VQVAE Unsupervised Unit Discovery and Multi-scale Code2Spec Inverter for Zerospeech Challenge 2019

Andros Tjandra[1,2], Berrak Sisman[3], Mingyang Zhang[3], Sakriani Sakti[1,2] ,
Haizhou Li[3], Satoshi Nakamura[1,2]

[1] Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan
[2] RIKEN, Center for Advanced Intelligence Project AIP (RIKEN AIP), Japan
[3] Department of Electrical and Computer Engineering, National University of Singapore (NUS), Singapore

Outline

- Background
- Previous Works
- Proposed Approach:
  → Vector Quantized Variational Autoencoder (VQVAE)
  → Codebook-to-Spectrogram Inverter (Code2Spec)
- Experiments
- Conclusion

Background

- **The ZeroSpeech 2019 challenge:**

  Confronts the problem of constructing a speech synthesizer without any text or phonetic labels: TTS without T

- **Two objectives:**

  → Discover subword units in an unsupervised way
  (Encodes them as efficient as possible -- low-bitrate)

  → Using the encoded representation, synthesize the speech
  to a different target speaker

Previous Works

- **Top performance in ZeroSpeech 2015 & 2017:**

  → Unsupervised clustering with DPGMM [Chen et al., 2015; Heck et al. 2017]

  → But, DPGMM is too sensitive to acoustic variations [Wu et al., 2018]

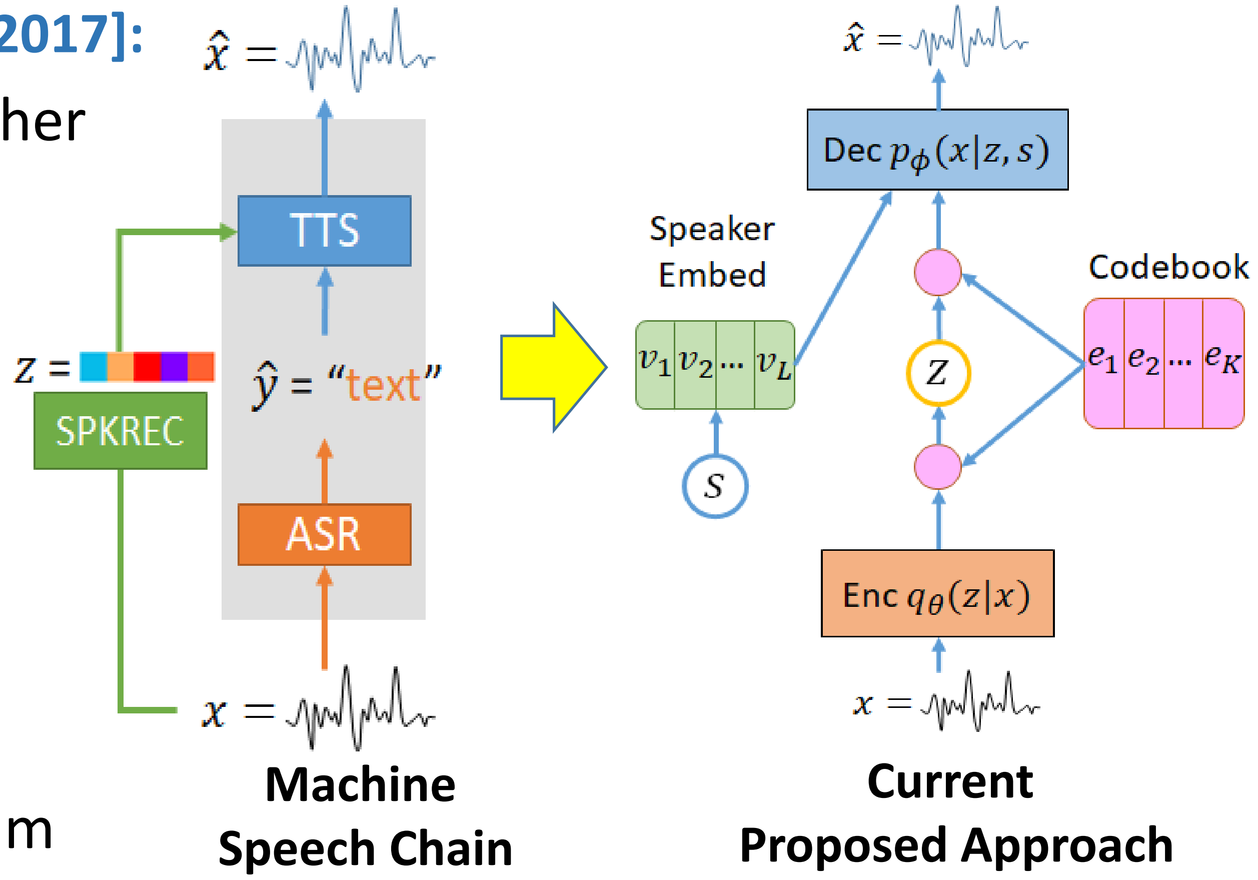
  → It is difficult to synthesize speech from DPGMM-based unit

