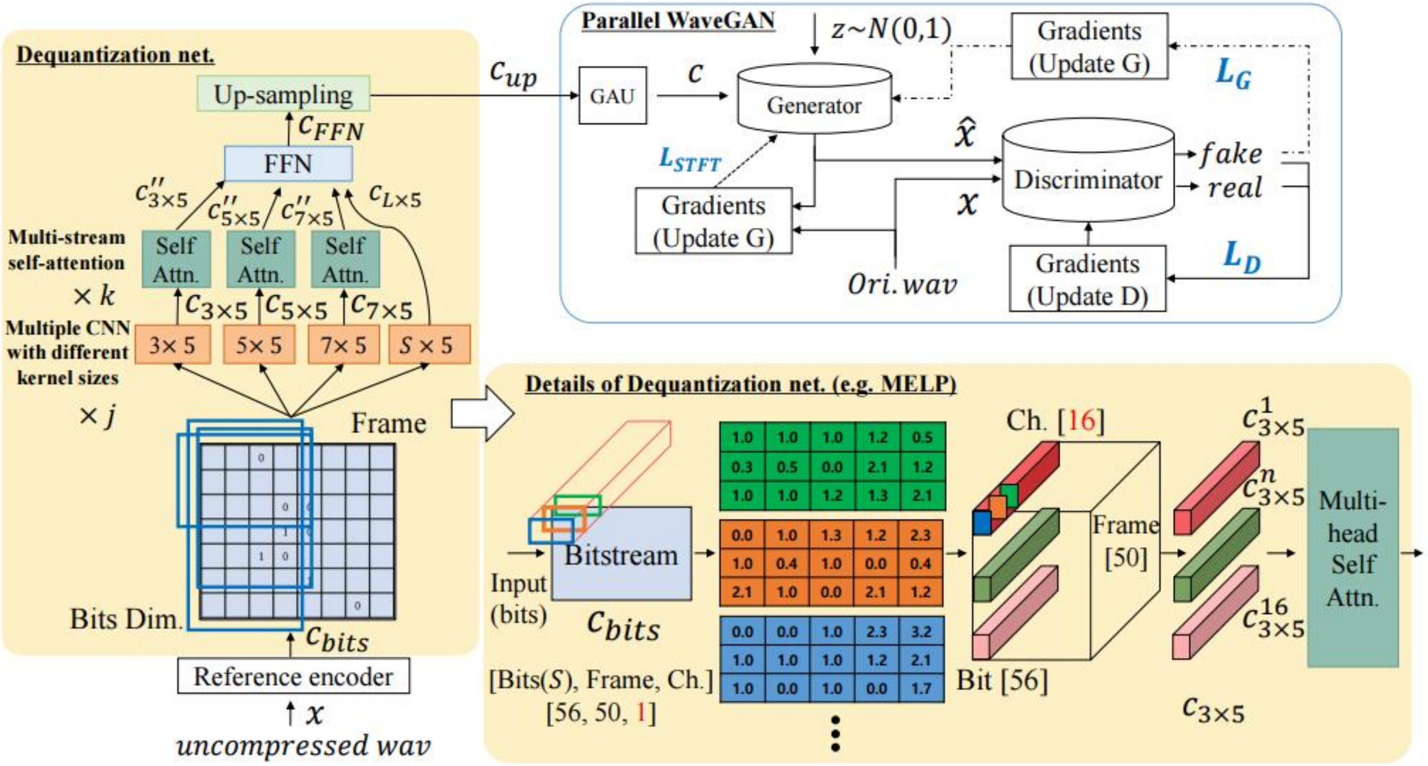
- So, for the reconstruction of the speech at the receiver, the conventional codec needs full knowledge on the bit allocation, transmission order, and quantization for each feature.

□ For designing a decoder without such a prior knowledge, three issues have to be resolved to reconstruct the input speech from the bitstream.

- We need to find out the relationship between bits without explicit knowledge on the bit allocation and transmission order.
- We also need to de-quantize the bitstream without a prior knowledge on the quantization scheme.
- Finally, the speech waveform can be recovered from the de-quantized bitstream without any information on the speech production model used in the codec.

Neurally Optimized Decoder

□ The proposed model consists of the generative neural vocoder and the dequantization network.



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

Fig. 1: Proposed model architecture

Neurally Optimized Decoder

B. Speech Production Model Based on GAN

□ It has been reported that neural vocoders based on generative model produce high-quality speech conditioned on various types of compressed features such as mel-spectrogram, LPC coefficients, and even discrete representations.

□ The proposed model is based on Parallel WaveGAN [5] which introduced high-quality speech generation with fast inference.

