ADL Final Project - ZeroSpeech2019 - TTS without T

Feng-guang Su
b04901070@ntu.edu.tw
National Taiwan University
Department of Electrical Engineering

Cheng-ping Hsieh
b04901020@ntu.edu.tw
National Taiwan University
Department of Electrical Engineering

Chung-ming Chien
b04901102@ntu.edu.tw
National Taiwan University
Department of Electrical Engineering

ABSTRACT

We compare two schema, the Multilabel-Binary Vectors (MBV) autoencoder [4] and the Vector Quantized Variational Autoencoder (VQVAE) [6], in which discrete representations of subword units could be discovered from speech without any parallel data, including text label, phoneme label and alignment. Both models apply autoencoding mechanism but with different revisions to extracting latent features from speech while ensuring discreteness. The efficiency of each encoding is evaluated according to the bitrate, while the quality can be checked by human by inspecting the generated spectrogram and listening the generated audios. Monophonic audio generation and interpolation of the MBV encoding is performed to further study the characteristic of the latent distribution. By combining the two methods, we aim to utilize their strengths and achieve a better performance in the ZeroSpeech 2019 Challenge, in terms of either bitrate or quality.

KEYWORDS

text to speech, unsupervised learning, subword unit discovery, voice conversion, adversarial learning, disentangled linguistic feature

ACM Reference Format:
Feng-guang Su, Cheng-ping Hsieh, and Chung-ming Chien. 2018. ADL Final Project - ZeroSpeech2019 - TTS without T. In *Proceedings of CSIE 5431.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/1122445.1122456

1 INTRODUCTION

Recently, text to speech (TTS) has emerged due to the observation of user-interface application and has gradually gained more and more popularity. The underlying training methods on this task usually require the large quantity of labeled training data, including text labels or phoneme labels. However, it is quiet challenging and costly to collect high-quality parallel corpora. For the low-resourced languages where the transcriptions are not available, such speech synthesizing systems must be trained in an unsupervised manner. The ZeroSpeech 2019 Challenge: TTS withous T, addresses this issue and demands an end-to-end voice conversion system with discrete latent representation of speech. Several approaches are proposed to achieve this task while trying to strike a balance between discretization and audio quality. The MBV method and the VQVAE share a similar autoencoder backbone while trying to discretize the continuous output of the encoder in a different manner, followed by a decoding process conditioned on speaker identity to achieve voice conversion. The differences of the two model reflect on the information density of the discrete tokens and the quality of generated audio segments. In this work we figure out and improve the performances of these approaches and explore the trade-offs.

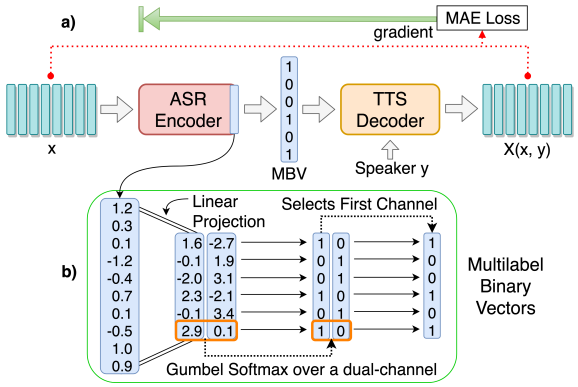

Figure 1: The ASR-TTS autoencoder framework where MBV are learned as discrete linguistic representations. [4]

2 RELATED WORK

We take the MBV approach [4] as our backbone. They present an ASR-TTS autoencoder framework with multilabel-binary vectors to learn distinct linguistic units discovery, as shown in Fig. 1. After then, they proposed to use additional adversarial training to generate a mask that augments the output of the TTS-Decoder for voice conversion. To learn multilabel-binary vectors, the ASR-Encoder is trained to map input acoustic feature sequence to a latent discrete encoding z, which is an n-dimensional binary vector consisting of arbitrary number of zeros and ones and defined as:

$$z = [e_1, e_2, ..., e_n] \in \mathbb{R}_n, e_i \in 0, 1 \tag{1}$$

In order to obtain the binarized differentiable vector z, they linear project a continuous output vector into a $\mathbb{R}^{nx2}$ space, where n is the dimension of the MBV. With dual-channel projection, they perform categorical reparameterization trick with Gumbel-Softmax [1] on the channel, which is equivalent to asking the model to determine whether an attribute is observed in a given input. The whole training procedure simply uses reconstruction mechanism with mean absolute error. Because the speaker identity is provided to the TTS-Decoder and the discreteness z possesses, the MBV encodings is able to learn an abstract space that is invariant to speaker identity and only encodes the content of speech, without using any form of linguistic supervision.