Achieving the best trade-off between unit discovery and speech synthesize is necessary

Proposed Method

- **Machine Speech Chain [Tjandra et al., 2017]:**
  → Enables ASR and TTS to assist each other when they receive unpaired data
  → Optimize both models with reconstruction loss

- **Inspired by a similar idea, we propose:**
  → Frame-based vector quantized variational autoencoder (VQ-VAE)
  → Multi-scale codebook-to-spectrogram inverter (Code2Spec)



Vector Quantized Variational Autoencoder (VQVAE)

- **Three Components of VQ-VAE** **[van den Oord et al. 2017]**
  - **Encoder** $q_\theta(z|x)$

    Read speech features $x \in \mathbb{R}^D$ and output latent variable $z \in \{1..K\}$
  - **Codebook** $E = [e_1, .., e_K] \in \mathbb{R}^{K \times D_e}$

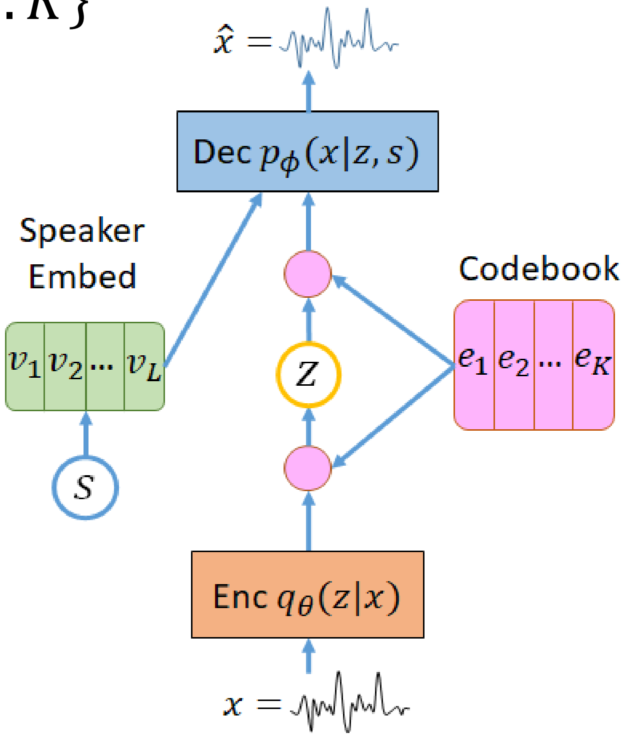
    Discretization is done by choosing the closest codebook

$$q_\theta(z=c|x) = \begin{cases} 1 & \text{if } c = \text{argmin}_i \|\hat{z} - e_i\|_2 \\ 0 & \text{else} \end{cases}$$

$$e_c = \sum_{i=1}^{K} q_\theta(z=i|x)\, e_i$$

  - **Decoder** $p_\phi(x|z,x)$ reconstruct the speech features conditioned by codebook z and speaker s

$$p_\phi(x|z,s) = p_\phi(x|e_c, v_s)$$

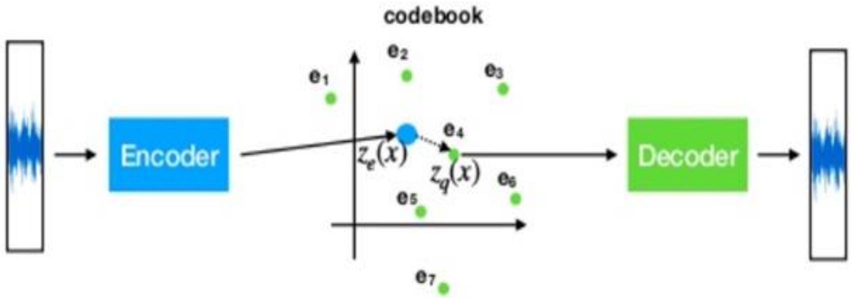


Vector Quantized Variational Autoencoder (VQVAE)

- **Training Objective:**

$$\mathcal{L}_{VQ} = \underbrace{-\log p_\phi(x|z,s)}_{\text{Reconstruction loss}} + \underbrace{\|\mathrm{sg}(\hat{z}) - e_c\|_2^2}_{\text{Embedding loss}} + \underbrace{\gamma\|\hat{z} - \mathrm{sg}(e_c)\|_2^2}_{\text{Consistency loss}}$$