- Parallel WaveGAN was trained with conditional GAN (cGAN) framework where the generator (Wavenet) produces the speech conditioned on mel-spectrogram to fool the discriminator D.
- On the other hand, the discriminator is a binary classifier that decides whether an input sample is a real or generated speech.

[5] E. Song, R. Yamamoto, M.-J. Hwang, J.-S. Kim, O. Kwon, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.. 2020, pp. 6199–6203.

Neurally Optimized Decoder

□ The modified loss functions are

- $L_D = -\mathbb{E}_{x\sim\mathbb{P}_{ori}(x)}[logD(x)] - \mathbb{E}_{z\sim\mathbb{P}_z(z),c\sim\mathbb{P}_{bits}(c)}\left[\log\left(1 - D\left(G(z,c)\right)\right)\right]$
- $L_G = \lambda_{adv}\mathbb{E}_{z\sim\mathbb{P}_z(z),c\sim\mathbb{P}_{bits}(c)}\left[\log(1 - D\left(G(z,c)\right))\right] + \frac{1}{M}\Sigma_{m=1}^{M} L_{STFT}^{m}\left(x, G(z,c)\right)$

□ L_{STFT} includes the spectral convergence loss and log STFT magnitude loss [5].

- $L_{STFT}(x,\tilde{x}) = L_{sc}(x,\tilde{x}\ ) + L_{mag}(x,\tilde{x})$
- $L_{sc}(y,\tilde{y}_2) = \frac{\||STFT(x)|-|STFT(\tilde{x})|\|_F}{\|STFT(x)\|_F}$
- $L_{mag}(y,\tilde{y}_2) = \frac{1}{T}\|\log|STFT(y)| - \log|STFT(\tilde{y}_2)|\|_1$

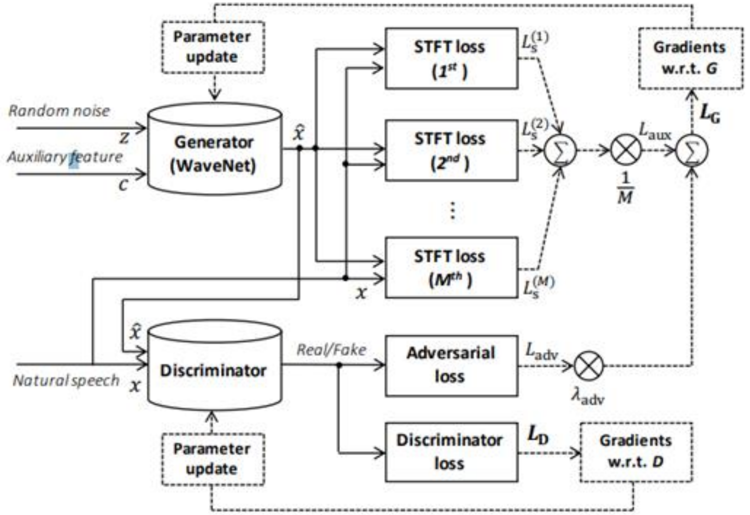


Fig. 1: An illustration of our proposed adversarial training framework with the multi-resolution STFT loss.

[5] E. Song, R. Yamamoto, M.-J. Hwang, J.-S. Kim, O. Kwon, and J.-M. Kim, “Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram,” in Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.. 2020, pp. 6199–6203.

Neurally Optimized Decoder

C. Dequantization Network

□ Compared to the speech generation conditioned on mel-spectrogram, this task is more challenging in that it requires restoring the input speech from more compressed features.

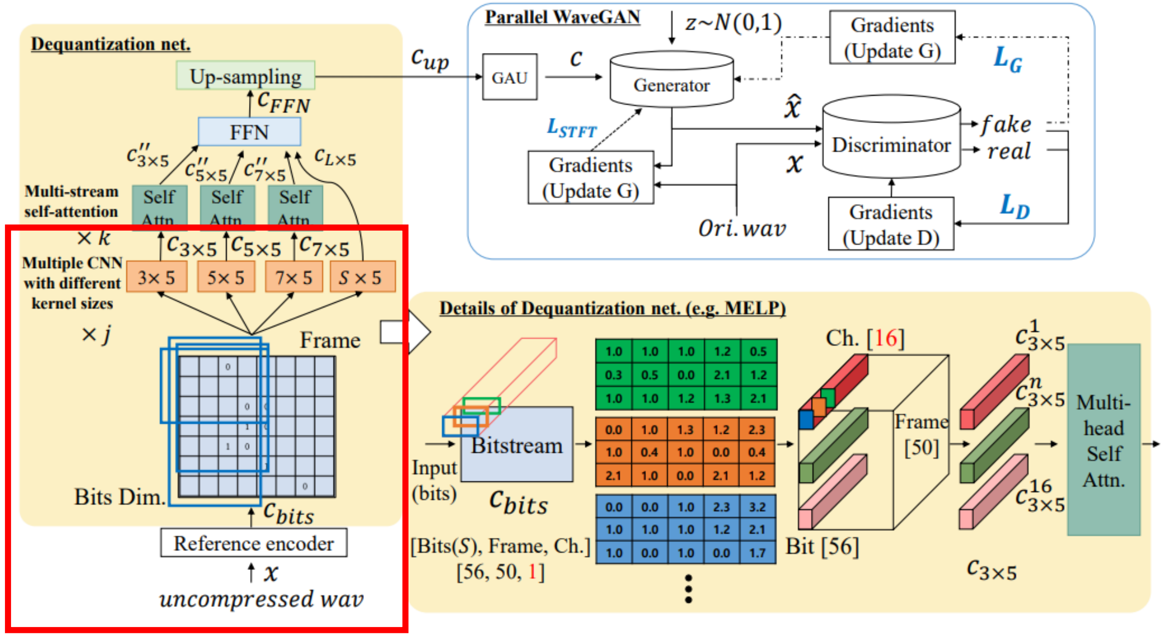
□ Authors propose a dequantization network that finds related bits and de-quantize them accurately.

