*INTERSPEECH 2020*
October 25–29, 2020, Shanghai, China

# Nonparallel Emotional Speech Conversion Using VAE-GAN

*Yuexin Cao, Zhengchen Liu, Minchuan Chen, Jun Ma, Shaojun Wang, Jing Xiao*

Ping An Technology
{LIUZHENGCHEN871, CHENMINCHUAN109}@pingan.com.cn

## Abstract

This paper proposes a nonparallel emotional speech conversion (ESC) method based on Variational AutoEncoder-Generative Adversarial Network (VAE-GAN). Emotional speech conversion aims at transforming speech from one source emotion to that of a target emotion without changing the speaker's identity and linguistic content. In this work, an encoder is trained to elicit the content-related representations from acoustic features. Emotion-related representations are extracted in a supervised manner. Then the transformation between emotion-related representations from different domains is learned using an improved cycle-consistent Generative Adversarial Network (CycleGAN). Finally, emotion conversion is performed by eliciting and recombining the content-related representations of the source speech and the emotion-related representations of the target emotion. Subjective evaluation experiments are conducted and the results show that the proposed method outperforms the baseline in terms of voice quality and emotion conversion ability.

**Index Terms**: Emotional Speech Conversion, Variational AutoEncoder, Generative Adversarial Network, Supervised Learning, Style Transfer

## 1. Introduction

Human speech conveys more information, such as speaker identity and speaking style, than just linguistic content. Voice conversion (VC) is a technique which changes speaker identity of the speech while preserving its linguistic content [1]. Some other techniques have been studied to modify the speaking style, such as speaker's emotion [2] or accent [3]. In this paper, we focus on emotional speech conversion (ESC), which aims at transforming speaker's emotion of a speech utterance while keeping speaker identity and linguistic content unchanged. ESC can be applied to many fields, such as generating more natural and expressive speech, improving user experience in human-computer interactions, or hiding negative emotions in social occasions.

Various methods have been proposed for ESC, including rule-based approaches [2], Gaussian Mixture Model (GMM)-based approaches [4] and neural network-based approaches [5]. While these methods have demonstrated their effectiveness, they require accurately-aligned parallel data. Collecting parallel data and aligning the source and target utterances can be costly and time-consuming. For VC task, many methods have been studied to sidestep these issues [6, 7, 8, 9, 10, 11]. Methods based on Variational AutoEncoder (VAE) and its variants were proposed to disentangle and model latent representations for speech [6, 7, 8, 9]. In these methods, the conversion process was decomposed into encoding and decoding stages, and aligned frame pairs or even parallel corpora were no longer necessary . In [10], adversarial training was adopted to the VAE-based framework to make the generated spectra more realistic. An improved version of cycle-consistent Generative Adversarial Network (CycleGAN) was proposed in [11], making the method free of parallel data.

Although methods using nonparallel data for VC have been widely studied, corresponding researches for ESC are still inadequate. A pioneering work was done in [12], which achieved nonparallel training using an unsupervised style transfer technique. A variant of GAN was adopted to improve the quality of the converted speech. That work is inspiring, but there still remains a gap between the converted speech and the real target in terms of quality and emotion fidelity, partly because of the architecture of the GAN and the unsupervised way in the extraction of the style code. Recently, a framework using multitask learning with text-to-speech (TTS) was proposed [13], focusing mainly on the preservation of the linguistic content.

The contribution of this paper is two-fold. Firstly, an improved CycleGAN (noted as CycleGAN2 below for briefness) proposed for VC is adopted in VAE-GAN framework for ESC. Secondly, a supervised strategy to extract more reliable emotion-related representations is proposed. Subjective evaluations are conducted on IEMOCAP database [14] and the experimental results show that the proposed method outperforms the baseline in terms of voice quality and emotion conversion ability of the converted speech.

The rest of this paper is organized as follows. Section 2 provides a brief overview of the models used in our work. Section 3 introduces our proposed methods in detail. Experimental results and conclusions are given in Section 4 and 5, respectively.

