INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Non-parallel Voice Conversion using Weighted Generative Adversarial Networks

Dipjyoti Paul[1], *Yannis Pantazis*[2], *Yannis Stylianou*[1]

[1]Department of Computer Science, University of Crete, Greece
[2]Institute of Applied and Computational Mathematics, FORTH, Greece

dipjyotipaul@csd.uoc.gr, pantazis@iacm.forth.gr, yannis@csd.uoc.gr

Abstract

In this paper, we suggest a novel way to train Generative Adversarial Network (GAN) for the purpose of non-parallel, many-to-many voice conversion. The goal of voice conversion (VC) is to transform speech from a source speaker to that of a target speaker without changing the phonetic contents. Based on ideas from Game Theory, we suggest to multiply the gradient of the Generator with suitable weights. Weights are calculated so that they increase the power of fake samples that fool the Discriminator resulting in a stronger Generator. Motivated by a recently presented GAN based approach for VC, StarGAN-VC, we suggest a variation to StarGAN, referred to as Weighted StarGAN (WeStarGAN). The experiments are conducted on standard CMU ARCTIC database. WeStarGAN-VC approach achieves significantly better relative performance and is clearly preferred over recently proposed StarGAN-VC method in terms of speech subjective quality and speaker similarity with 75% and 65% preference scores, respectively.

Index Terms: Voice conversion, generative adversarial networks, training algorithm.

1. Introduction

The aim of voice conversion (VC) is to modify the para/non-linguistic information contained in the speech uttered by a source speaker, while keeping the linguistic contents unchanged. Various tasks such as personalized *Text-to-Speech* (TTS) systems, entertainment, speaking assistance and speech enhancements [1, 2, 3, 4] are benefited by the application of VC.

Voice conversion can be formulated as a regression problem of estimating a mapping function from source to target speech. A large number of popular statistical approaches like *linear multivariate regression* (LMR) [5], *Gaussian mixture model* (GMM) [6], *joint density* GMM (JD-GMM) [7] were introduced more than two decades ago which proved quite successful. Over the time, several non-linear spectral mapping techniques based on *restricted Boltzmann machine* (RBM) [8], *feed-forward deep neural networks* (DNNs) [9, 10], *recurrent* DNNs [11] and *non-negative matrix factorization* (NMF) [12] have also been proposed. However, most of these conventional VC methods require aligned parallel source and target speech data for training. In many scenarios, it is troublesome to collect parallel utterances. Even when parallel data is accessible, the required alignment procedures introduces artifacts and leads to speech-quality degradation. To overcome these limitations, numerous attempts have been made to develop non-parallel VC methods. *Sequence-to-sequence* (Seq2Seq) learning has proved to be outstanding at various research tasks and was successfully adopted in VC [13, 14, 15]. Seq2Seq VCs mainly uses multiple modules such as Automatic speech recognition (ASR) and TTS which are trainable with pairs of speech and its transcript rather than the source-target speech. These approaches converts both acoustic features and duration of the source speech. Nonetheless, these techniques consist of several training procedures and they are expensive in terms of both external data and computation.

Conditional variational autoencoders (CVAEs) approach were recently adopted for VC [16, 17]. CVAEs are an extended version of *variational autoencoders* where the encoder and decoder networks can take additional auxiliary input variable. The VC has experienced significant improvements following the introduction of *generative adversarial networks* (GANs). The VAE-GAN framework is an alternate approach for non-parallel VC that overcomes the weakness of VAEs [18]. Furthermore, a variation of GANs named *cycle-consistent* GAN (CycleGAN) was presented in [19]. CycleGAN utilizes a frame-by-frame approach which is designed to learn forward and inverse mappings simultaneously using an *adversarial loss* and *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 [20] which was originally proposed as a method for simultaneously learning images among multiple domains [21]. It possess a unified model architecture which allows simultaneous training of multiple domains i.e., many-to-many mapping within a single network.

