PROCEEDINGS of the
23rd International Congress on Acoustics
9 to 13 September 2019 in Aachen, Germany

Weighted generative adversarial network for many-to-many voice conversion

Dipjyoti Paul[(1)], Yannis Pantazis[(2)], Yannis Stylianou[(3)]

[(1)]University of Crete, Greece, dipjyotipaul@csd.uoc.gr
[(2)]Institute of Applied and Computational Mathematics FORTH, Greece, pantazis@iacm.forth.gr
[(3)]University of Crete, Greece, yannis@csd.uoc.gr

Abstract

The goal of voice conversion (VC) is to convert speech from a source speaker to that of a target, without changing phonetic contents. VC usually relies on parallel data for training, which limits its practical applications. Existing approaches are also limited in handling multiple speakers, since different models should be built independently for every speaker pair. To tackle that, a variant of Generative Adversarial Network (StarGAN-VC) were introduced that allows many-to-many mapping instead of learning all the pairwise transformations. Moreover, StarGAN-VC can handle non-parallel data, i.e., speakers do not need to utter the same sentences. In this paper, we suggest an algorithmic variation of StarGAN training where suitable weights are introduced. Weights which modify the Generator's gradient value aim to put more power to fake samples that fool the Discriminator. The suggested algorithm results in a stronger Generator. We refer to this variation as weighted-StarGAN (weStarGAN). In weStarGAN, the convergence of the training performance is accelerated. More importantly, the proposed algorithm achieves significant improvement against baseline StarGAN-VC concerning speech subjective quality for both speech quality and speaker similarity.

Keywords: Voice conversion, Generative adversarial networks, Training algorithm.

1 INTRODUCTION

Voice conversion (VC) modifies the para/non-linguistic information contained in the speech uttered by a source speaker, while keeping the linguistic contents unchanged. Statistical approaches like *Gaussian mixture model* (GMM) [1] and *joint density* GMM (JD-GMM) [2] were among the first successful attempts. Over the time, several non-linear spectral mapping techniques based on *restricted Boltzmann machine* (RBM) [3], *feed-forward deep neural networks* (DNNs) [4] and *recurrent* DNNs [5] were also proposed. Significant improvements have been witnessed following the introduction of *generative adversarial networks* (GANs). A variation of GANs named *cycle-consistent* GAN (CycleGAN) was presented in [6]. CycleGAN is designed to learn forward and inverse mappings simultaneously using both an *adversarial loss* and a *cycle-consistency loss*. One of the drawback of CycleGAN-VC is the ability to learns only one-to-one mappings. To resolve this issue, *Star generative adversarial network* based VC (StarGAN-VC) was recently introduced [7] which was originally proposed as a method for simultaneously learning images among multiple domains [8].

This paper extends the work of StarGAN-VC and proposes a novel training algorithm inspired by the *Weighted* GAN (WeGAN) [9]. Furthermore, to increase the stability in the existing StarGAN-VC, *Wasserstein GANs with gradient penalty* (WGAN-GP) is implemented [10]. The proposed *Weighted* StarGAN (WeStarGAN) approach introduces an effective weight factor for each sample in WGAN-GP. The new algorithm puts more weight to generated samples whose data distribution is closer to the real speech distribution. Such samples are more likely to fool the Discriminator. Simultaneously, it reduces the weights of generated speech samples that are confidently discriminated as fake speech. By doing so, WeStarGAN enhances the robustness of the weak Generator by adding weights to the training process and the Generator is able to improve source to target speech conversion by generating good quality target speech samples.

2 WEIGHTED STARGAN

In conventional StarGAN, the Generator G converts the attribute of source $\mathbf{x}$ speaker domain into target $\mathbf{y}$ speaker domain conditioned on the target domain label c, $\mathbf{y}' = G(\mathbf{x}, c)$. Here, $\mathbf{x}$ and $\mathbf{y}$ are the acoustic feature sequences of speech belonging to attribute domains source and target speakers, respectively. The Discriminator learns to distinguish between real and fake speech samples, while the Generator learns to generate fake speech that are indistinguishable from real speech samples. An auxiliary classifier is also introduced that allows the Discriminator to control multiple domains. Fig. 1 illustrates the training process of StarGAN-VC approach.