## 2. Related work

### 2.1. Variational AutoEncoder (VAE)

VAE can be viewed as a variant over vanilla AutoEncoder, which has a more understandable and controllable latent space. The training process of a VAE, after introducing some necessary simplification, approximation and the re-parameterization trick, is equivalent to finding the optimal parameters that maximize the variational lower bound:

$$\{\theta^*, \phi^*\} = \arg\max_{\theta,\phi}\{-D_{KL}(q_\phi(\hat{z}|x)||p(z)) + \log p_\theta(x|\hat{z}, y)\}, \quad (1)$$

where $\theta$ and $\phi$ is the set of decoder parameters and encoder parameters, respectively; $D_{KL}(\cdot||\cdot)$ is the Kullback-Leibler divergence; $q_\phi(\cdot)$ is the variational posterior; $p(\cdot)$ is the true prior; $x$ is the training data; $y$ is the attribute representation; $z$ is the latent representation; $\hat{z}$ is the drawn sample.

### 2.2. CycleGAN2

Considering two acoustic feature sequences, $x \in R^{Q\times T_x}$ and $y \in R^{Q\times T_y}$ from source $X$ and target $Y$, where $Q$ is the feature dimension and $T_x$ and $T_y$ are the sequence lengths, respectively. CycleGAN2 seeks to learn a mapping function $G_{X\to Y}$ , which converts $x \in X$ into $y \in Y$ without the need of parallel



http://dx.doi.org/10.21437/Interspeech.2020-1647

data. This method uses an adversarial loss [15] and a cycle-consistency loss [16]. Additionally, to preserve the linguistic information, an identity-mapping loss [17] is added.

**Adversarial loss:** Conventional adversarial loss is used to make the generated feature $G_{X\to Y}(x)$ indistinguishable from the real target $y$:

$$L_{adv}(G_{X\to Y}, D_Y) = E_{y\sim P_Y(y)}[\log D_Y(y)] + E_{x\sim P_X(x)}[\log(1 - D_Y(G_{X\to Y}(x)))], \quad (2)$$

where the discriminator $D_Y$ and the generator $G_{X\to Y}$ try to find an equilibrium in this min-max game. $L_{adv}(G_{Y\to X}, D_X)$ has an analogous form. This adversarial loss helps to alleviate the over-smoothing effect, but this is not enough, because the L1-norm used in the cycle-consistency loss (see below) still causes over-smoothing. To mitigate this problem, an additional discriminator $D_X^{'}$ is introduced and an additional adversarial loss is defined on the circularly generated feature:

$$L_{adv2}(G_{X\to Y}, G_{Y\to X}, D_X^{'}) = E_{x\sim P_X(x)}[\log D_X^{'}(x)] + E_{x\sim P_X(x)}[\log(1 - D_X^{'}(G_{Y\to X}(G_{X\to Y}(x))))]. \quad (3)$$

Similarly, $D_Y^{'}$ and $L_{adv2}(G_{Y\to X}, G_{X\to Y}, D_Y^{'})$ can be introduced. As two adversarial losses ($L_{adv}(\cdot)$ and $L_{adv2}(\cdot)$) are defined, the authors in [11] call them *two-step adversarial losses*.

**Cycle-consistency loss:** The output of $G_{X\to Y}(x)$ is guided to follow the distribution of the real target under the restriction of the two-step adversarial losses. The linguistic consistency between input and output features, however, may not be guaranteed. Therefore, a cycle-consistency loss is used as a further regularization for the mapping:

$$L_{cyc}(G_{X\to Y}, G_{Y\to X}) = E_{x\sim P_X(x)}[||G_{Y\to X}(G_{X\to Y}(x)) - x||_1] + E_{y\sim P_Y(y)}[||G_{X\to Y}(G_{Y\to X}(y)) - y||_1], \quad (4)$$

where the bidirectional mappings are trained simultaneously, and thus training can be more stable. $||\cdot||_1$ means the L1-norm.

**Identity mapping loss:** An identity-mapping is adopted to regularize the generator to be close to an identity mapping when real samples of the target domain are provided as the input:

$$L_{id}(G_{X\to Y}, G_{Y\to X}) = E_{y\sim P_Y(y)}[||G_{X\to Y}(y) - y||_1] + E_{x\sim P_X(x)}[||G_{Y\to X}(x) - x||_1]. \quad (5)$$

The intuition behind this is that the model is supposed to preserve the input if it already looks like from the target domain.

CycleGAN2 employs an architecture called 2-1-2D Convolutional Neural Networks (CNN) for the generator network. A 2D CNN is thought to be better suited for converting features while preserving the original structures, as it restricts the converted region to local. In this network, 2D convolution is used for downsampling and upsampling, and 1D convolution is used for the main conversion process (i.e., residual blocks [18]).

For the discriminator, CycleGAN2 uses PatchGAN [19], which adopts convolution at the last layer and determines the realness on the basis of the patch. The main advantage is that less parameters are needed.

