2018 International Conference on Computational Science and Computational Intelligence (CSCI)

Voice Conversion using Conditional CycleGAN

Dongsuk Yook and In-Chul Yoo
Artificial Intelligence Laboratory
Korea University
Seoul, Korea
{yook, icyoo}@ai.korea.ac.kr

Seungho Yoo
Business School
Korea University
Seoul, Korea
seung1343@gmail.com

***Abstract*—Voice conversion (VC) modifies characteristics of speech, such as gender and speaker identities. The VC can be applied to various tasks including speaking assistance and speaker anonymization. Generally, such VC techniques require parallel speech data for training, which is very expensive. Recently, voice conversion has been accomplished using CycleGAN, which does not require parallel speech data. In this paper, we further extend the idea of using CycleGAN to convert multiple speakers' voices by conditioning the CycleGAN using speaker identity information.**

Keywords-voice conversion; generative adversarial networks (GAN), CycleGAN, Conditional CycleGAN (CC-GAN)

I. Introduction

Voice conversion (VC) is the task of converting input speech to output speech by altering various speech characteristics, including gender and speaker identities while preserving the linguistic information. Such tasks can be accomplished with machine learning techniques using parallel speech data. The parallel speech data consists of paired utterances that have the same linguistic contents (e.g., words), but different speakers' voice. The machine learning algorithms then map the input speaker's voice to the target speaker's voice. The collection of parallel data required in this process is very expensive process, however.

Recently, the cycle-consistent adversarial network (CycleGAN) [1], which is a variant of the generative adversarial network (GAN) [2], was proposed for unpaired image-to-image translation. A voice conversion algorithm utilizing CycleGAN has demonstrated successful voice conversion using non-parallel source and target speeches [3]. In this paper, we further extend the idea of CycleGAN based VC to convert multiple speakers' voices by conditioning the CycleGAN using speaker identity information.

II. Voice Conversion using CycleGAN

Generative adversarial network (GAN) [2] has shown superior performances in various tasks. It consists of two modules, a generator and a discriminator. Using the MNIST database, for example, a generator module is trained to generate random realistic handwritten digit images, while the discriminator is trained to distinguish the actual training data from the generated random data. A variant of GAN, called conditional generative adversarial network (cGAN) [4], utilizes additional information such as class labels for the training of the generator and discriminator. The CycleGAN [2][5][6], adds cycle-consistency loss term to the loss function of the GAN for image conversion using unpaired data.

The CycleGAN algorithm can be applied to voice conversion problems using non-parallel speech data [3][7][8]. Reference [8] uses fully connected feed forward neural network for the generator and discriminator, while [7] uses convolutional neural networks (CNN) for the generator and discriminator. The CycleGAN-VC [3], which shows promising VC performance, uses a gated convolutional neural network [9]. The gated CNN has shown state-of-the-art performance in language modeling and speech modeling tasks. The CycleGAN-VC [3] can be implemented as a DiscoGAN model [5].

III. Voice Conversion using Conditional CycleGAN

Conventional VC methods based on CycleGAN [3][7][8] transform the speech from a single source speaker to another single target speaker. In this paper, we extend the conventional CycleGAN based VC in order to convert the speech from multiple source speakers into multiple target speakers using only a single GAN model. We can do this by conditioning the CycleGAN with speaker identity information as shown in Figure 1. We call it a conditional CycleGAN (CC-GAN). At each layer of CC-GAN, speaker identity vector Y is appended to the layer's output vector to form the input to the next layer. The input vector is also augmented with the speaker identity vector. A generator is typically composed of down-sampling layers and up-sampling layers. In generator G_{ab}, for example, speaker a's identity vector Y_a is fed into the down-sampling layers, while speaker b's identity vector Y_b is fed into the up-sampling layers. The speaker vector Y_b is also fed into the discriminator D_b. Figure 1 shows a case of one-to-one mapping between speakers a and b. Since the generator and discriminator are conditioned on the speaker identity vector Y, however, the same model can be used for other speakers as well by changing the content of Y. In this way, we can build a general generator that takes source speech, a source speaker identity vector, and a target speaker identity vector, to produce target speaker's voice.

978-1-7281-1360-9/18/$31.00 ©2018 IEEE
DOI 10.1109/CSCI46756.2018.00290

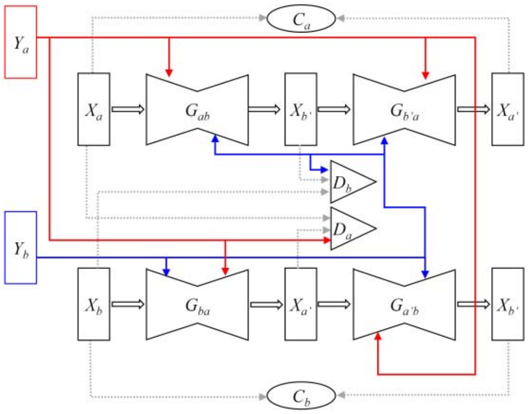

Fig. 1. Conditional CycleGAN architecture. *X*, *G*, *D*, *C*, and *Y* represent speech vector, generator, discriminator, cycle consistency loss computation, and speaker identity vector, respectively.