For the performance of MBV encoding, they can obtain a lower bit rate due to less dimensions of vectors and only ones and zeros used in the representation. However, there is a trade-off between bit rate and voice quality. According to their voice results, the voice with additional adversarial learning is more clear than only using autoencoder framework. However, we observe severe artifacts under both settings, shown below.

- Source voice from speaker A: https://bit.ly/2Xafl7w
- Generated voice to speaker B: https://bit.ly/2FHYRsK
- Generated voice to speaker B: https://bit.ly/2JgM0yD

In this study, we focus on dealing with the problems of voice quality, and thus pay more emphasis on the effectiveness and efficiency of finding distinct linguistic units instead of the voice conversion itself. Considering the trade-off of bitrate and voice quality, we aim to obtain a better result.

3 APPROACH

Before training the ASR-TTS autoencoder, each raw speech segment is transformed into a 2D spectrogram with short-time Fourier transform (STFT). The spectrogram could be viewed as a 2D image and handled with computer vision strategies. To improve the quality of the synthesized speech, we first modified the MBV method with adversarial learning, in which the generated audio can be improved by a discriminator. Besides, we also implement VQVAE architecture to obtain discrete linguistic units, which had been widely used for the discrete representation of the output of an autoencoder. The details of the methods we introduced are shown below.

3.1 Multilabel-Binary Vector (MBV) with Adversarial Learning

Due to the lack of constraints on MBV framework, we propose to add adversarial loss instead of only using reconstruction loss for the training of autoencoder. With a discriminator to distinguish whether the reconstructed spectrogram is real or fake, we can obtain a more human-like speech quality. The whole architecture is shown in Fig. 2, which is similar to VAEGAN [3].

3.2 Vector Quantized Variational Autoencoder (VQVAE)

The vector quantized variational autoencoder(VQVAE) [7], proposed in 2017, is a variant of variational autoencoder(VAE) [2]. VQVAE has been proved to be a powerful VAE-like framework with discrete latent representation, while preserving the ability to capturing abstract features. Various studies have shown competitive results against the original VAE framework [5, 7].

Here, we replace MBV architecture in [4] with the VQVAE framework [7] and propose 2 versions of model architectures. In the first version, we revise the training procedure of the VQVAE model, where a vanilla VAE is used for training and the vector quantization is conducted only for testing. That is, by directly feeding z into the decoder while training, we expect the model to learn a better representation and converge more efficiently in the early age of training. And by quantizing the encoded representation to the nearest embedding e_c in the *codebook* while testing, the latent space can be further discretized, enabling a much lower bitrate. Through the loss functions proposed in [7], the encoded latent feature z and the e_c are pulled toward each other, effectively resulting in a clustering of the encoded features, which indirectly discretizes the latent representation.

For comparison, in the second version, we directly use the training procedure of VQVAE, feeding the nearest embedding e_c into the decoder for both training and testing, and at the same time backpropagating the gradients to the encoder.

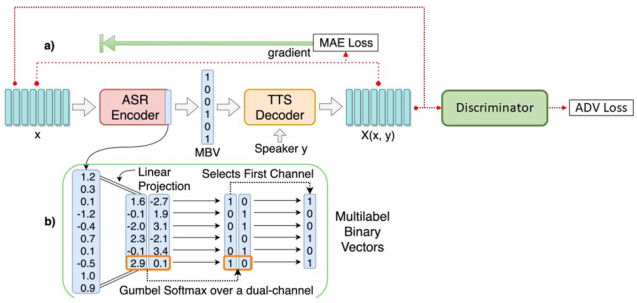

Figure 2: The MBV autoencoder framework with adversarial learning.

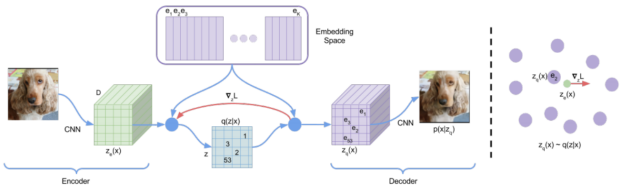

Figure 3: The VQVAE architecture which corresponds to our VQVAEv2 model.

The loss function of the VQVAE framework is composed of 3 parts, the reconstruction loss of the spectrogram, the codebook loss, and the commitment loss, as proposed in [7]:

$$Loss = \| x - Dec(Enc(x), y) \| + \| z - sg(e_c)) \| + \gamma \| sg(z) - e_c) \| \quad (2)$$

where x is the raw speech, y is the speaker identity, $Dec(\cdot)$ and $Enc(\cdot)$ stands for the decoder and encoder network with the discretization mechanism included, and the function sg$(\cdot)$ stops the gradients. Noticeably, $Enc(\cdot)$ equals to z in VQVAEv1, and equals to e_c in the VQVAEv2. We set the the size of *codebook* to be 64, the dimension of *codebook* to be 64 ,and γ to be 0.25 for stable training.

