

Nonparallel Emotional Speech Conversion

Jian Gao¹, Deep Chakraborty², Hamidou Tembine¹, Olaitan Olaleye³

¹ Department of Computer Science and Engineering, New York University, USA

² College of Information and Computer Sciences, University of Massachusetts Amherst, USA

³ Signify (formerly Philips Lighting) Research, North America, USA

{jg4631, tembine}@nyu.edu, dchakraborty@cs.umass.edu, olaitan.olaleye@signify.com

Abstract

We propose a nonparallel data-driven emotional speech conversion method. It enables the transfer of emotion-related characteristics of a speech signal while preserving the speaker's identity and linguistic content. Most existing approaches require parallel data and time alignment, which is not available in many real applications. We achieve nonparallel training based on an unsupervised style transfer technique, which learns a translation model between two distributions instead of a deterministic one-to-one mapping between paired examples. The conversion model consists of an encoder and a decoder for each emotion domain. We assume that the speech signal can be decomposed into an emotion-invariant content code and an emotion-related style code in latent space. Emotion conversion is performed by extracting and recombining the content code of the source speech and the style code of the target emotion. We tested our method on a nonparallel corpora with four emotions. The evaluation results show the effectiveness of our approach.

Index Terms: Emotional Speech Conversion, Non-parallel training, Style Transfer, Autoencoder, GANs

1. Introduction

Voice transformation (VT) is a technique to modify some properties of human speech while preserving its linguistic information. VT can be applied to change the speaker identity, i.e., voice conversion (VC) [1], or to transform the speaking style of a speaker, such as emotion and accent conversion [2]. In this work, we will focus on emotion voice transformation. The goal is to change emotion-related characteristics of a speech signal while preserving its linguistic content and speaker identity. Emotion conversion techniques can be applied to various tasks, such as enhancing computer generated speech, hiding negative emotions for people, helping film dubbing, and creating more expressive voice messages on social media.

Traditional VC approaches cannot be applied directly because they change speaker identity by assuming pronunciation and intonation to be a part of the speaker-independent information. Since the speaker's emotion is mainly conveyed by prosodic aspects, some studies have focused on modelling prosodic features such as pitch, tempo, and volume [3, 4]. In [5], a rule-based emotional voice conversion system was proposed. It modifies prosody-related acoustic features of neutral speech to generate different types of emotions. A speech analysis-synthesis tool STRAIGHT [6] was used to extract fundamental frequency (F_0) and power envelope from raw audio. These features were parameterized and modified based on Fujisaki model [7] and target prediction model [8]. The converted features were then fed back into STRAIGHT to resynthesize speech waveforms with desired emotions. However, this method requires temporal aligned parallel data that

is difficult to obtain in real applications; and the accurate time alignment needs manual segmentation of the speech signal at phoneme level, which is very time consuming.

To address these issues, we propose a nonparallel training method. Instead of learning one-to-one mapping between paired emotional utterances (x_1, x_2), we switch to training a conversion model between two emotional domains ($\mathcal{X}_1, \mathcal{X}_2$).

Inspired by disentangled representation learning in image style transfer [9, 10], we assume that each speech signal $x_i \in \mathcal{X}_i$ can be decomposed into a content code $c \in \mathcal{C}$ that represents emotion-invariant information and a style code $s_i \in \mathcal{S}_i$ that represents emotion-dependent information. \mathcal{C} is shared across domains and contains the information we want to preserve. \mathcal{S}_i is domain-specific and contains the information we want to change. In conversion stage, we extract content code of the source speech and recombine it with style code of the target emotion. A generative adversarial network (GAN) [11] is added to improve the quality of converted speech. Our approach is nonparallel, text-independent, and does not rely on any manual operation.

We evaluated our approach on IEMOCAP [12] for four emotions: angry, happy, neutral, sad; which are widely studied in emotional speech analysis literature. To our knowledge, this is the first attempt of nonparallel emotion conversion on this dataset, though [13] tried to generate synthetic feature representations of emotional speech. We evaluate the model's conversion ability by the percentage change from source emotion to target emotion. A subjective evaluation on Amazon MTurk with hundreds of listeners was conducted. It shows our model can effectively change emotions and keep the speaker identity.

The rest of the paper is organized as follows: Section 2 presents the relation to prior work. Section 3 gives a detailed description of our model. Experiments and evaluation results are reported in Section 4. Finally, we conclude in Section 5.

2. Related Work

2.1. Emotion-related features

Previous emotion conversion methods directly modify parameterized prosody-related features that convey emotions. [14] first proposed to use Gaussian mixture models (GMM) for spectrum transformation. A recent work [5] explored four types of acoustic features: F_0 contour, spectral sequence, duration and power envelope, and investigated their impact on emotional speech synthesis. The authors found that F_0 and spectral sequence are the dominant factors in emotion conversion, while power envelope and duration alone has little influence. They further claimed that all emotions can be synthesized by modifying the spectral sequence, but did not provide a method to do it. In this paper, we focus on learning the conversion models for F_0 and spectral sequence.