□ The dequantization network has a multiple convolutional neural network (CNN) layer with different kernel sizes and a multi-stream self-attention layer.

Neurally Optimized Decoder

□ Multiple CNN Layer

- The uncompressed wav x is converted to 2D bitstreams c_{bits} $(Bits * Frame)$.
- In order to group the quantized bits from the same speech feature, the proposed network makes various combinations of bits using CNN filters with different kernel sizes.
- The multiple CNN structure reflects different receptive fields that can extract local embeddings by 2D convolution and global embeddings by 1D convolution.



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

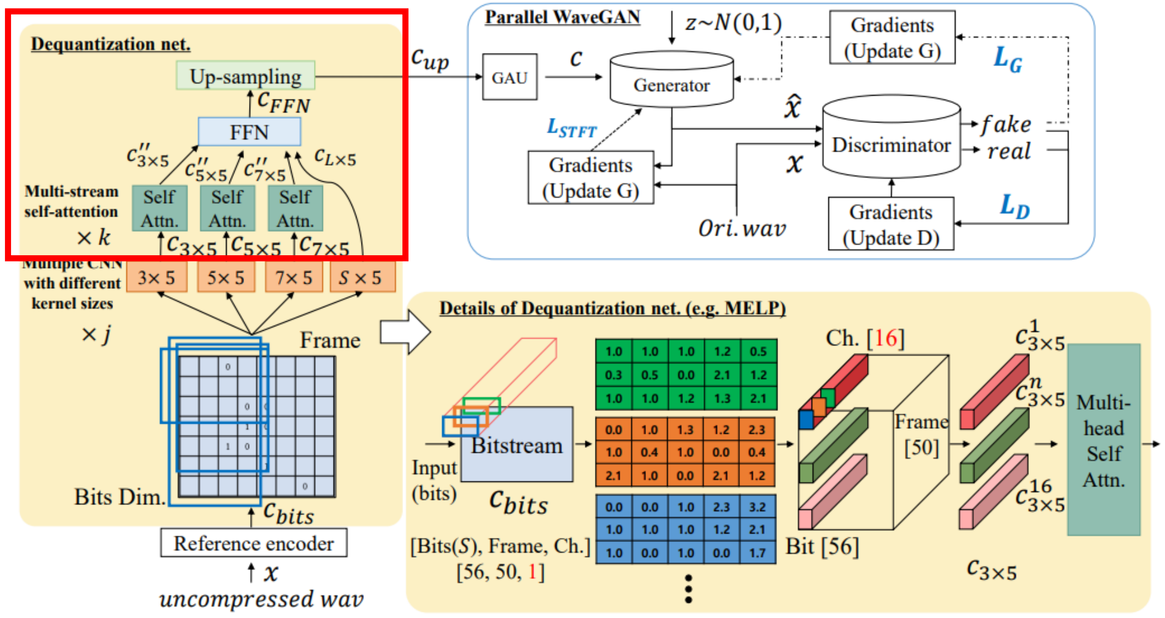
⋮

Fig. 1: Proposed model architecture

Neurally Optimized Decoder

□ Multi-head Self-attention [6]

- The output embeddings of the 2D CNN layers represent position-wise local features, but dequantization also requires long-range relations among the local features.
  - Furthermore, the bit allocation differs with the voiced-and unvoicedness of the frame.
- To attack the two problems, authors apply a multi-head self-attention for each local embedding individually to capture long-range relationships from different representation subspaces.



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

⋮

Fig. 1: Proposed model architecture

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998-6008.