Even though, a significant amount of research has been provided in the literature for non-parallel methods, generation of high quality audio quality is still very challenging and has room for improvement. This paper extends the work of StarGAN-VC and proposes a novel training algorithm inspired by *Weighted* GAN (WeGAN) [22]. Furthermore, existing StarGAN-VC utilizes three loss functions with conventional GAN approach. However, it lacks stable training which can be overcome by *Wasserstein GANs with gradient penalty* (WGAN-GP) [23]. In our proposed approach, we introduce a new and effective weight factor for WGAN-GP. The proposed *Weighted* StarGAN (WeStarGAN) algorithm improves the training of the Generator by transferring ideas from Game Theory. The new algorithm puts more weight to generated samples whose data distribution are more closer to the real samples and are more likely to fool the Discriminator. Simultaneously, it reduces the weights of generated samples that are confidently discriminated as fake. By doing so, WeStarGAN enhances the robustness of the weak Generator by adding weights to the training process and we expect that the inferred Generator is stronger favorably affecting the convergence properties. Experimental results based on subjective performance evaluation confirms that our proposed method achieves better speaker similarity and perceptual speech quality than baseline StarGAN-VC system.

Copyright © 2019 ISCA

http://dx.doi.org/10.21437/Interspeech.2019-2869

2. Generative Adversarial Networks

2.1. Preliminaries

Given the Discriminator D and Generator G, both parameterized via neural networks, $D(x)$ computes the probability of sample x being real while $G(z)$ is the sample produced by the Generator given noise input z. In order to be trained, the following objective function of the two-player *zero-sum game* has to be optimized:

$$\min_{G}\max_{D}\mathbb{E}_{x\sim p_{\text{data}}}[\log D(x)]+\mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))].$$

where p_{data} is the distribution to be learned, while p_z is the noise input distribution.

2.2. Star Generative Adversarial Networks

Our proposed algorithm is adopted from StarGAN approach [21] which was proposed for multi-domain image-to-image translation and slightly differs from StarGAN-VC approach [20] in terms of both cost functions and DNN architectures.

The objective is to train a single Generator G that learns mappings among multiple domains i.e., many-to-many speaker conversion. To achieve this, we train G to convert 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, $\mathbf{x} \in \mathbb{R}^{F\times D}$ and $\mathbf{y} \in \mathbb{R}^{F\times D}$ are the acoustic feature sequences of speech belonging to attribute domains $\mathbf{x}$ and $\mathbf{y}$. The target domain label c is generated randomly so that G can learn the flexibility to transform the source speech. An auxiliary classifier is introduced that allows the Discriminator to control multiple domains. Fig. 1 illustrates the training process of StarGAN-VC approach.

We applied three losses in the objective function, Adversarial Loss, Domain Classification Loss and Reconstruction Loss.

Adversarial Loss: G generates an fake data $G(\mathbf{x}, c)$ conditioned on both the source speaker's data $\mathbf{x}$ and the target domain label c, while D tries to distinguish between real and fake data. While training, G tries to minimize this objective, while the Discriminator D tries to maximize it. Moreover, we implemented Wasserstein GAN with gradient penalty [23] which uses a penalty term in the loss and provides strong performance and stability. The modified adversarial loss for D is defined as,

$$\begin{aligned}\mathcal{L}^{D}_{adv\text{-}gp} = E_{x\sim p_{\text{src}}}[-D(\mathbf{x})] + \mathbb{E}_{x\sim p_{\text{src}},c} D(G(\mathbf{x}, c))\\ + \lambda_{gp} E_{\hat{x}}[(||\nabla_{\hat{x}} D(\hat{\mathbf{x}})||_2 - 1)^2],\end{aligned}$$

$$\mathcal{L}^{G}_{adv} = -\mathbb{E}_{x\sim p_{\text{src}},c}[D(G(\mathbf{x}, c))],$$

where $\hat{\mathbf{x}}$ is sampled uniformly along a straight line between a pair of real and generated data samples and λ_{gp} is a constant value.

Domain Classification Loss: An auxiliary lassifier is implemented on top of D which imposes the domain classification loss while optimizing the cost function. Two loss terms are incorporated here: domain classification loss of real speech data which optimizes D and a domain classification loss of fake speech data which optimizes G. The losses are as follows,

$$\mathcal{L}^{real}_{cls} = \mathbb{E}_{x\sim p_{\text{src}},c'}[-\log D_{cls}(c'|\mathbf{x})],$$
$$\mathcal{L}^{fake}_{cls} = \mathbb{E}_{x\sim p_{\text{src}},c}[-\log D_{cls}(c|G(\mathbf{x}, c))],$$

where $D_{cls}(c'|\mathbf{x})$ represents a probability distribution of a real data $\mathbf{x}$ over domain labels computed by D. D learns to classify a real data to its corresponding original domain c'. Whereas, $D_{cls}(c|G(\mathbf{x}, c))$ represents the probability distribution of a fake data $G(\mathbf{x}, c)$ over domain labels computed by D. G tries to minimize this objective to generate data that can be classified as target domain c.

