CycleGAN-VC: Non-parallel Voice Conversion Using Cycle-Consistent Adversarial Networks

Takuhiro Kaneko, Hirokazu Kameoka

Abstract— We propose a non-parallel voice-conversion (VC) method that can learn a mapping from source to target speech without relying on parallel data. The proposed method is particularly noteworthy in that it is general purpose and high quality and works without any extra data, modules, or alignment procedure. Our method, called CycleGAN-VC, uses a cycle-consistent adversarial network (CycleGAN) with gated convolutional neural networks (CNNs) and an identity-mapping loss. A CycleGAN learns forward and inverse mappings simultaneously using adversarial and cycle-consistency losses. This makes it possible to find an optimal pseudo pair from non-parallel data. Furthermore, the adversarial loss can bring the converted speech close to the target one on the basis of indistinguishability without explicit density estimation. This allows to avoid over-smoothing caused by statistical averaging, which occurs in many conventional statistical model-based VC methods that represent data distribution explicitly. We configure a CycleGAN with gated CNNs and train it with an identity-mapping loss. This allows the mapping function to capture sequential and hierarchical structures while preserving linguistic information. We evaluated our method on a non-parallel VC task. An objective evaluation showed that the converted feature sequence was near natural in terms of global variance and modulation spectra, which are structural indicators highly correlated with subjective evaluation. A subjective evaluation showed that the quality of the converted speech was comparable to that obtained with a Gaussian mixture model-based parallel VC method even though CycleGAN-VC is trained under disadvantageous conditions (non-parallel and half the amount of data).