In fact, if there are *n* speakers, n^2 different mapping functions can be trained with a single CC-GAN. Since not only G_{ab} and $G_{a'b}$ (G_{ba} and $G_{b'a}$) but also G_{ab} and G_{ba} share the same parameters, there is only one physical GAN in CC-GAN that can compute n^2 different mapping functions. This is illustrated in Figure 2, where α and β are the speaker indices. The single GAN is trained with n^2 paired speaker data. The cycle consistency can be computed between the same speaker indices.

IV. Experiments

We used a CycleGAN-VC model [3] as our baseline. Input speeches were downsampled to 16kHz. Input features consisting of 24 Mel-cepstral coefficients (MCEPs), logarithm fundamental frequency (log F0), and aperiodicities (APs) were extracted every 5ms.

In the speech synthesis step, F0 parameters were obtained by logarithm Gaussian normalized transformation, and AP parameters were used unaltered. Source and target sequences were randomly segmented in sizes of 127 frames. We used female data (SF1) and male data (TM3) from obtained from an English speech dataset VCC2016 [10] for the preliminary experiments. Figure 3 shows the source, converted by CycleGAN, converted by CC-GAN, and target speech cepstral trajectories of an example utterance. It shows that CC-GAN can compute two different mapping functions successfully using only a single physical GAN.

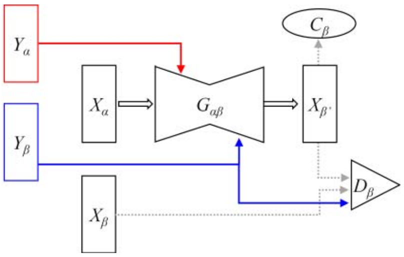

Fig. 2. Conditional CycleGAN implemented as a single physical GAN.

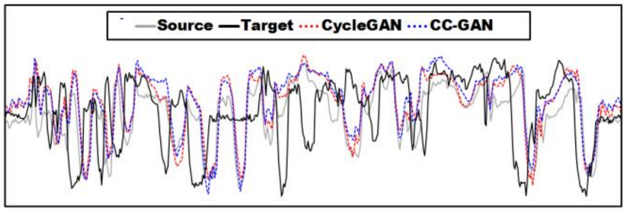

Fig. 3. Cepstral trajectories of an example utterance.

V. Conclusion

In this paper, we proposed a novel voice conversion method that can transform multiple speaker's voices using a single model called CC-GAN. The proposed method extends the conventional CycleGAN by conditioning it with speaker identity information. Preliminary experiments confirm the idea that use of CC-GAN is feasible for at least two speakers. Future studies are planned that include building a CC-GAN using hundreds of speakers in an effort to create a system that will convert any source speaker to any other target speakers.

Acknowledgment

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT and Future Planning (NRF-2017R1E1A1A01078157). Also, it was partly supported by the Institute for Information & Communications Technology Promotion (IITP) grant funded by the Korean government (MSIP) (No.2018-0-00269, A research on safe and convenient big data processing methods).

References

[1] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in Proc. ICCV, 2017, pp. 2223-2232.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in Proc. NPIS, 2014, pp. 2672-2680.

[3] T. Kaneko and H. Kameoka, "Parallel-data-free voice conversion using cycle-consistent adversarial networks," arXiv:1711.11293, 2017.

[4] M. Mirza and S. Osindero, "Conditional generative adversarial nets," arXiv:1411.1784, 2014.

[5] T. Kim, M. Cha, H. Kim, J. Lee, and J. Kim, "Learning to discover cross-domain relations with generative adversarial networks," in Proc. ICML, 2017, pp. 1857-1865.

[6] Z. Yi, H. Zhang, P. Tan, and M. Gong, "DualGAN: Unsupervised Dual learning for image-to-image translation," in Proc. ICCV, 2017, pp. 2868-2876.

[7] Y. Gao, R. Singh, and B. Raj, "Voice impersonation using generative adversarial networks," in Proc. ICASSP, 2018, pp. 2506-2510.

[8] F. Fang, J. Yamagishi, I. Echizen, and J. L. Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," in Proc. ICASSP, 2018, pp. 5279-5283.

[9] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in Proc. ICML, 2017, pp. 933-941.

[10] T. Toda, L. H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016," in Proc. Interspeech, 2016, pp. 1632-1636.