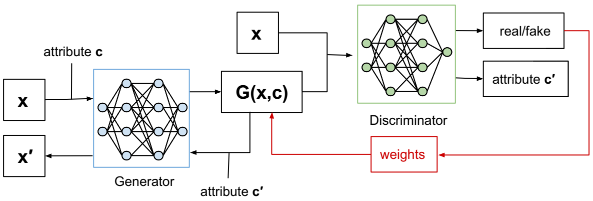

Figure 1: *Overview of StarGAN (in black), consisting of two modules, a Discriminator D (identical neural network architecture is used for Classifier except the last convolutional layer) and a Generator G. The weights (in red) are introduced during the training optimization process in our proposed algorithm.*

Reconstruction Loss: The adversarial and classification losses assist G to generate speech that are realistic and can be classified to its correct target domain. However, this does not guarantee on preserving the content of the linguistic information while changing only the speaker domain-related information. To alleviate this problem, a reconstruction loss is introduced to the Generator, defined as,

$$\mathcal{L}_{rec} = \mathbb{E}_{x\sim p_{\text{src}},c,c'}[||\mathbf{x} - G(G(\mathbf{x}, c), c')||_1],$$

where $G(\mathbf{x}, c)$ is the generated data conditioned on $\mathbf{x}$ and the target domain label c and $G(G(\mathbf{x}, c), c')$ is reconstruct the original speech $\mathbf{x}$ which is conditioned on $G(\mathbf{x}, c)$ and the original domain label c'. We applied $L1$ norm as a reconstruction loss.

The overall objective functions to be minimized with respect to G an D can be written as

$$\mathcal{L}_D = \mathcal{L}^{D}_{adv\text{-}gp} + \lambda_{cls}\mathcal{L}^{real}_{cls},$$

$$\mathcal{L}_G = \mathcal{L}^{G}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cls}\mathcal{L}^{fake}_{cls},$$

where λ_{rec} and λ_{cls} are the hyper-parameters for domain classification loss and reconstruction loss, respectively.

2.3. Training StarGAN with Weights (WeStarGAN)

In [22], authors presented a training algorithm based on weights that improved the performance of vanilla GANs. Instead of equally-weighted 'fake' samples, a weight to each sample is assigned which multiplies to the respective gradient term of the Generator. 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 samples that are more probable to fool the Discriminator and simultaneously reduces the weight of samples that are confidently discriminated as fake. A theoretical argument reveals that the optimal Generator with weights achieves lower or equal loss value than the optimal Generator with equally-weighted samples for a fixed Discriminator. Hence, it is expected the inferred Generator is stronger favourably affecting both the point and the speed of convergence with minor additional computational cost. The proposed algorithm is presented in Fig. 2.

We extend the training algorithm to Wasserstein GANs with gradient penalty (WGAN-GP). In WGAN-GP, the discriminator does not return the probability of a sample being real but a

```
Algorithm 1

for number of iterations do
    for k steps do
        Sample $\{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$ from the source data distribution $p_{src}(\mathbf{x})$.
        Update the Discriminator to minimize the objective function:
        $\frac{1}{m}\sum_{i=1}^{m}[-D(\mathbf{x_i}) + D(G(\mathbf{x_i}, c))]$
        $- \frac{1}{m}\sum_{i=1}^{m} \lambda_{cls}\log D_{cls}(c'|\mathbf{x_i})$
        $+ \frac{1}{m}\sum_{i=1}^{m} \lambda_{gp}(||\nabla_{\hat{x_i}} D(\hat{\mathbf{x}}_\mathbf{i})||_2 - 1)^2]$
    end
    Sample $\{\mathbf{x_1}, \ldots, \mathbf{x_m}\}$ from the source data distribution $p_{src}(\mathbf{x})$.
    Normalize:
    $\bar{D}_i = D(G(\mathbf{x_i}, c)) - \frac{1}{2m}[\sum_{j=1}^{m} D(\mathbf{x_j}) + D(G(\mathbf{x_j}, c))]$.
    Compute the unnormalized weights:
    $w_i = e^{\eta min(0, \bar{D}_i)}, \ i = 1, ..., m.$
    Normalize:
    $w_i = \frac{w_i}{\sum_{j=1}^{m} w_j}, \ i = 1, ..., m.$
    Update the Generator to minimize the objective function:
    $\sum_{i=1}^{m} -w_i D(G(\mathbf{x_i}, c)$
    $+ \frac{1}{m}\sum_{i=1}^{m} \lambda_{rec}||\mathbf{x_i} - G(G(\mathbf{x_i}, c), c')||_1$
    $- \frac{1}{m}\sum_{i=1}^{m} \lambda_{cls} \log D_{cls}(c|G(\mathbf{x_i}, c))]$
end
```

