Speech Enhancement
for Low Bit Rate Speech Codec

Ju Lin, Kaustubh Kalgaonkar, Qing He, Xin Lei

ICASSP, 2022.

May. 20. 2022. (FRI)

Youngwon Choi

GIST SAPL
Speech and Audio Processing Lab.

Contents

1. Introduction

Introduction

- □ Speech codecs typically compress speech signal to compact bitstream by using hand-crafted features t hat eliminate redundant and/or unnecessary information.

- □ Traditional parametric coding of speech facilitates low rate but the resulting speech often sounds with a robotic character.

- □ Recently, deep learning techniques have been introduced to mitigate the limitations of traditional low bit rate speech codecs.
  - These approaches may be classified into end-to-end and neural augmented speech codecs.

- □ More recently, generative adversarial networks (GAN) have been applied into speech codecs with non-autoregressive decoders that can generate high-quality speech with lower computational cost.

Introduction

- In this paper, authors propose a neural extension to Codec2 (a parametric codec designed for low bit-rates).
  - This work is an attempt to explore if it is possible to enhance the output of existing low bit rate codecs using some additional information provided in form of embeddings.
  - In addition, the proposed neural enhancement does not break the existing speech coder, which could be also desirable.

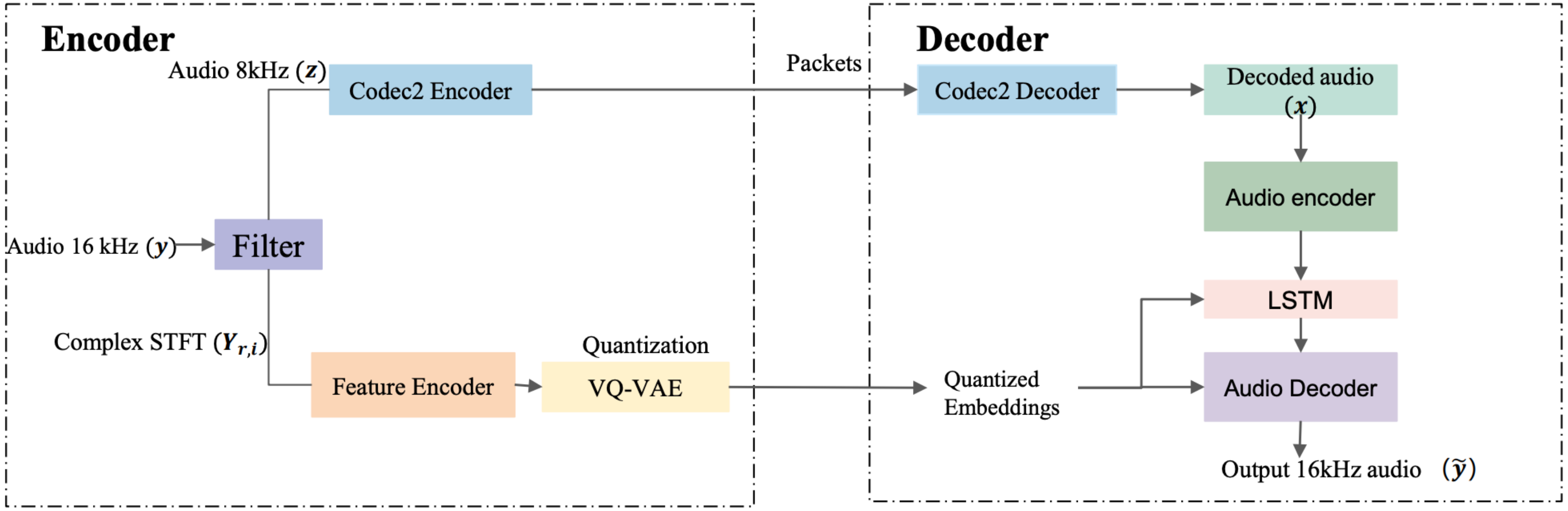


Fig. 1. The proposed neural extension framework to Codec2.

2. Proposed Approach

End-To-End Neural Audio Coding

Overview

□ The codec encoder consists of two branches; the first branch works on 8kHz speech signal and compresses the audio using the Codec2 encoder, the second branch uses the fullband(wideband?) speech signal to extract the compressed neural embeddings.

