Amphibian Sounds Generating Network Based on Adversarial Learning

Sangwook Park, Mounya Elhilali, Senior Member, IEEE, David K. Han, Senior Member, IEEE, and Hanseok Ko, Senior Member, IEEE

Abstract—This letter proposes a generative network based on adversarial learning for synthesizing short-time audio streams and investigates the effectiveness of data augmentation for amphibian call sounds classification. Based on Fourier analysis, the generator is designed by a multi-layer perceptron composed of frequency basis learning layers and an output layer, and a discriminator is constructed by a convolutional neural network. Additionally, regularization on weights is introduced to train the networks with practical data that includes some disturbances. Synthetic audio streams are evaluated by quantitative comparison using inception score, and classification results are compared for real versus synthetic data. In conclusion, the proposed generative network is shown to produce realistic sounds and therefore useful for data augmentation.

Index Terms—Generative model, adversarial networks, Wasserstein distance, audio stream generation.

recordings have been explored by collecting sounds from various frog and toad species [3]–[6].

As data collection is an expensive and intricate effort, the idea of using generative networks to augment data has begun to attract the attention of researchers [7]–[9]. Among several approaches, Generative Adversarial Networks (GANs) composed of generator and discriminator have been remarkably effective in image generation [10]–[14]. The generator produces an image from a random seed, and the discriminator decides whether the image was synthesized by the generator or not. With help of regularization schemes and tuning [15]–[18], the training of GAN may converge, and the generator produces data indistinguishable by the discriminator.

Although GAN-based data generation has been effective for image-related tasks, using GAN networks for audio generation still poses many fundamental challenges. Generating real audio