### 2.3. VAE-GAN

VAE [20] and GAN [15] are two mainstream generative models which have been studied deeply. These two models have their own advantages and drawbacks. In image generation tasks, for example, VAE tends to generate normal but blurred samples; the outputs of the GAN are usually quite clear but sometimes can be weird-looking. Therefore, VAE-GAN [21] has been proposed to overcome the drawbacks of each model, while their advantages are maintained. A typical architecture of a VAE-GAN is illustrated in Fig. 1.

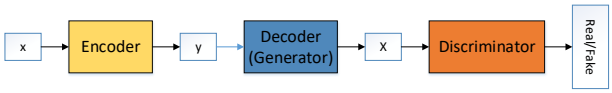



Figure 1: *Overview of VAE-GAN system. The decoder of VAE is viewed as the generator of GAN.*

### 2.4. Feature Selection and Representation Disentanglement in ESC

Previous studies on ESC usually handle acoustic features extracted by a vocoder such as STRAIGHT [22] or WORLD [23]. In [2], four acoustic features, namely $F_0$ contour, spectral sequence, duration and power envelop were explored. In [12], $F_0$ contour, spectral sequence and aperiodicity were extracted and conversion methods for the former two features were investigated. This paper focuses on the conversion models for $F_0$ and spectral features.

## 3. Proposed Method

The schematic diagram of the proposed method is shown in Fig. 2. Three acoustic features, namely spectral features, $F_0$ values and aperiodicity, are used. WORLD [23] is adopted as the vocoder for acoustic feature extraction and speech waveform reconstruction in this paper. The spectral features at each frame are represented by Mel-cepstral coefficients (MCEPs). $F_0$ values and aperiodicity are also extracted by WORLD analysis. Considering their different properties, the three acoustic features are converted using separate models in our method. We convert $F_0$ values through logarithm Gaussian normalized transformation [24] defined as:

$$f_{trg} = \exp((\log f_{src} - \mu_{src}) * \frac{\sigma_{trg}}{\sigma_{src}} + \mu_{trg}), \quad (6)$$

where $f_{src}$, $\mu_{src}, \sigma_{src}$ and $f_{trg}$, $\mu_{trg}, \sigma_{trg}$ are the $F_0$, mean and standard deviation from the source and target emotion set, respectively. We do not modify aperiodicity since it has little impact on emotion conversion.

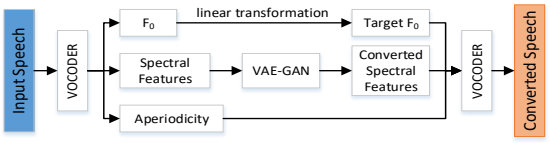



Figure 2: *Diagram of the proposed conversion method*

For spectral features, we train the VAE-GAN-based conversion model shown in Fig. 3. The main idea of the conversion method is to extract content-related representations in an unsupervised manner, and to extract emotion-related representations in a supervised way. In this paper, emotion labels are explicitly used during training and conversion. Emotion-related representations are extracted from an embedding layer, which

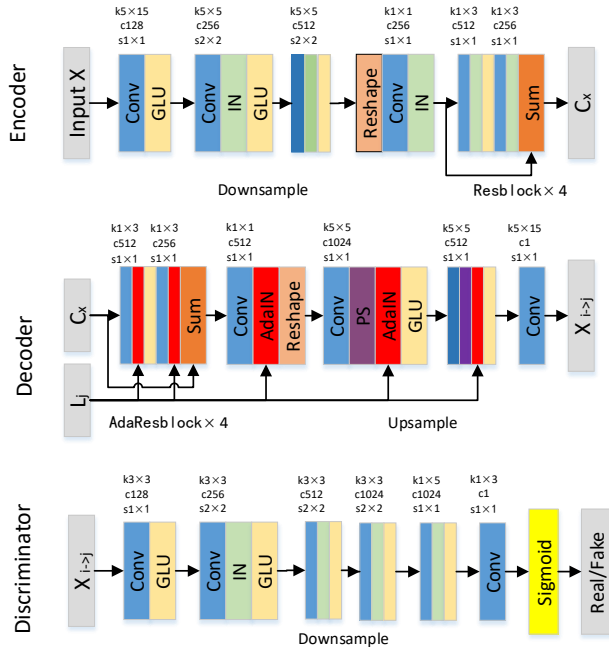



Figure 3: *Model Architecture of Encoder, Decoder and Discriminator. k, c and s denote kernel size, number of channels and stride size. "Conv", "GLU", "IN", "AdaIN", "PS", "Sigmoid" denote convolution, gated linear unit, instance normalization, adaptive instance normalization, pixel shuffler and sigmoid layers, respectively. $C_x$ and $L_j$ denote the content of input $X$ and the target label of emotion $j$, respectively. Note that the last layer of the decoder is a convolution layer.*

takes a one-hot vector of emotion as input. In the conversion stage, the content representations from the source speech and the emotion-related representations from the target speech are injected into the network to generate the desired speech. Theoretically, this idea is based on two assumptions:

*1. Different speech of the same emotion share the same emotion-related representations.*

*2. The emotion-related representations from speech of different emotions are different.*

The conversion model consists of three parts: an encoder, a decoder (or generator) and a discriminator. The encoder learns to encode the spectral features into content-related representations. The emotion labels of spectral feature segments are categorical features and encoded into emotion-related representations with embedding layers. These representations are then fed into the decoder, together with the output of the encoder. The discriminator receives the output of the decoder to judge whether it is from real data. The combination of the latter two parts, i.e. the decoder and the discriminator, can be viewed as a variant of CycleGAN2.

The encoder outputs the mean and variance of the latent vector, denoted as $\mu_z$ and $\sigma_z^2$, respectively. A KL loss is introduced to reduce the gap between the variational posterior and the true prior. In this work, $z$ is chosen to follow the isotropic standard normal distribution. So the KL loss is defined in a closed-form:

$$L_{KL}(q_\phi(\hat{z}|x)||p(z)) = -\frac{1}{2}\sum_{d=1}^{D}(1+\log\sigma_{z_d}^2-\mu_{z_d}^2-\sigma_{z_d}^2), \tag{7}$$

where $D$ is the dimension of the latent space. The goal of the proposed method is to minimize the following loss function:

$$\begin{aligned} L &= L_{adv}(G_{X\to Y}, D_Y) + L_{adv}(G_{Y\to X}, D_X) \\ &+ L_{adv2}(G_{X\to Y}, G_{Y\to X}, D_X^{'}) + L_{adv2}(G_{Y\to X}, G_{X\to Y}, D_Y^{'}) \\ &+ \lambda_{cyc}L_{cyc}(G_{X\to Y}, G_{Y\to X}) + \lambda_{id}L_{id}(G_{X\to Y}, G_{Y\to X}) \\ &+ L_{KL}(q_\phi(\hat{z}|x)||p(z)), \end{aligned} \tag{8}$$

where $\lambda_{cyc}$ and $\lambda_{id}$ are weights, and each term is presented in Eqn.2-5 and Eqn.7.

# 4. EXPERIMENTS

## 4.1. Experimental Setup

The IEMOCAP database [14] was used in our experiments. This database was recorded from ten actors in dyadic sessions for the purpose of studying expressive dyadic interaction from a multimodal perspective. The corpus contains approximately 12 hours of audio data and the emotion of each audio file is annotated into categorical labels by multiple annotators. Nine different emotions are included in the corpus. In this paper, we investigated four of them: angry, happy, sad and neutral. Since the number of labeled speech utterances in terms of their emotion is limited in general scenarios, we randomly selected 30 utterances for each of the four emotions from one speaker to form the training set.

The waveforms of the database were in 48 kHz PCM format and were downsampled to 16 kHz. Training samples with fixed length of 128 frames were randomly selected from the raw audio sequences. 24-order Mel-cepstral coefficients (MCEPs), 1-order logarithmic $F_0$ values and 513-order aperiodicity features were extracted using WORLD vocoder. The window length was 40ms and the frame shift was 5ms. The source and target MCEPs were normalized to zero-mean and unit-variance per dimension, using the statistics of the training set. $\lambda_{cyc}$ and $\lambda_{id}$ were empirically set to 10 and 5, respectively.

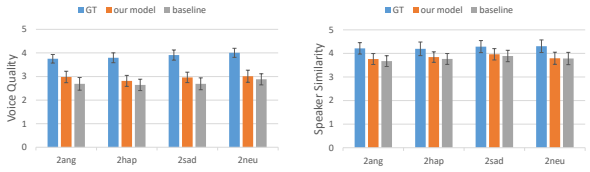
## 4.2. Network Architectures

The encoder network comprised of four convolutional layers and four residual blocks, and four adaptive residual blocks with four convolutional layers were used for the decoder network. There were six conventional layers and one sigmoid layer in the discriminator network. Instance normalization (IN) [25] was used to remove the emotion and speaker information while preserving content information for the encoder. All models were trained using Adam optimizer [26] with the momentum term $\beta_1$ of 0.5 and $\beta_2$ of 0.999. The batch size was set to 1. The initial learning rate was set to 0.0002 for the encoder and decoder and 0.0001 for the discriminator, respectively. The detailed information of network architecture is depicted in Fig. 3.