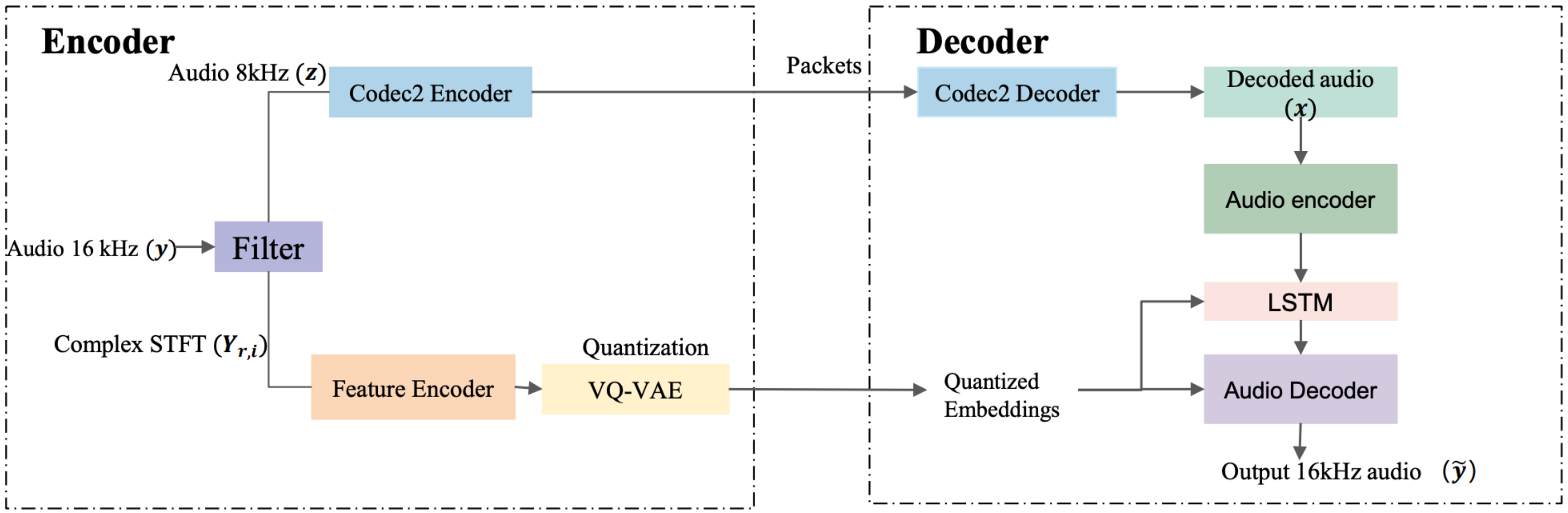


Fig. 1. The proposed neural extension framework to Codec2.

Experiments

2.1 Codec2

- Codec2 is an open-sourced parametric speech coder, which belongs to the sinusoidal coder family and can run at various update rates from 450bps to 3.2kbps.

- Codec2 operates on narrow-band speech with a sampling rate of 8 kHz.

- In this paper, authors use Codec2 at 1.2kbps and 2.4kbps.

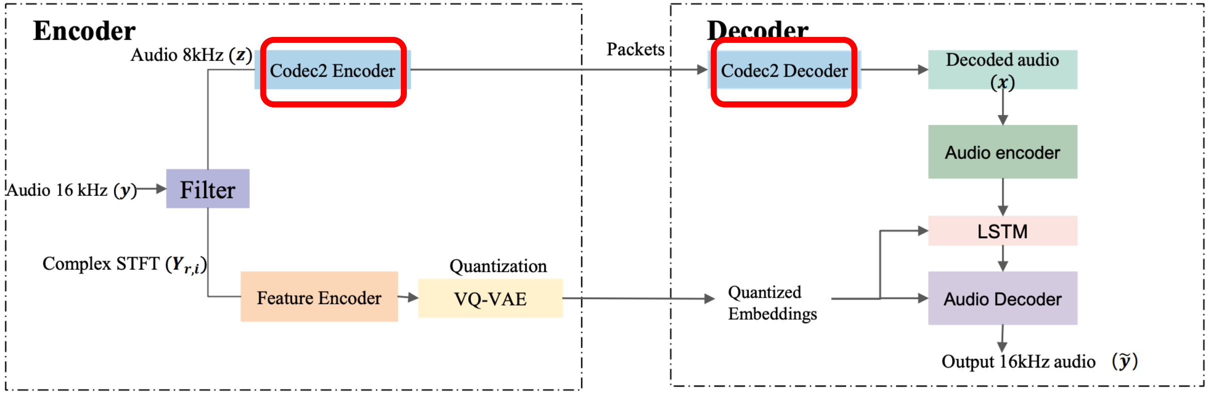


Fig. 1. The proposed neural extension framework to Codec2.