Figure 2: *Training algorithm of WeStarGAN. For a direct comparison with the StarGAN, we follow the formulation of [20].*

continuous regression-type value. Taking this fact into account, we uniformly scale the output of the Discriminator $D(G(\mathbf{x}, c))$ based on the output of Discriminator conditioned on both real and fake data and translate the data around axis 0. The normalized output is then employed to the weight function. The proper weights for WeStarGANs Generator are defined by

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

where η corresponds to the hyper-parameter which weighs the factor of the weight values. Note that, the normalized $\bar{D}_i$ is only used to estimate the weights. We empirically set $\eta = 0.1$ for our experiments.

The choice for the weights is dictated by the fact that we focuses on improving the Generator training by putting more attention on the data that are closer to real distribution. Therefore, when Discriminator output is $\bar{D}_i < 0$, the weight decreases by exponential factor. On the other hand, when $\bar{D}_i > 0$. our algorithm takes into account the samples which almost follow the real data distribution.

3. Experimental Setup

3.1. Experimental conditions

The experiments have been conducted with the CMU Arctic database [24] that consists of speech spoken by two male speakers (rms and bdl) and two female speakers (clb and slt) and are divided into two subsets i.e., training and evaluation, without overlap. As there are four speakers involves in our experiments, the attribute c is represented as a four-dimensional one-hot vector depending upon the target speaker attribute. Although, the database contains parallel speech, we randomly select training data as our system operates on non-parallel data, The sampling rate of the speech signals is 16 kHz. For each utterance, 36 dimension mel-cepstral coefficients (MCCs), logarithmic fundamental frequency ($logF0$), and aperiodicities (APs) were extracted for every 5 ms using the WORLD analyzer [25]. The $logF0$ is converted using the logarithm normalized transformation and the aperiodicities are used directly without any modification. Once the training process is completed, we use WORLD vocoder to generate speech from converted features.

3.2. Network architectures

In the Generator, an acoustic feature sequence is inserted and the outputs is an acoustic feature sequence of the same length. We normalize the source and target MCCs per dimension. The generator network comprises of three convolutional layers (conv), six residual blocks and three transposed convolutional layers (Dconv), and seven conv layers is used for Discriminator. Whereas, in [20], five conv layer and five Dconv layers are considered in Generator and two separate five conv layers are used for Discriminator and Classifier networks. Instance normalization [29] is used for the generator but no normalization is used for the discriminator. All models are trained using Adam optimizer with $\beta1 = 0.5$ and $\beta2 = 0.999$. The batch size is set to 32. The overview of network architecture is depicted in Fig 3.

4. Results and Discussion

In this section, we present the experimental results to evaluate the performance of voice converted speech samples. To assess the performance based on subjective evaluation experiments, we conducted listening tests for the speech quality (i.e., naturalness) and speaker similarity of the converted speech to the target speech. Our proposed WeStarGAN-VC were compared against recently proposed 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 vocal quality 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. All the converted speech samples were presented randomly from the evaluation set. Furthermore, the evaluation samples contains both intra-gender pairs and cross-gender pairs.

The evaluation results of the preference test are demonstrated in Fig. 4. The proposed WeStarGAN algorithm obtained the majority of preferences for best conversion in terms of sound quality and speaker similarity. For speaker similarity, the result shows that 17% preferences were given to 'No preference' option which indicates similar speaker characteristics in the speech samples generated using both the approaches. Nevertheless, the proposed method performs better with 65% preference. Moreover, WeStarGAN significantly outperforms baseline in generating good speech quality. The significant improvement in speech quality might be attributed to the fact that weights are only multiplied to the fake samples of the adversarial loss function which is responsible for generating real-like speech samples. On the contrary, no weights are introduced to the domain classification loss that is responsible for speaker mapping. We finally remark that WeStarGAN has the potential to be used for the training of lighter Generators which are