## 4.3. Experimental Results

Three groups of subjective evaluations were conducted to investigate the performance of the proposed method on voice quality, speaker similarity and emotion conversion ability, respectively [1]. The method proposed in [12] was chosen as the baseline.

[1]Examples of converted speech in our experiments can be found at https://siyizhou.github.io/CYX2019/index.html




(a) voice quality (b) speaker similarity

Figure 4: *MOS test for voice quality and speaker similarity with 95% confidence interval. (a): voice quality. (b): speaker similarity. "2ang" means the target emotion is angry. "GT" means the ground truth.*

#### 4.3.1. Voice Quality and Speaker Similarity

To evaluate voice quality and speaker similarity, Mean Opinion Scores (MOS) tests were conducted. Each conversion to one emotion from the other three emotions were investigated, making 12 groups of conversion ($4 \times 3 = 12$). For each of the 12 groups of conversion, 5 utterances from the test set were randomly selected, and there were 180 stimuli in total. Thirty listeners were recruited and they were asked to assign a score of 1-5 (5: excellent and 1: bad) to each stimulus. Two-tailed paired t-test was conducted to evaluate the significance of each MOS test.

The evaluation results of the voice quality and speaker similarity MOS tests are shown in Fig. 4. Fig. 4(a) shows that the proposed method outperforms the baseline in all four groups of conversion on voice quality. The t-test shows that all the p-values were less than 0.05, indicating the differences are significant. This demonstrates that the proposed method can improve the quality of converted speech, mainly because of the two-step adversarial loss adopted in CycleGAN2 which alleviates the over-smoothing effect. It is worth mentioning that the gap between the proposed method and the ground truth is big. This can be due to the background noise and reverberation in the recordings, which degrades the modeling accuracy greatly. From Fig. 4(b), we can see that both the proposed method and the baseline can achieve close quality with natural speech in terms of speaker similarity. Notably, all the ratings for the proposed method have exceeded 90% of those for original recordings, which proves the effectiveness of the proposed conversion system. The superiority of the proposed method over the baseline, however, is not significant according to the t-test. We expect to boost the performance by introducing some modules which explicitly handle speaker identity's information in the future.

#### 4.3.2. Emotion Conversion Ability

The emotion conversion ability of the proposed method was evaluated subjectively. We assessed all the possible conversion combinations of the four emotions (12 conversion pairs in total) and randomly selected 10 utterances from the test set for each conversion pair. Thirty subjects were asked to listen to these utterances and judge which emotion each utterance belonged to. We counted the subjects' votes and the results are summarized in Table 1. We can see that not all of the converted speech could be perceived correctly in terms of emotion. One reason is that part of emotions are conveyed by linguistic information, which can not be handled by the proposed method. The results of conversion with angry, sad and neutral as target are acceptable, which demonstrates the emotion conversion ability of the proposed method. However, conversion to happy did not perform well. This suggests that different emotions have their own properties, and specific operation for happy may be needed.

We also compared the proposed method and the baseline on emotion conversion ability. Thirty subjects participated the test and the results are exhibited in in Fig. 5. The proposed method outperforms the baseline in almost all the conversions except happy-to-angry and has a significantly higher average percentage change (45.81% vs 41.86%). This indicates the importance of the supervised manner used, which makes the extracted emotion-related representations more reliable.

Table 1: *Percentage change from source emotion to target emotion. Higher value indicates stronger conversion ability. For example, the first row means that in the conversion from neutral to angry, 44.67% converted speech are labeled "angry", 0%, 22.00% and 33.33% converted speech are labeled "happy", "sad" and "neutral", respectively, by the listeners.*