Experiments

2.2 Feature Encoder

- The feature encoder takes full-band (wideband?) complex STFT $Y_{r,i}$ as input.
  - STFT is extracted for each 10ms frame of incomin g audio for sync with Codec2 encoder.
- Authors use two designs for the feature encoder.
  - Split Frequency(SF): explicitly split the frequency bins into low and high frequency parts which are independently encoders.
    - Each encoder consists of five convolutional blocks.
  - Split Channel(SC): use a single encoder across the entire spectrum.
    - Used Six convolutional blocks.

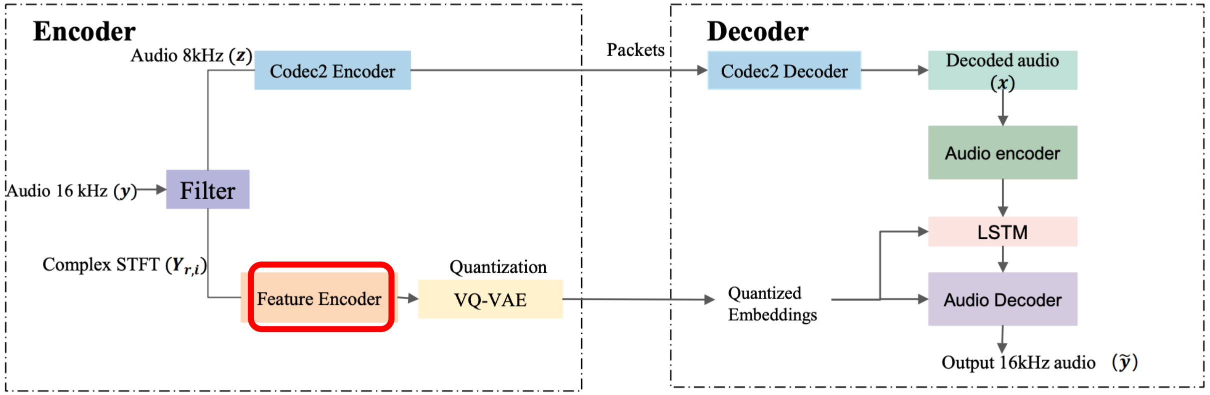
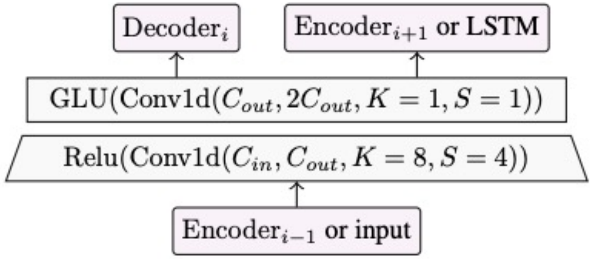


Fig. 1. The proposed neural extension framework to Codec2.



- Convolutional blocks are similar to the block used in [1], compos ed of a 2-D (1-D?) convolutional layer, followed by batch normaliza tion and GLU(gated linear units) as activation function.

[1] Alexandre Defossez, Nicolas Usunier, L ´ eon Bottou, and Fran- ´ cis Bach, "Music source separation in the waveform domain," arXiv preprint arXiv:1911.13254, 2019.

Experiments

2.3 Vector-Quantization Layer

□ Configurations

– Split Frequency(SF): employs different codebooks ac ross section of the spectrum.

– Split Channels(SC): employs different codebooks acr oss cluster of channels.

□ Codebook types:

– Two 9-bit codebooks, 1.8kbps

– Four 6-bit codebooks, 2.4kbps

□ VQ-VAE loss

$$L_{vq} = \left|\left|\mathrm{sg}[z_e] - \widetilde{z}_e\right|\right|_2^2 + \beta\left|\left|z_e - \mathrm{sg}[\widetilde{z}_e]\right|\right|_2^2,$$

z_e: output of feature encoder, $\widetilde{z}_e$: quantized embedding

(In this paper, $\beta = 0.25$)

□ The first term is optimized by an exponential movi ng average k-means.