4 EXPERIMENTS

In this section, we conduct several experiments to compare different methods on distinct linguistic units discovery, e.g. MBV, MBV with adversarial learning (ADV), VQVAEv1, VQVAEv2 , and continuous latent representation.

4.1 Same-Speaker Reconstruction (Training)

Same-speaker reconstruction simply reduce the task to an discrete autoencoding setting, which is just the case in the training of the model. Due to the similar reconstruction mechanism among different autoencoders, we first simply compare their reconstruction loss, spectrogram, and voice during training. Considering the reconstruction loss in Fig. 4, we can find the continuous and VQVAEv1 cases have similar loss, lower than all other methods. The results show that continuous vectors can learn latent features effectively

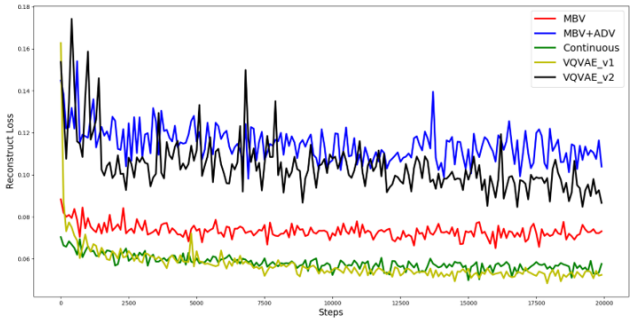

Figure 4: The reconstruction loss of different methods.

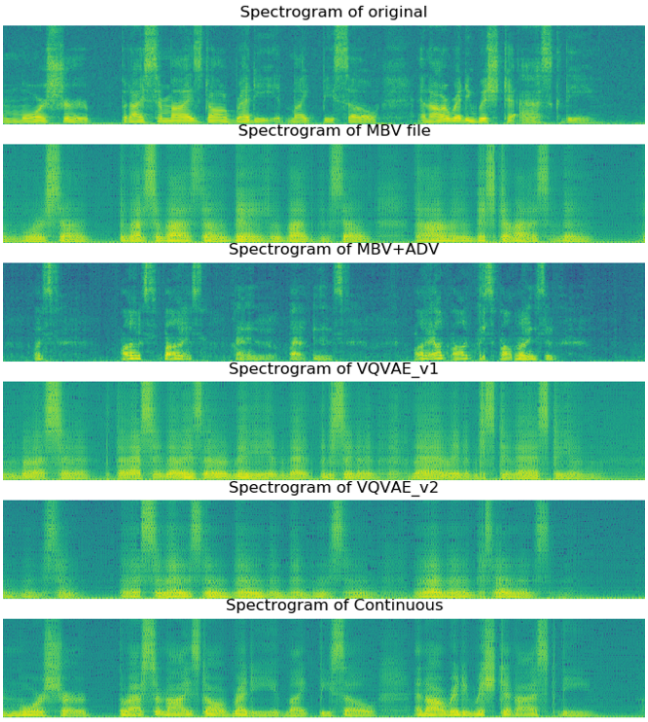

Figure 5: The spectrogram generated by different methods while training.

in the embedding space. In the other way, MBV with ADV have the highest reconstruction loss because a trade-off may arise between the adversarial loss and the reconstruction loss.

As shown in Fig 5, we can observe that MBV, VQVAEv1, VQVAEv2 have rather blurred spectrogram while MBV with ADV method has the clearest one. This result indicates that sharper waveform can be obtained with the aids of additional adversarial loss. However, this constraint suppress the reconstruction loss, resulting in the loss of some detailed information in the spectrogram.

The links to the voices generated by different methods are listed in Table 1. It is apparent that the voice reconstructed from continuous latent representation achieve the best quality while others still

Table 1: The same-speaker reconstructed voice by different methods.

Methods	Voice Link
Original	https://bit.ly/2NnaF9R
MBV	https://bit.ly/2ZVLlJx
MBV+ADV	https://bit.ly/2FGJJM8
VQVAEv1	https://bit.ly/31Zw0cN
VQVAEv2	https://bit.ly/2NobHCB
Continuous	https://bit.ly/2IWKcvD

Table 2: The generated speech of the voice conversion task via different methods.

Source	https://bit.ly/2XGKAGP	
Methods	**Target Man**	**Target Woman**
MBV	https://bit.ly/2XbDBBm	https://bit.ly/2KRmIKC
MBV+ADV	https://bit.ly/2FHdZGS	https://bit.ly/2XGbGhe
VQVAEv1	https://bit.ly/2Xj7xk4	https://bit.ly/2IZ6IUH
VQVAEv2	https://bit.ly/2X8J2kI	https://bit.ly/2YobgJx
Continuous	https://bit.ly/2XCajA3	https://bit.ly/2J093Pf

suffers from artifacts except for the MBV with ADV. With adversarial training, the synthesized speech sounds more human-like as well as the spectrogram does. Moreover, we can find that the VQVAEv1 model may generate a similar voice for different phonemes, which possibly means that our codebook does not learn complete representation for all phonetic combinations.

