論　文　要　旨
Thesis Abstract

(yyyy/mm/dd)　2022 年 09 月 09 日

※報告番号	乙第 97 号	氏　名 (Name)	Tanasan Srikotr

主論文題名（Title）

The Improved Speech Spectral Envelope Compression
Based on VQ-VAE with Adversarial Technique

内容の要旨（Abstract）

The speech vocoder or speech coding is designed to reduce the speech information in the sender before transmitting it via transmission media to the receiver and to reproduce the speech information. This kind of methodology saves the transmission system bandwidth and increases the number of users in the transmission system. However, we must carefully design the speech vocoder; otherwise, the quality of the reconstructed speech waveform deteriorates.

The famous speech coding is the Linear Predictive Coding (LPC-10) and Code-Excited Linear Prediction (CELP) based on speech analysis and speech synthesis system. Spectral envelopes are the critical speech parameter in speech processing. However, many methods based on Cepstrum and LPC cannot always synthesize natural-sounding speech. This dissertation extracts the high-quality spectral envelope from the WORLD vocoder to examine the speech quantization performance based on deep learning — the full spectral envelopes estimated from the WORLD vocoder can synthesize the high-fidelity speech waveform. However, the spectral envelopes are hard to quantify to obtain the quantized spectral envelopes acceptable to synthesize the natural, high-quality speech waveform.

The proper conventional compression technique required to quantize the spectral envelope parameters is Vector Quantization (VQ). Lately, deep learning technologies have shown an advantage compared to conventional VQ. The Vector Quantized Variational AutoEncoder (VQ-VAE) is an end-to-end compression technique based on the deep learning method. The VQ-VAE is the quantization version of the Variational AutoEncoder (VAE). The difference between a VAE and a VQ-VAE is that the VAE learns continuous z-latent representations, whereas the VQ-VAE learns discrete z-latent representations. The compression based on deep learning widely introduces the VQ-VAE because the VQ-VAE provides better performance than conventional VQ methods such as LBG or K-means.

This dissertation's whole study focuses on the advantage of deep learning in reducing the reconstruction errors of speech spectral envelope quantization compared to the conventional VQ and the VQ-VAE.

The first part of the study in this dissertation examined the effect of deep learning architecture on VQ based on deep learning. The conventional VQ and the VQ based on deep learning were compared for the spectral envelope quantization performance. The spectral envelope parameters were extracted from a high-quality vocoder named WORLD at 48 kHz sampling frequency in the experiments. The quantization performance in four target bitrate operations varied from low to high bitrates was evaluated. We proposed the Multi-layers Perceptron Vector Quantized Variation AutoEncoder (MLP-VQ-VAE). It reduced the memory sizes of z-latent representations and embedding space (codebook) by around 1.6 times compared to the conventional VQ and 21.4 times for the VQ-VAE. It also decreased the average Log Spectral Distortion (LSD) by around 1.1 points in dB lower than the conventional VQ and around 2.5 points in dB than the VQ-VAE.

The second study was about the techniques of VQ in VQ-VAE and investigated the possibility of improving the reconstruction performance. We proposed the Sub-band Vector Quantized Variational AutoEncoder (Sub-band VQ-VAE) and the Predictive Vector Quantized Variational AutoEncoder (Predictive VQ-VAE). The spectral envelope quantization performance of the WORLD vocoder at 48 kHz sampling frequency was compared. The experimental results for the four target bitrates showed that the Sub-band VQ-VAE reduced the average LSD by around 1.3 points in dB compared to the conventional VQ-VAE. The Predictive VQ-VAE results indicated that it had a lower distortion in terms of LSD than the VQ-VAE, about 2.58 points in dB for the four target bitrates.

The last study in the dissertation was the advanced deep learning training technique in VQ-VAE. The collaborative design of the VQ-VAE and the Generative Adversarial Network (GAN) worked together in the spectral envelope quantization of the WORLD vocoder, operated at 16 kHz sampling frequency. We proposed the three different methods in deep learning trainings with GAN architectures: the VAEGAN implemented in VQ-VAE, the VQ-VAE-EMGAN, and the VQ-VAE-EMDEC. They were compared with the VQ-VAE for the quantization performance in four target bitrate operations. The proposed training methods with GAN showed the effectiveness. The VQ-VAE-EMDEC reduced the average LSD by around 0.98 points in dB, the average L2 z-latent error by around 0.11 and in terms of reconstructed speech waveform, it also improved the Perceptual Evaluation Speech Quality (PESQ) by around 0.32 compared to the VQ-VAE.