1. **Reconstruction loss** between speech and generated speech
   $\rightarrow$ To optimize encoder and decoder parameter
2. **Embedding Loss** or the update loss of the codebook dictionary
   $\rightarrow$ To optimize the move of embedding to the encoder output
3. **Consistency Loss**
   $\rightarrow$ As the volume of the embedding space is dimensionless, this loss is to forces encoder to generate a representation near the codebook



[Source: https://www.slideshare.net/fukuabca/vqvae]

Codebook-to-Spectrogram (Code2Spec)

- **Code2Spec Inverter Module**

→ Generates magnitude linear spectrogram given the codebook using multiscale 1D convolution

$$\boxed{\hat{M} = \text{Code2Spec}([e[1], e[1], .., e[T_z], e[T_z]])}$$

→ Training Objective:

1. **MSE as reconstruction loss**

$$\mathcal{L}_{MSE} = \|M - \hat{M}\|_2^2$$

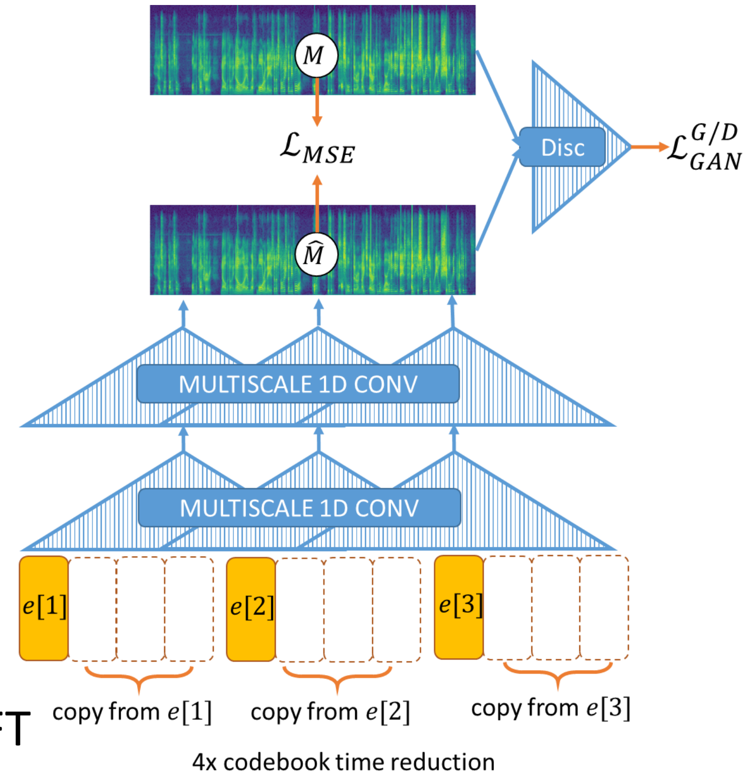
2. **GAN as auxiliary loss**

$$\mathcal{L}_{GAN}^{G} = \begin{cases} -\text{Disc}(\hat{M}) & \text{WGAN [Arjovsky et al., 2017]} \\ (\text{Disc}(\hat{M}) - 1)^2 & \text{LSGAN [Mao et al., 2017]} \end{cases}$$

$$\mathcal{L}_{GAN}^{D} = \begin{cases} \text{Disc}(\hat{M}) - \text{Disc}(M) & \text{WGAN} \\ \text{Disc}(\hat{M})^2 + (\text{Disc}(M) - 1)^2 & \text{LSGAN} \end{cases}$$

→ Waveform Generation:

Reconstruct the missing phase with Griffin-Lim algorithm & apply STFT



Experimental Set-up

- **Dataset:** Default ZeroSpeech 2019 Data on English & Surprise Languages
- **Feature Extraction:**
  → Mel-spectrogram (80 dimensions, 25-ms window size, 10-ms time-steps)
  → MFCC (13 dims + $\Delta + \Delta^2$)
- **Feature representations:**
  → Directly using features (no model involves)
  → K-Means
  → GMM with diagonal covariances
  → VQ-VAE
- **Stride size to reduce the time length:**
  → Stride size: 1, 2, 4, 8

Results: Baseline & Experiment on Direct Features

- **Baseline & Topline from ZeroSpeech**

Feature	ABX	Bit rate
Baseline	35.63	71.98
Topline	29.85	37.73

- **Direct feature representation (ABX with DTW cosine distance)**

Feature	ABX	Bit rate
Mel-Spec	30.291	1738.38
MFCC	21.114	1737.47

MFCC produced better performances on the ABX metric than the Melspectrogram.
But, the bit rate still remains too high.