| | | **ang** | **hap** | **sad** | **neu** |
|---|---|---|---|---|---|
| 2ang | neu_ang | **44.67%** | 0 | 22.00% | 33.33% |
| | sad_ang | **48.33%** | 0 | 27.33% | 24.33% |
| | hap_ang | 38.00% | **46.00%** | 0 | 16.00% |
| 2hap | ang_hap | **50.33%** | 37.00% | 3.33% | 9.33% |
| | sad_hap | 6.00% | 38.00% | **41.67%** | 14.34% |
| | neu_hap | 0 | 34.33% | 27.00% | **38.67%** |
| 2sad | hap_sad | 7.00% | **34.67%** | 31.33% | 27.00% |
| | ang_sad | 27.00% | 0 | **47.33%** | 25.67% |
| | neu_sad | 4.33% | 9.67% | **49.67%** | 36.33% |
| 2neu | hap_neu | 0 | 36.00% | 3.33% | **60.67%** |
| | ang_neu | 29.33% | 0 | 7.33% | **63.33%** |
| | sad_neu | 0 | 0 | 43.00% | **57.00%** |

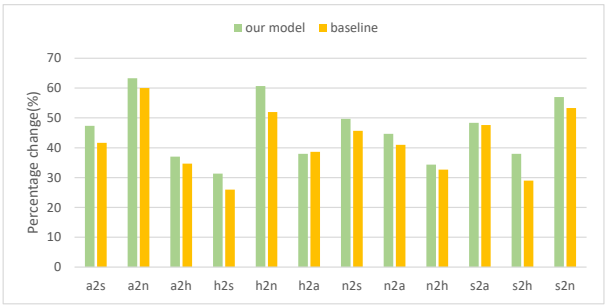



Figure 5: *Comparison of emotion conversion ability. Emotion state angry, happy, neural and sad is denoted as "a", "h", "n" and "s" respectively. For example, "a2s" means the conversion from angry to sad.*

## 5. Conclusions

We propose a novel method for emotional speech conversion. In this method, a VAE-GAN network is used to elicit the content-related and emotion-related representations in the training stage and recombine them in the conversion stage. Convincing emotion labels are used to make supervised learning possible. Future works will focus on seeking more effective techniques to extract emotion-related representations and adopting a neural vocoder to further boost the voice quality of converted speech.

# 6. References


[1] S. H. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.

[2] X. Yawen, H. Yasuhiro, and A. Masato, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.

[3] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic posteriorgrams," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5314–5318.

[4] R. Aihara, R. Takashima, T. Takiguchi, and Y. Ariki, "GMM-based emotional voice conversion using spectrum and prosody features," *American Journal of Signal Processing*, vol. 2, no. 5, pp. 134–138, 2012.

[5] Z. Luo, T. Takiguchi, and Y. Ariki, "Emotional voice conversion using deep neural networks with MCC and F0 features," in *IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1–5.

[6] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," *Interspeech*, 2017.

[7] 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.

[8] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "ACVAE-VC: Non-parallel many-to-many voice conversion with auxiliary classifier variational autoencoder," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1432 – 1443.

[9] P. L. Tobing, Y.-C. Wu, T. Hayashi, K. Kobayashi, and T. Toda, "Non-parallel voice conversion with cyclic variational autoencoder," *Interspeech*, 2019.

[10] 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," *Interspeech*, 2017.

[11] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6820–6824.

[12] J. Gao, D. Chakraborty, H. Tembine, and O. Olaleye, "Nonparallel emotional speech conversion," *Proc. Interspeech 2019*, 2019.

[13] T.-H. Kim, S. Cho, S. Choi, S. Park, and S.-Y. Lee, "Emotional voice conversion using multitask learning with text-to-speech," *ICASSP*, 2020.

[14] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[15] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, X. Bing, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, 2014.

[16] T. Zhou, P. Krahenbuhl, M. Aubry, Q. Huang, and A. A. Efros, "Learning dense correspondence via 3D-guided cycle consistency," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 117–126.

[17] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," in *Proceedings of the International Conference on Learning Representations*, 2017.

[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[19] C. Li and M. Wand, "Precomputed real-time texture synthesis with Markovian generative adversarial networks," in *European conference on computer vision*. Springer, 2016, pp. 702–716.

[20] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," *International Conference on Learning Representations, ICLR*, 2014.

[21] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, "CVAE-GAN: fine-grained image generation through asymmetric training," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2745–2754.

[22] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, 1997, pp. 1303–1306.

[23] A. V. D. Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio," *SSW*, p. 125, 2016.

[24] K. Liu, J. Zhang, and Y. Yan, "High quality voice conversion through phoneme-based linear mapping functions with STRAIGHT for mandarin," in *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, vol. 4. IEEE, 2007, pp. 410–414.

[25] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoPR*, vol. abs/1607.08022, 2016.

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR*, 2015.