During the training of weStarGAN-VC using the stochastic grandient descend algorithm, a weight is multiplied to the Discriminator output provided with fake speech samples. Therefore, instead of equally-weighted 'fake' samples, weights are introduced in the Generator objective function while optimizing using two-player minimax game. The proper weights for WeStarGAN's Generator are defined by

$$w_i = e^{\eta min(0,\bar{D}_i)}$$

where η corresponds to the factor of the weight values empirically set to η=0.1. The value $\bar{D}_i$ is calculated from the output of the Discriminator $D(G(\mathbf{x_i}, c))$ after rescaling and mean value-subtraction. The weights are designed to impose more strength to samples that fool the Discriminator and thus are closer to the real data. Intuitively, the weighted algorithm puts more weight to fake target speaker's samples that are more probable to look alike source speaker's sample and simultaneously reduces the weight of speech samples that are confidently discriminated as fake.

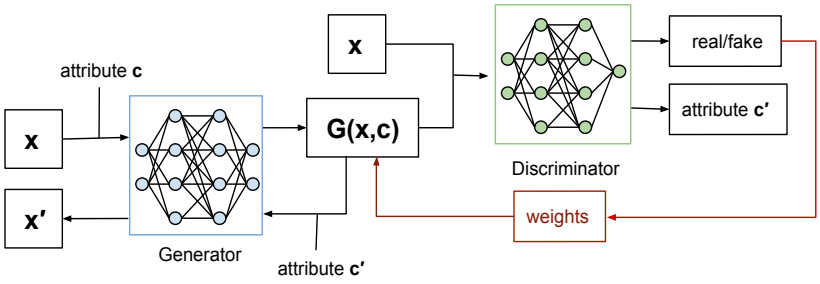

Figure 1. Overview of StarGAN-VC, consisting of two modules, a Discriminator D and a Generator G. The weights are only introduced during the training optimization process.

3 SUBJECTIVE EVALUATION

To assess the performance based on subjective evaluation experiments, we conducted listening tests for the naturalness or speech quality and speaker similarity of the converted speech to the target speech. Our proposed WeStarGAN-VC were compared against baseline StarGAN-VC architecture. Two separate listening tests are reported, 'ABX' and 'AB' test. In the 'ABX' test, experimental subjects have to decide whether a given sentence 'X' is closer in speaker similarity to one of a pair of sentences 'A' and 'B', which are converted speech samples obtained with the proposed and baseline methods, not necessarily in that order. Whereas, the 'AB' test compares the speech quality or naturalness of the converted speech. Fifteen native and non-native English listeners participated in our listening tests. For speaker similarity, the proposed method performs better with 65% preference whereas 17% preferences are given to 'No preference' option which indicates similar speaker characteristics in the speech samples generated using both the approaches. Moreover, WeStarGAN significantly outperforms baseline in generating good speech quality with 75% preference.

ACKNOWLEDGEMENTS

This project has received funding from the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu).

REFERENCES

[1] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol 6 (2), 1998, pp 131–142.

[2] A. Kain and M. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on*, vol 1, 1998, pp 285–288.

[3] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *IEICE TRANSACTIONS on Information and Systems*, vol 97 (6), 2014, pp 1403–1410.

[4] S. Desai, A. Black, B. Yegnanarayana, and K. Prahallad, "Spectral mapping using artificial neural networks for voice conversion," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol 18 (5), 2010, pp 954–964.

[5] L. Sun, S. Kang, K. Li, and H. Meng, "Voice conversion using deep bidirectional long short-term memory based recurrent neural networks," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp 4869–4873.

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

[7] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: Non-parallel many-to-many voice conversion with star generative adversarial networks," *arXiv preprint arXiv:1806.02169*, 2018.

[8] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo, "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 8789–8797.

[9] Y. Pantazis, D. Paul, M. Fasoulakis, and Y. Stylianou, "Training generative adversarial networks with weights," *arXiv preprint arXiv:1811.02598*, 2018.

[10] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, "Improved training of wasserstein gans," in *Advances in Neural Information Processing Systems*, 2017, pp. 5767–5777.