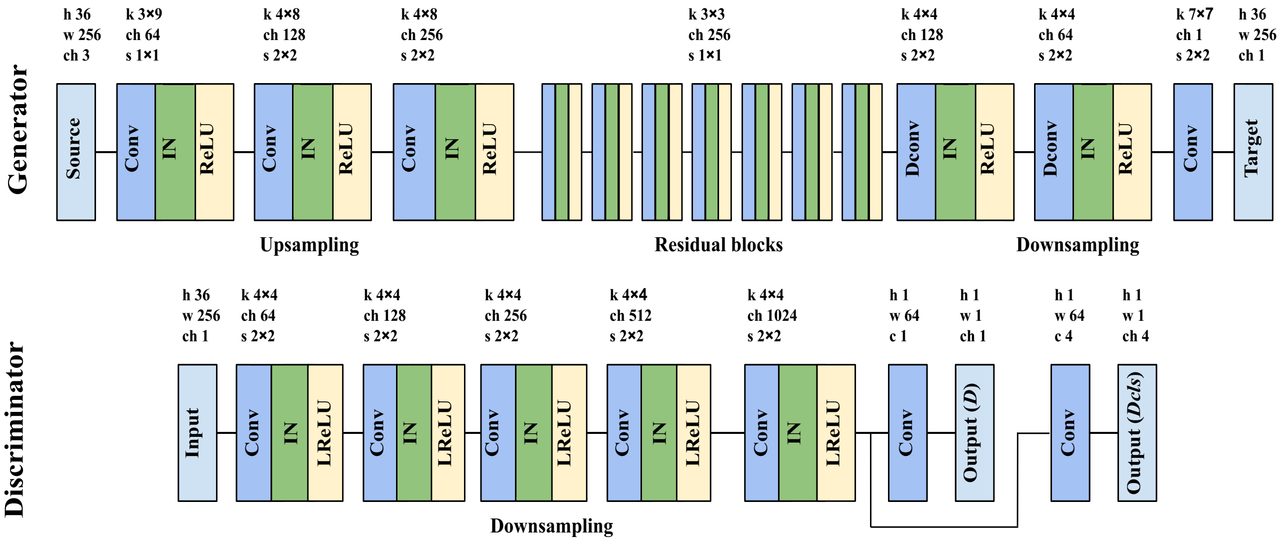

Figure 3: *Overview of StarGAN, consisting of two modules, a discriminator D and a generator G. In the input and output layers, h, w, and ch represent height, width, and number of channels, respectively. In each convolutional layer, k, c, and s denote kernel size, number of output channels and stride size, respectively. "Conv", "IN", "ReLU", "LReLU", "Deconv" denote convolution, instance normalization, rectified linear unit, leaky rectified linear unit and transposed convolution respectively. D_{cls} provides a probability distribution over domain labels where domain corresponds to the number of speakers used to train VC.*

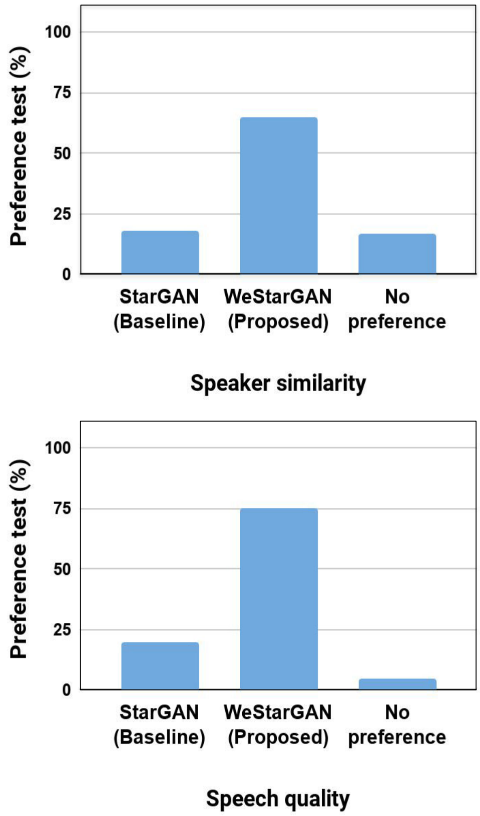

Figure 4: *Subjective preference test in (%) for speaker similarity and speech quality.*

necessary in cases such as operating on mobile devices. Indeed, constraints in computational power which affects the battery consumption as well as capability to respond in real-time limits the use of very deep and complicated neural networks. However, the use of lighter neural networks results in a less expressive and flexible Generator which may affect the quality of the converted speech. The application of the weighted algorithm aims to alleviate such issue by enhancing the capacity of the Generator.