Results: K-Means, GMM & Proposed VQ-VAE

- **K-Means continuous representation**

Model	ABX / Bitrate			
	#C	1T	2T	4T
K-Means (cont, DTW cos)	64	23.56 / 553	25.97 / 280	29.41 / 136
	128	23.16 / 649	24.24 / 321	28.12 / 161
	256	21.90 / 744	23.73 / 369	27.17 / 182

- **GMM posterior representation**

Model	ABX / Bit rate			
	#C	1T	2T	4T
GMM (post, DTW KL)	64	20.81 / 1647	22.67 / 676	29.82 / 257
	128	19.61 / 1705	23.06 / 704	31.19 / 281
	256	18.93 / 1691	23.39 / 757	32.99 / 306

- **Proposed Approach: VQ-VAE codebook representation**

Model	ABX / Bit rate				
	#CL	1T	2T	4T	8T
VQ-VAE (cont, DTW cos)	64	27.46 / 606	25.51 / 302	26.15 / 138	28.81 / 70
	128	27.65 / 686	24.29 / 347	25.04 / 165	30.87 / 79
	256	27.63 / 787	24.37 / 349	24.17 / 184	30.51 / 79
	512	27.69 / 871	23.59 / 400	24.63 / 180	32.02 / 74

VQ-VAE model has the best tradeoff between ABX and bit-rate compared to K-Means, GMM and direct MFCC features

Conclusion

- **VQ-VAE model has the best tradeoff between ABX and bit-rate compared to K-Means, GMM and direct MFCC features**
- **Things we tried but didn't work well:**

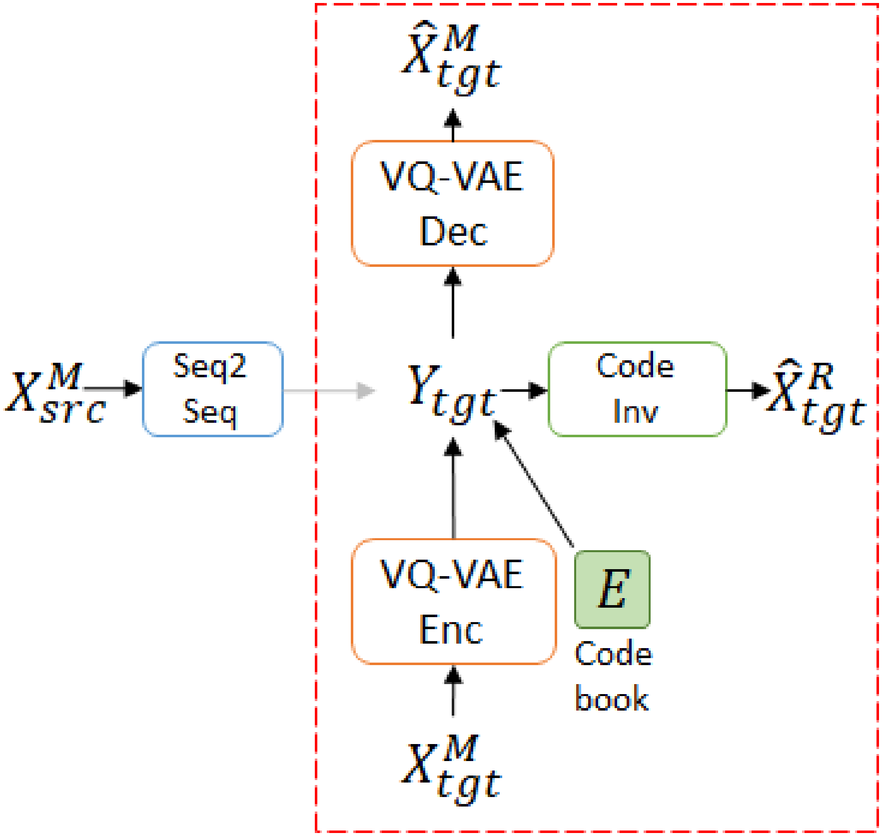
  → **Wavenet vocoder**
  The codebook every 20 or 40 ms perhaps too sparse

  → **GAN speech enhancement**
  Effective for achieving high-quality VC with clean speech, not for distorted speech
- **Our best submission:**

  → **VQ-VAE+Code2Spec with 256 codebooks and 2 & 4 time-stride**

  → Significantly improved performance from baseline (even the topline):
  the intelligibility - CER (**Rank 1st**), the naturalness - MOS (**Rank 3rd**),
  and the discrimination - ABX scores (**Rank 4th**)

Sequel: S2ST without T



VQ-VAE & Code2Spec

Andros Tjandra, Sakriani Sakti, Satoshi Nakamura, "SPEECH-TO-SPEECH TRANSLATION BETWEEN UNTRANSCRIBED UNKNOWN LANGUAGES," ASRU, pp. to appear, 2019

Thank you