4.2 Voice Conversion

For the evaluation of voice conversion, we test the results with unseen source speaker and known target speaker, as shown in Table 2. The target man and woman speaker can be easily distinguished among all methods, which show the autoencoder has the ability to encode an unseen voice into a speaker-invariant latent space and decode into the voice of the target speaker. The linguistic and speaker similarity between the synthesized speech and the source speech are evaluated by our team members and simply ranked in Table 3.

We can find an obvious trade-off between content preservation (linguistic similarity) and style transfer (speaker similarity). However, the *continuous* latent representation obtains the best performance compared with the discrete counterparts. We are still working on finding a more effective discretization method which can disentangle speech content from style information and carry the required linguistic content efficiently, expecting to improve on both linguistic and speaker similarity.

4.3 Latent Space Interpolation

We also generate the monophonic speech segments by feeding multi-hot vectors into the decoder of our MVB model to explain the latent discrete representation. In table 4 we can clearly observe the interpolation on unique linguistic units. For example, the voice

Table 3: Trade-off

Methods	Bitrate	# Distinct Symbols	S.S. Rank	L.S. Rank
MBV	47.4	52	4	2
MBV+ADV	49	56	2	5
VQVAEv1	44.7	31	3	4
VQVAEv2	57.4	64	4	2
Continuous	138.5	16849	1	1

S.S.= Speaker similarity, L.S.= Linguistic similarity

Table 4: Latent Space Interpolation

MBV	Voice Link
10000	https://bit.ly/2XMkWkd
00001	https://bit.ly/2KPJW3C
10001	https://bit.ly/2KNoCvO

generated from vector [1,0,0,0,0,1] does sound like the linear combination of that from [1,0,0,0,0,0] and that from [0,0,0,0,0,1]. It is of importance that this interpolation experiment shows that the model definitely acquires the linear nature of voices.

4.4 Trade-off

Last, we measure the bitrate, the number of distinct symbols, and the human evaluation of the generated voices from all methods. From 3, we can observe that the quality of the generated voices is generally in inverse proportional to the bitrate and the number of distinct symbols. It is quite intuitive that the more distinct symbols are used, the clearer the generated voices are. However, we can easily find that our proposed model VQVAE-v1 generates better quality of voice but using fewer distinct symbols than the MBV method and VQVAE-v2. We suspect the reason for the better performance of VQVAE-v2 to be the flexibility of training procedure. By directly feeding z into the decoder while training, the encoder can learn more precisely and produce high-quality latent features, thus enhancing the ability of the codebooks to capturing the characteristic of the latent features by acquiring the average of similar latent vectors, after which the discretization can be viewed as a simple clustering task.

5 CONCLUSION

In this work, we study, explore, reproduce and compare several works on the TTS without T task. Without the aids of any parallel label, unsupervised autoencoding scheme is generally utilized, with various of discretization techniques applied to the latent space in order to find an efficient discrete representation of speech. In the experiments of the MVB methods, we address that adversarial training is useful to reduce artifacts of the generated speech but hurts the reconstruction quality and thus the linguistic information. We also show that the interpolation of the latent multi-hot labels produce meaningful results, indicating that in spite of the discreteness, the encoded vectors still possess the linear nature of voice. In the VQVAE experiments, we successfully improve the performance by force-feeding the continuous latent representation, which is not yet quantized, into the decoder network while training. In the eye of the voice conversion task, we observe a trade-off between the linguistic and speaker similarity, which is an inevitable result of such autoencoding approaches. In the future, we aim to develop a more clever autoencoding models specializing at such TTS without T task based on those we have studied, hope to surpass the performances of the existing ones.

REFERENCES

[1] Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144* (2016).
[2] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).
[3] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2015. Autoencoding beyond pixels using a learned similarity metric. *arXiv preprint arXiv:1512.09300* (2015).
[4] Andy T. Liu, Po chun Hsu, and Hung yi Lee. 2019. Unsupervised End-to-End Learning of Discrete Linguistic Units for Voice Conversion. arXiv:arXiv:1905.11563
[5] Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. 2018. Theory and Experiments on Vector Quantized Autoencoders. arXiv:arXiv:1805.11063
[6] Andros Tjandra, Berrak Sisman, Mingyang Zhang, Sakriani Sakti, Haizhou Li, and Satoshi Nakamura. 2019. VQVAE Unsupervised Unit Discovery and Multi-scale Code2Spec Inverter for Zerospeech Challenge 2019.
[7] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. 2017. Neural Discrete Representation Learning. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 6306–6315. http://papers.nips.cc/paper/7210-neural-discrete-representation-learning.pdf