5. Conclusion

In this paper, we proposed WeStarGAN, a novel algorithmic variation of StarGAN capable of performing non-parallel multi-domain voice conversion task. With minor additional computational cost, the suggested approach managed to improve the training process by devising a stronger generator at each mini-batch iteration. This development is very crucial because our approach can overcome the limitation of using a weaker generator and still can successfully be trained to generate good quality speech samples. In addition, we extended the weighting approach to the more stable WGAN-GP model. Subjective evaluation revealed that the proposed method obtained higher sound quality and speaker similarity than the baseline method. As future directions, we list a more extensive study in terms of network architectures and investigation of adding weights to the discriminator, too.

6. Acknowledgements

The first author has received funding from the EUs H2020 research and innovation programme under the MSCA GA 67532 (the ENRICH network: www.enrich-etn.eu). The second author acknowledges support of this work by the project Innovative Actions in Environmental Research and Development (PErAn) (MIS 5002358) which is implemented under the “Action for the Strategic Development on the Research and Technological Sector”, funded by the Operational Programme ”Competitiveness, Entrepreneurship and Innovation” (NSRF 2014-2020).

7. References

[1] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 1. IEEE, 1998, pp. 285–288.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech Communication*, vol. 54, no. 1, pp. 134–146, 2012.

[3] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[4] D. E. Eslava and A. M. Bilbao, "Intra-lingual and cross-lingual voice conversion using harmonic plus stochastic models," *Barcelona, Spain: PhD Thesis, Universitat Politechnica de Catalunya*, 2008.

[5] H. Valbret, E. Moulines, and J. Tubach, "Voice transformation using PSOLA technique," in *ICASSP*, vol. 1. IEEE, 1992, pp. 145–148.

[6] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *Speech and Audio Processing, IEEE Transactions on*, vol. 6, no. 2, pp. 131–142, 1998.

[7] 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. IEEE, 1998, pp. 285–288.

[8] T. Nakashika, T. Takiguchi, and Y. Ariki, "Voice conversion based on speaker-dependent restricted boltzmann machines," *IEICE TRANSACTIONS on Information and Systems*, vol. 97, no. 6, pp. 1403–1410, 2014.

[9] 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, no. 5, pp. 954–964, 2010.

[10] L. Chen, Z. Ling, L. Liu, and L. Dai, "Voice conversion using deep neural networks with layer-wise generative training," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 22, no. 12, pp. 1859–1872, 2014.

[11] 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)*. IEEE, 2015, pp. 4869–4873.

[12] Z. Wu, T. Virtanen, E. S. Chng, and H. Li, "Exemplar-based sparse representation with residual compensation for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 10, pp. 1506–1521, 2014.

[13] H. Miyoshi, Y. Saito, S. Takamichi, and H. Saruwatari, "Voice conversion using sequence-to-sequence learning of context posterior probabilities," *arXiv preprint arXiv:1704.02360*, 2017.

[14] J. Zhang, Z. Ling, L.-J. Liu, Y. Jiang, and L.-R. Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2019.

[15] K. Tanaka, H. Kameoka, T. Kaneko, and N. Hojo, "Atts2s-vc: Sequence-to-sequence voice conversion with attention and context preservation mechanisms," *arXiv preprint arXiv:1811.04076*, 2018.

[16] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from non-parallel corpora using variational auto-encoder," in *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 2016, pp. 1–6.

[17] Y. Saito, Y. Ijima, K. Nishida, and S. Takamichi, "Non-parallel voice conversion using variational autoencoders conditioned by phonetic posteriorgrams and d-vectors," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5274–5278.

[18] C. C. Hsu, H. T. Hwang, Y. C. Wu, Y. Tsao, and H. M. Wang, "Voice conversion from unaligned corpora using variational autoencoding wasserstein generative adversarial networks," *arXiv preprint arXiv:1704.00849*, 2017.

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

[20] 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.

[21] 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.

[22] Y. Pantazis, D. Paul, M. Fasoulakis, and Y. Stylianou, "Training generative adversarial networks with weights," *in European Signal Processing Conference, EUSIPCO*, 2019 (accepted).

[23] 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.

[24] J. Kominek and A. W. Black, "The cmu arctic speech databases," in *Fifth ISCA workshop on speech synthesis*, 2004.

[25] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.