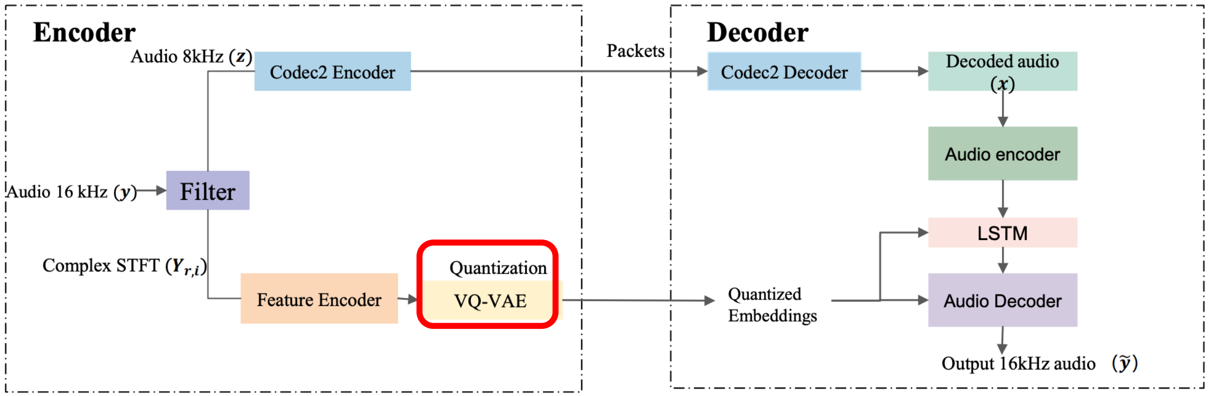


Fig. 1. The proposed neural extension framework to Codec2.

Experiments

2.4 Complex Convolutional Recurrent Network

- □ Audio-encoder
  - – Similar to feature-encoder.
  - – 5 Conv2d blocks.
    - ▪ Each with 128 channels, filters of size 2×6 and a stride s etting of 1×2.
- □ LSTM layer
  - – Two LSTM layer.
    - ▪ Each with 512 hidden nodes.
- □ Audio-decoder
  - – 5 Transposed Convolution 2d blocks
    - ▪ Same channel, filter and stride settings with audio-encoder except for the last block as two channel output.
  - – Skip connection.
  - – Convert the low-resolution features generated by the LSTM layers into high-resolution spectrograms.

$\tilde{z}_e$

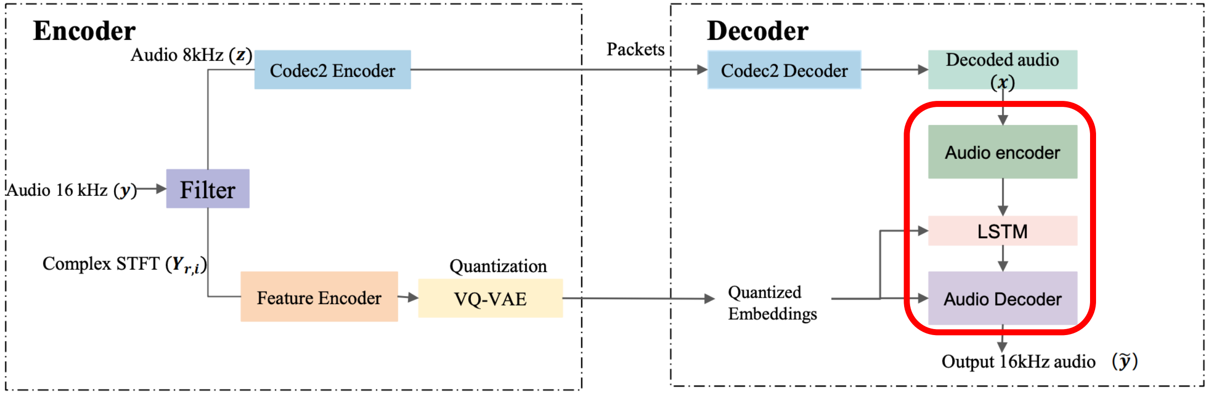
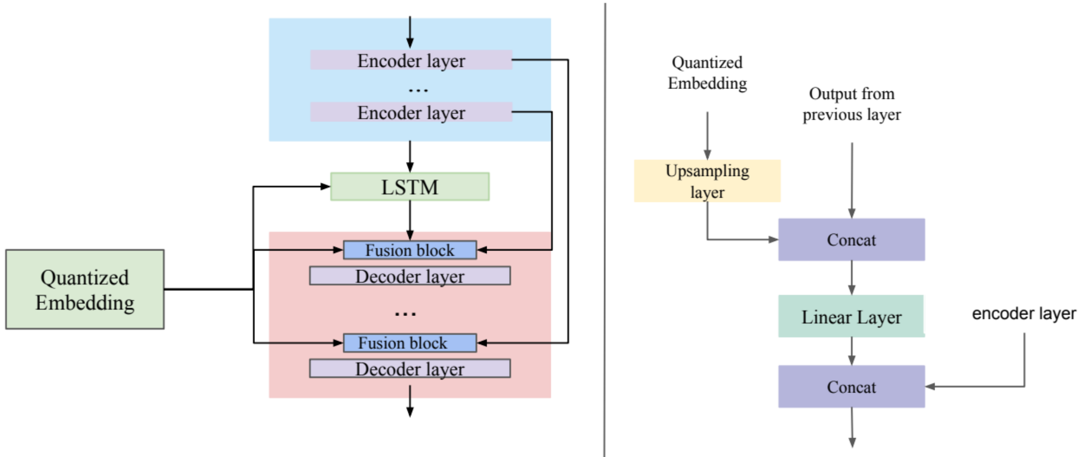