Neurally Optimized Decoder

□ Multi-head Self-attention [6]

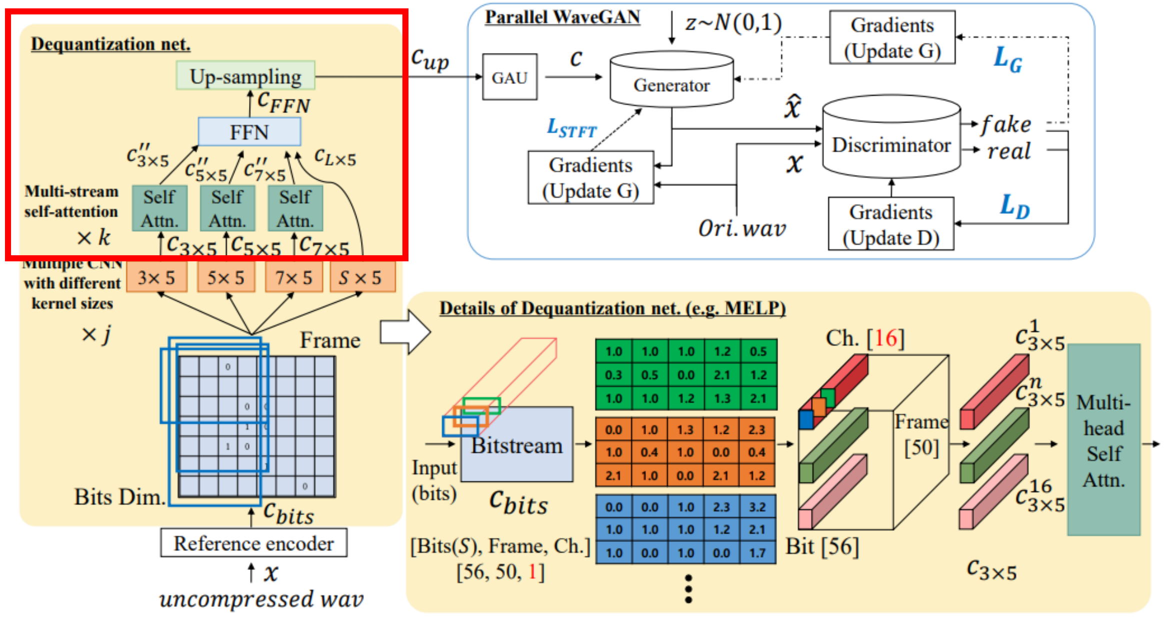
– Authors apply a 2D positional encoding to provide the same bits pattern with different representations

- $\widetilde{c_i} = MHSA\left(PE(c_i)\right)$
- $c_i' = Layernorm(c_i + \widetilde{c_i})$
- $c_i'' = Layernorm\left(c_i' + FFN(c_i')\right)$
- $c_{FFN} =$ $FFN_2\left(FFN_1\left(concat(c_{3\times5}'', c_{5\times5}'', c_{7\times5}'', c_{S\times5})\right)\right)$

  – PE : Sinusoidal positional encoding

  – $MHSA$: Multi-head self attention

  – FFN : Feed forward network



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

⋮

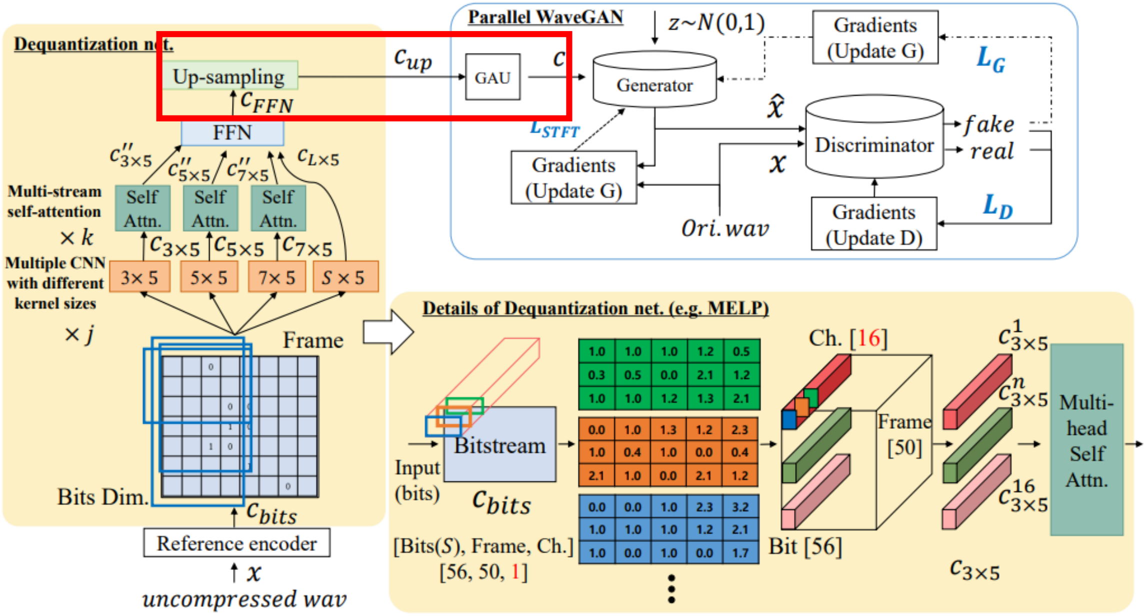
Fig. 1: Proposed model architecture

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998-6008.

Neurally Optimized Decoder

□ Gated Activation Unit (GAU)

- Authors used a gated activation unit (GAU) for the conditioning, which decides the amount of the information of each embedding vector used to reconstruct the original speech at each frame.
  - $c_{up} = Upsample(c_{FFN})$
  - $c = tanh(W_f * c_{up}) \odot \sigma(W_g * c_{up})$
    - σ : sigmoid function
    - ⊙ : element-wise product



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

⋮

Fig. 1: Proposed model architecture

3. Experiments

Experiments

A. Database of Speech and Codecs

□ Database

- Speech database : TIMIT
  - Train: 4,620 utterances, 462 speakers
  - Test: 1,620 utterances, 162 speakers
  - Downsampled from 16kHz to 8kHz
- Noise dataset: ITU-TP.501 noise dataset
  - 12 different noises
  - Artificially generated with noise type
  - SNR value randomly chosen from 0 to 20 dB
- Bitstreams are produced by putting randomly sampled one second of the speech into the encoder of each codecs.

□ Codecs

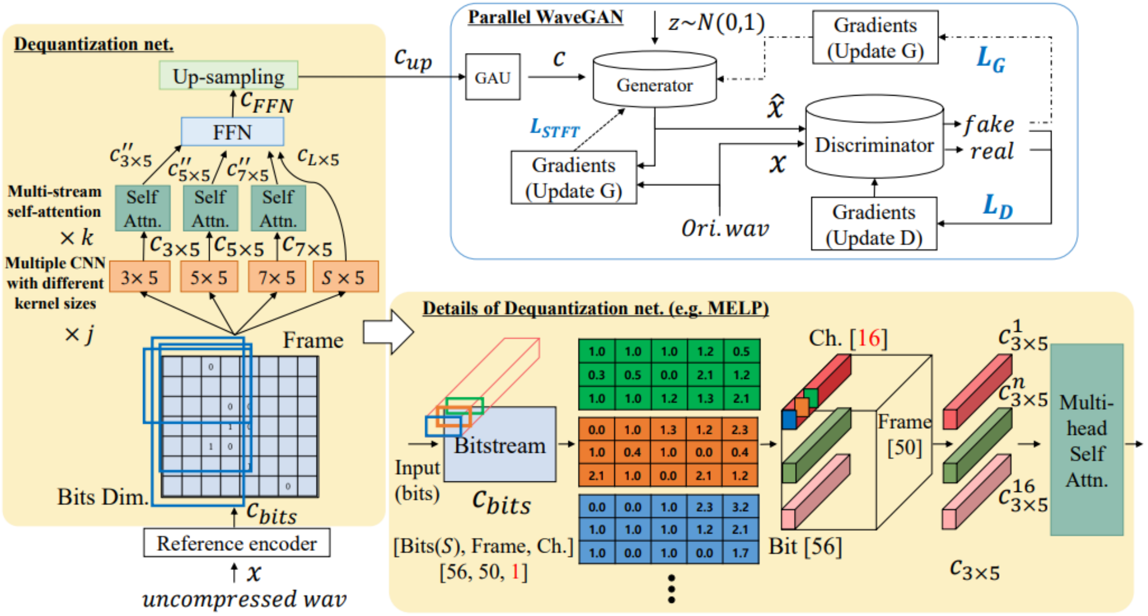
- MELP
  - Bitrates: 2400b/s
- AMBE
  - Bitrates: 2400b/s
- SPEEX
  - Bitrates: 2150b/s

Experiments

B. Model and Training Configurations

□ Model

- CNN
  - j=2, with the output channel size of 8, 16.
- Attention parameters
  - k=3, d_{head}=4, d_{model} and d_{ff}=16.
- FFN layers
  - 2 with the output dimension of 1024, 56.
- Up-sampling
  - Up-sample scales to [3,3,5,4] for MELP
  - [4,4,5,2] for AMBE and SPEEX