Fig. 1. The proposed neural extension framework to Codec2.



(a) Overall framework for CCRN (b) Details of the fusion block in audio decoder

Experiments

2.4 Complex Convolutional Recurrent Network

- □ The quantized embedding is fused in both the LSTM layers and audio-decoder layers.
  - – The quantized embedding is first passed to a upsampling layer with stride of 2 before concatenating the output with the result of previous decoder layer.
- □ This output of the fusion layer is combined with the output of correspond encoder layer after a linear projection to match the dimensions.

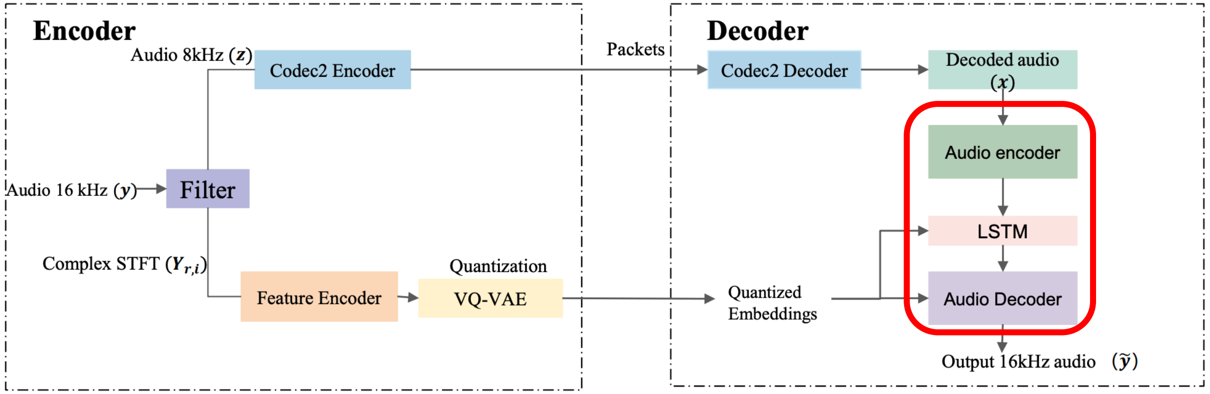
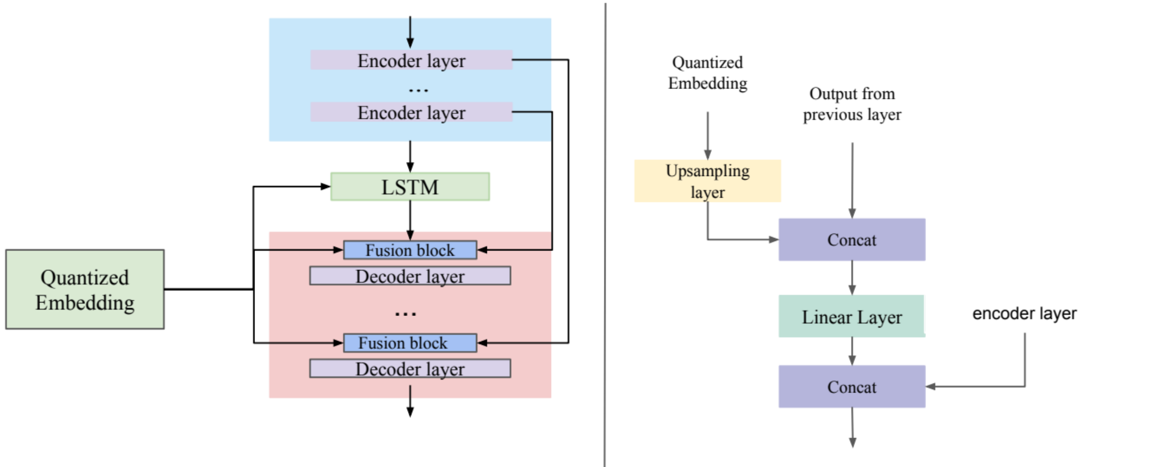


Fig. 1. The proposed neural extension framework to Codec2.



(a) Overall framework for CCRN

(b) Details of the fusion block in audio decoder

Experiments

2.5 Adversarial Training

□ In preliminary experiments, authors noticed that the reconstructed spectrograms had little variation in the high frequency band.

□ To address this problem, authors use LSGAN to fine-tune the models.

□ A generative adversarial network (GAN) consists of a generator network (G) and a discriminator netwo rk (D).

□ The two components are trained in an “adversarial” fashion: the discriminator tries to distinguish bet ween the samples produced by the generator from real samples, and the generator tries to fool the di scriminator by generating realistic samples.

Experiments

□ Generator

- Used the CCRN with VQ-VAE introduced above as the generator G.

□ Discriminator

- Discriminator D takes paired STFT magnitudes. ((target or enhanced) + decoded).
- Discriminator D consists of several Conv2d blocks (with ReLU activation) and two fully connected layers.
  - No activation function for the last FC layer.
- For discriminator D, authors investigate two configurations based on how real pair and fake pair data combined.
  - LSGAN-V1: channel-wise concatenation
  - LSGAN-V2: frequency-wise concatenation
    - only first 4kHz frequency bands of the upsampled decoded audio are used for LSGAN-V2.

Experiments

2.6 Loss Function

□ Generator loss (total loss)

– Consists of the reconstruction loss and the discriminator loss.

$$L_{total} = L_{recon} + L_{adv}$$

– The **reconstruction loss** includes an L_1 loss in the time domain, a weight STFT loss(WSTFT) and VQ-VAE loss mentioned before.

$$L_{recon} = \lambda_1 ||y - \tilde{y}||_1 + \lambda_2 L_{WSTFT}(Y, \tilde{Y}) + \lambda_3 L_{vq}$$

- x: upsampled decoded signal, y: original signal, $\tilde{y}$: enhanced signal
- $\lambda_1 = 1$, $\lambda_2 = 22$ (FFT scaling factor), $\lambda_3 = 1$

– WSTFT is proposed to emphasize the high frequency region.

- Author split the frequency bins into 4 sub-bands and each sub-band is assigned a specific weight.
- The hyperparameter w_k in equation below was set to (0.1, 1.0, 1.5, 1.5).

$$L_{WSTFT} = \sum_{k=1}^{4} w_k \, ||Y_k - \tilde{Y}_k||_1$$

– The **adversarial loss** for the generator is defined as:

$$L_{adv} = \frac{1}{2} \mathbb{E}_{(X_{r,i}, Y_{y,i}) \sim p_{data}(X_{r,i}, Y_{r,i})} \left[\left(D\left(G(X_{r,i}, Y_{r,i}), X\right) - 1 \right)^2 \right]$$

Experiments

2.6 Loss Function

□ Discriminator loss

– The discriminator network D seeks to distinguish real data from generated data by minimizing the following lo ss function:

$$L_D = \frac{1}{2}\mathbb{E}_{Y,X\sim p_{data}(Y,X)}[(D(Y,X)-1)^2] + \frac{1}{2}\mathbb{E}_{(X_{r,i},Y_{r,i})\sim p_{data}(X_{r,i},Y_{r,i})}\left[D\left(G\left(X_{r,i},Y_{r,i}\right),X\right)^2\right]$$

3. Experiments

Experiments

Experimental Setup

□ Dataset

- Training set: DNS Challenge dataset
  - 10-second segments, 204k in total
- Validation set: DNS development dataset
  - 10-second segments, 150 in total
- Test dataset
  - 15 sentences from Librispeech
  - 15 sentences from VCTK dataset
- Every data sample downsampled to 16kHz audio.