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

⋮

Fig. 1: Proposed model architecture

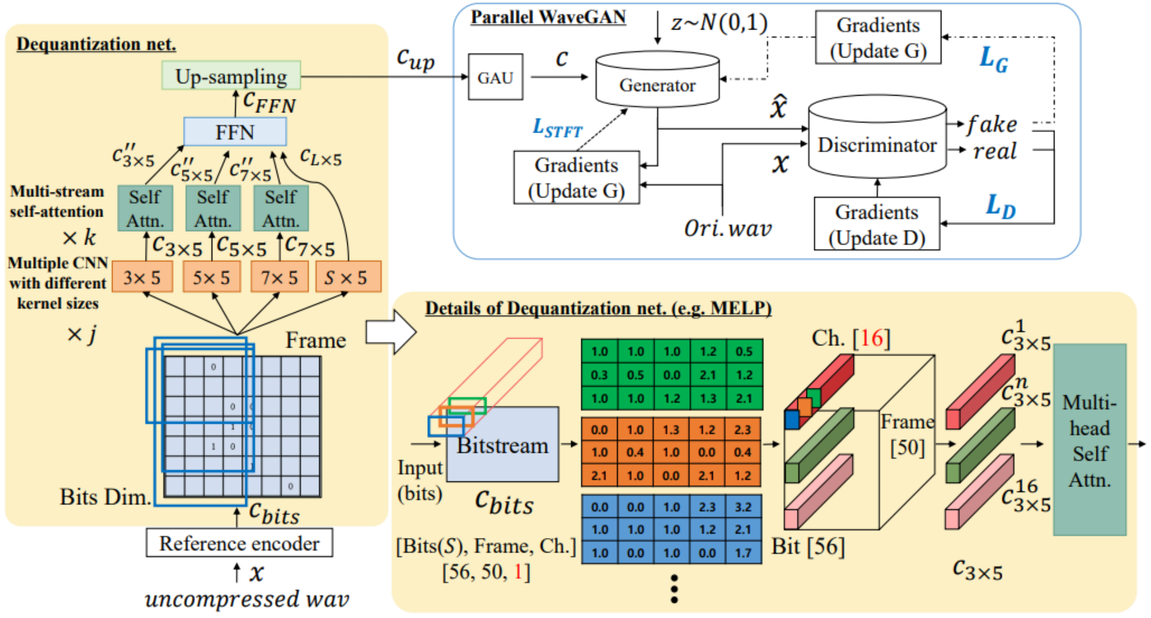
Experiments

B. Model and Training Configurations

□ Training

- Multi-resolution STFT loss
  - Setting followed [5], with halving the FFT size, window size, and frame shift
- Length of each audio clip : 1 second
- Batch size: 12
- Total 800K steps
  - Generator solely trained for first 100K steps
- Learning rate
  - $5e^{-4}$ for Generator (Initial)
  - $5e^{-5}$ for Discriminator (Initial)
  - Reduced by half for 75K steps for generator
  - Reduced by half for 100K steps for discriminator

- Other setting followed [5]



1.0	1.0	1.0	1.2	0.5
0.3	0.5	0.0	2.1	1.2
1.0	1.0	1.2	1.3	2.1

0.0	1.0	1.3	1.2	2.3
1.0	0.4	1.0	0.0	0.4
2.1	1.0	0.0	2.1	1.2

0.0	0.0	1.0	2.3	3.2
1.0	1.0	1.0	1.2	2.1
1.0	0.0	1.0	0.0	1.7

Fig. 1: Proposed model architecture

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention Is All You Need,” in Proc. Adv. Neural Inf. Process. Syst., 2017, pp. 5998-6008.

Experiments

C. Evaluations

□ Ablation Study

- Compared to the 1D convolution network, the PESQ score of the multiple CNN for various local embeddings improved from 3.06 to 3.25, beating the PESQ score of the reference decoder.
- The multi-head self-attention network with the multiple CNN showed the best PESQ score, which means that more complex relationship between the local embeddings can be represented.
- For the comparison of upper-bound performance, authors evaluated 1D convolution network with mel feature in neural vocoder for speech synthesis.