- STFT Filter in the encoder side
  - Hanning windows of 20ms
  - Hop size of 10ms
  - FFT length of 512 points
- Loss calculating STFT
  - Hanning windows of 32ms
  - Hop size of 16ms.

□ Evaluation metrics

- Mean Opinion Scores (MOS)
  - 30 sentences(test dataset)
  - 20 raters
  - MOS scores within 95% confidence intervals

Experiments

Experimental Setup

□ Baseline systems (codecs)

- Opus, NB, 6kbps
- Codec2, WB, 3.2kbps

□ (Additional) Model Configuration

- The discriminator in LSGAN-V1 consists of six Conv2d blocks with same filter sizes of 2*5 and output channels of [8,32,64,128,128,128].
  - Followed by two FC layers. (256->1)
- LSGAN –V2 uses '[discriminator in LSGAN-V1] + Conv2d blocks with output channel 64 and kernel size 1*1' to reduce feature dimensions.

□ Training hyperparameter

- Adam Optimizer
  - Initial learning rate with 0.0002
- Trained first 100 epochs using only reconstruction loss.
  - Batch size of 10.
- Fine-tuned above pretrained model with combination of reconstruction loss LSGAN loss for 30-60 epochs to avoid model collapse problem.
  - For computational efficiency, authors extract four seconds segment to compute the discriminator loss.

Experiments

Results

□ Comparison with baseline systems

– Compared Original Audio, Opus(sample rate: 12kHz), Codec2(sample rate: 8Hz) and proposed idea.

System	Codec2	VQ-VAE	Total bitrates	MOS
Original Audio (16 kHz)		-		4.10 ± 0.070
Opus	-		6kbps	3.38 ± 0.088
Codec2	-		3.2kbps	3.26 ± 0.098
Ours (LSGAN-V2)	1200	2400 (SC)	3.6kbps	3.58 ± 0.082
Ours (LSGAN-V2)	2400	2400 (SC)	4.8kbps	3.67 ± 0.083

Table 1. Performance in terms of MOS score for the proposed and baseline systems.

Experiments

Results

□ Impact of the LSGAN

– Authors' hypothesis is that upsampled audio only contains 4kHz speech spectrum due to Codec2 constraints a nd channel-wise concatenation (in LSGAN-V1) with paired data will leave the high frequency spectrum empty

ID	System	Codec2	VQ-VAE	Total bitrates	MOS
O1	Ours (w/o LSGAN)	2400	2400 (SC)	4.8kbps	3.44 ± 0.093
O2	Ours (LSGAN-V1)	2400	2400 (SC)	4.8kbps	3.53 ± 0.082
O3	Ours (LSGAN-V2)	2400	2400 (SC)	4.8kbps	3.67 ± 0.083

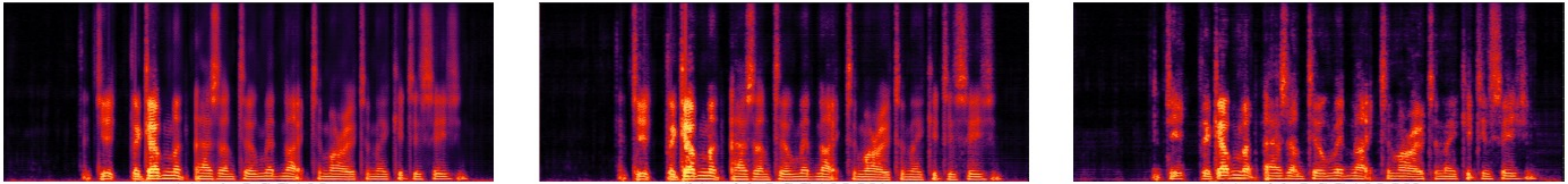
Table 2. Ablation studies for effectiveness of the LSGAN.

Experiments

Results

□ Impact of the LSGAN

- The output without LSGAN is considerably different than that of the ones with LSGAN fine-tuning.
- This difference can be observed in the high-frequency regions note in particular the **“over-smoothing”** effect t hat happens in the systems without adversarial training (a): the high-frequency part of the estimated spectrog ram exhibit patterns of “vertical bars” without much variation across frequency.



(a)w/o LSGAN (b) with LSGAN-V1 (c) with LSGAN-V2

Fig. 3. The spectrograms of estimated signal by the proposed approaches.

Experiments

Results

□ Bit allocation and its importance

– Comparing O4 to O5 and O5 to O6, authors could observe that allocating more bits to the embeddings is bette r than allocating more bits to the Codec2 parameters.

  - Besides, bit rate increase in the embeddings section will be accompanied with relatively larger compute increase than w hat would happen if we increased the Codec2 bitrate.

ID	System	Codec2	VQ-VAE	Total bitrates	MOS
O4	Ours (LSGAN-V2)	1200	2400 (SF)	3.6kbps	3.54 ± 0.082
O5	Ours (LSGAN-V2)	2400	1800 (SF)	4.2kbps	3.54 ± 0.084
O6	Ours (LSGAN-V2)	2400	2400 (SF)	4.8kbps	3.58 ± 0.083

□ Comparison of SC and SF (types of codebooks used in VQ-VAE)

– Authors observe that using SC can achieve better performance than using SF setting.

4. Conclusion

Conclusion

- □ In this work, authors have presented a hybrid speech codec that combines the traditional parametric c odec Codec2 and neural embeddings.
  - – This type of hybrid systems can be integrated with existing Codec2 systems with minimal integration effort.

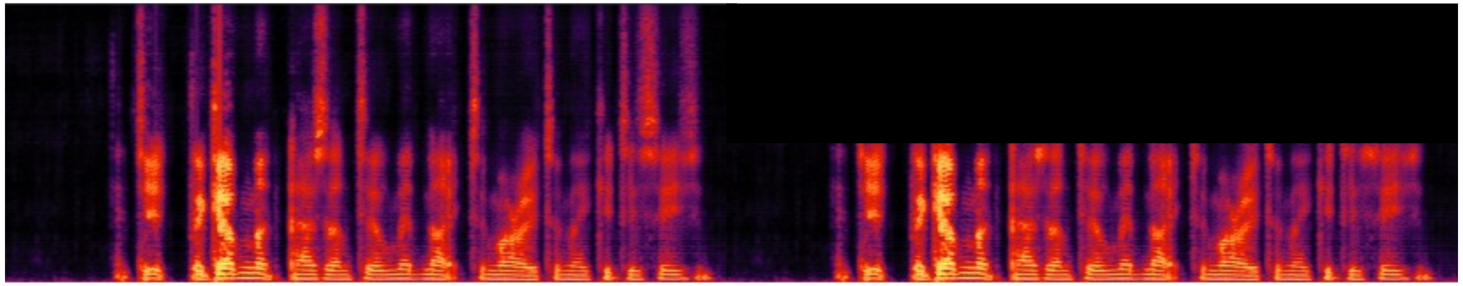
- □ In the future authors intend to explore architectures with better compute efficiency that do not sacrifi ce audio quality.

Thank you for listening
Q & A

Appendix

x: upsampled decoded signal, y: original signal, $\tilde{y}$: enhanced signal

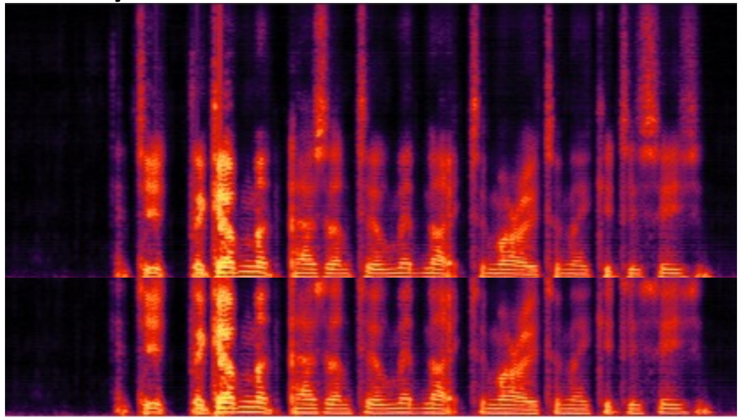
□ LSGAN-V1: channel-wise concatenation



y or $\tilde{y}$

x

□ LSGAN-V2: frequency-wise concatenation



y or $\tilde{y}$

x