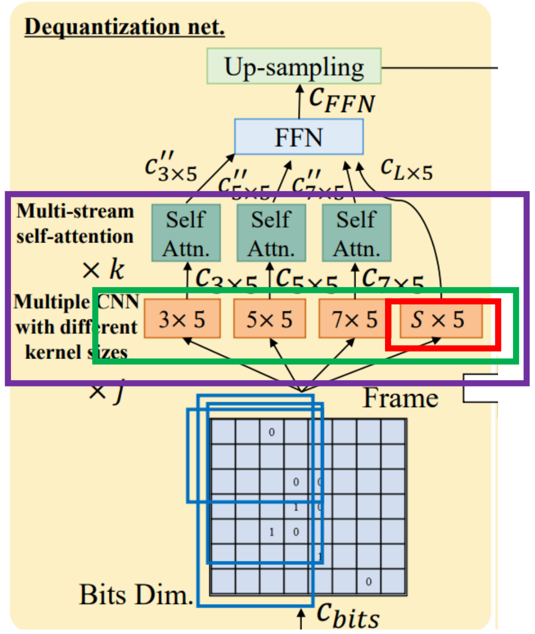
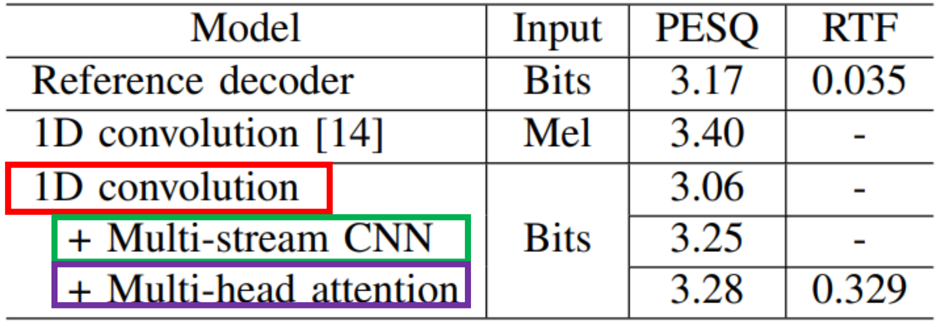


TABLE I: PESQ and RTF results for ablation study on MELP.



Model	Input	PESQ	RTF
Reference decoder	Bits	3.17	0.035
1D convolution [14]	Mel	3.40	-
1D convolution	Bits	3.06	-
+ Multi-stream CNN		3.25	-
+ Multi-head attention		3.28	0.329

Experiment

□ Objective Evaluation

– The proposed decoder improves PESQ in both clean and noisy condition for all codecs.

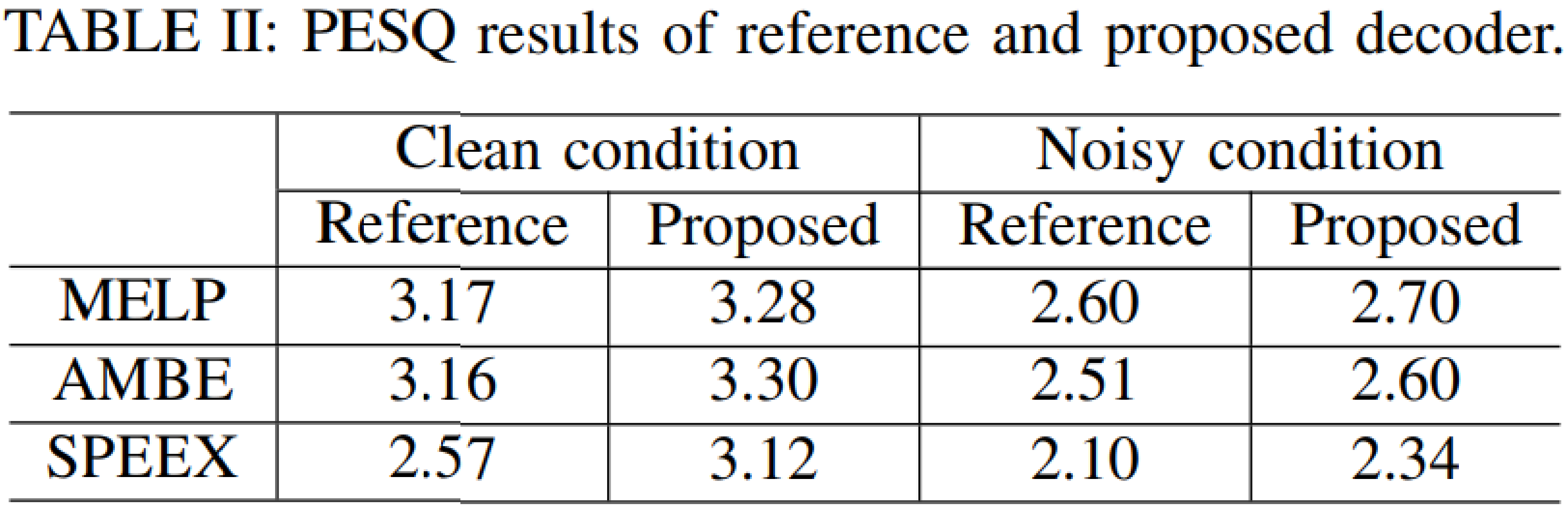


TABLE II: PESQ results of reference and proposed decoder.

	Clean condition		Noisy condition	
	Reference	Proposed	Reference	Proposed
MELP	3.17	3.28	2.60	2.70
AMBE	3.16	3.30	2.51	2.60
SPEEX	2.57	3.12	2.10	2.34

Experiments

□ Subjective Test

– Authors used modified MUSHRA test, which provides a labelled original speech to listeners.
  - ▪ 10 speech-expert listeners evaluated eight utterances from the TIMIT test set.

– The results show that the proposed neural decoder outperforms each reference decoder regarding the subjective quality without information of the reference codecs.
  - ▪ SPEEX (2150b/s) enhanced by the proposed decoder shows a subjective score around AMBE and beyond MELP (both 2400b/s).

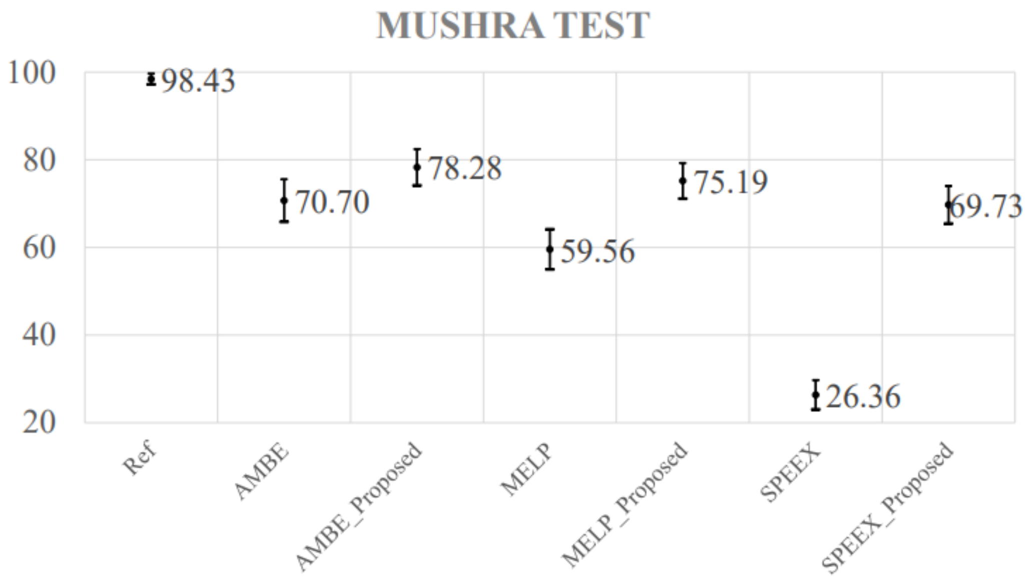


Fig. 2: Subjective quality test (MUSHRA scores).

Experiments

□ Speaker Transparency

- In order to assess speaker similarity, authors evaluated the comparison mean opinion score (CMOS) based on subjective listening tests.
- Randomly selected 30 pairs of utterances from TIMIT test set were rated by 15 speech-expert listeners.
- For all cases, the speaker similarity CMOS for the proposed decoder is higher than that of each reference decoder.

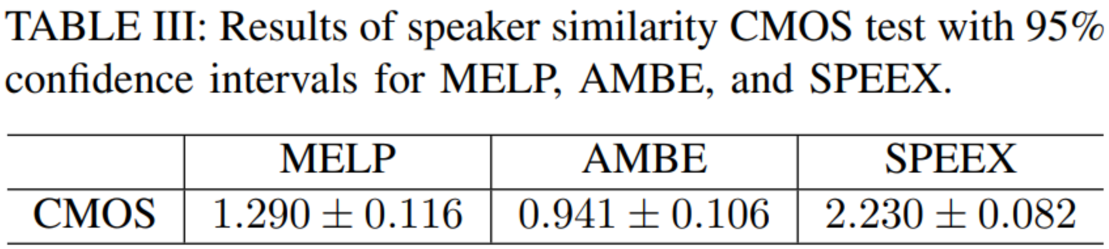


TABLE III: Results of speaker similarity CMOS test with 95% confidence intervals for MELP, AMBE, and SPEEX.

	MELP	AMBE	SPEEX
CMOS	1.290 ± 0.116	0.941 ± 0.106	2.230 ± 0.082

4. Conclusion

Conclusion

□ In this letter, authors proposed a neutrally optimized decoder, which directly reconstructs the original speech from bitstream without any information about the original codec.

□ To find out the relationship between bits and to de-quantize the bits without a prior knowledge, authors introduced a dequantization network including multiple CNN and multi-head self-attention layers.

□ Motivated by the high-quality speech restoration by generative model, the proposed decoder trained the Parallel WaveGAN conditioned on the embedding vector from the dequantization network.

□ From the results, the proposed neural decoder is generally applicable to various speech codecs and successfully reconstructs the original speech even outperforming the reference decoders in most cases.

Thank you for listening
Q & A