2.2. Nonparallel training approaches

Parallel data means utterances with the same linguistic content but varying in aspects to be studied. Since parallel data is hard to collect, nonparallel approaches have been developed. Some borrow ideas from image-to-image translation [15] and create GAN models [11] suitable for speech, such as VC-VAW-GAN [16], SVC-GAN [17], VC-CycleGAN [18], VC-StarGAN [13]. Another trend is based on auto-regressive models like WaveNet [19]. Although it can train directly on raw audio without feature extraction, the heavy computational load and huge amount of training data required is not affordable for most users.

2.3. Disentangled representation learning

Our work draws inspiration from recent studies in image style transfer. A basic idea is to find disentangled representations that can independently model image content and style. It is claimed in [9] that a Convolutional Neural Network (CNN) is an ideal representation to factorize semantic content and artistic style. They introduced a method to separate and recombine content and style of natural images by matching feature correlations in different convolutional layers. For us, the task is to find disentangled representations for speech signal that can split emotion from speaker identity and linguistic content.

3. Method

3.1. Assumptions

The research on human emotion expression and perception has two major conclusions. First, human emotion perception is a multi-layered process. [20] figured out that humans do not perceive emotion directly from acoustic features, but through an intermediate layer of semantic primitives. They introduced a three-layered model and learnt the connections by a fuzzy inference system. Some researchers found that adding middle layers can improve emotion recognition accuracy. Based on this finding, we suggest the use of multilayer perceptrons (MLP) to extract emotion-related information in speech signals.

Second, the emotion generation process of human speech follows the opposite direction of emotion perception. This means the encoding process of the speaker is the inverse operation of the decoding process of the listener. We assume that emotional speech generation and perception share the same representation methodology, that is, the encoder and decoder are inverse operations with mirror structures.

Let $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ be utterances drawn from two different emotional categories. Our goal is to learn a mapping between two distributions $p(x_1)$ and $p(x_2)$. Since the joint distribution $p(x_1, x_2)$ is unknown for nonparallel data, the conversion models $p(x_1|x_2)$ and $p(x_2|x_1)$ cannot be directly estimated. To solve this problem, we make two assumptions:

- (i). The speech signal can be decomposed into an emotion-invariant content code and an emotion-dependent style code;
- (ii). The encoder E and decoder G are inverse functions.

3.2. Model

Fig. 1 shows the generative model of speech with a partially shared latent space. A pair of corresponding speech (x_1, x_2) is assumed to have a shared latent code $c \in \mathcal{C}$ and emotion-related style codes $s_1 \in \mathcal{S}_1, s_2 \in \mathcal{S}_2$. For any emotional speech x_i , we have a deterministic decoder $x_i = G_i(c_i, s_i)$ and its inverse encoders $c_i = E_i^c(x_i), s_i = E_i^s(x_i)$. To convert emotion, we just extract and recombine the content code of the source speech

with the style code of the target emotion.

$$\begin{aligned} x'_{1 \leftarrow 2} &= G_1(c_2, s_1) = G_1(E_2^c(x_2), s_1) \\ x'_{2 \leftarrow 1} &= G_2(c_1, s_2) = G_2(E_1^c(x_1), s_2) \end{aligned} \quad (1)$$

It should be noted that the style code s_i is not inferred from one utterance, but learnt from the entire emotion domain. This is because the emotion style from a single utterance is ambiguous and may not capture the general characteristics of the target emotion. It makes our assumption slightly different from the cycle consistent constraint [21], which assumes that an example converted to another domain and converted back should remain the same as the original, i.e., $x'_{1 \leftarrow 2 \leftarrow 1} = x_1$. Instead, we apply a semi-cycle consistency in the latent space by assuming that $E_1^c(x'_{1 \leftarrow 2}) = c_1$ and $E_1^s(x'_{1 \leftarrow 2}) = s_1$.

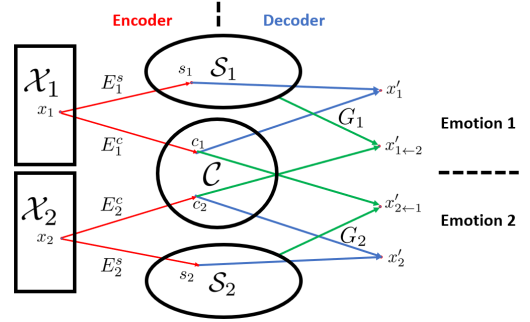


Figure 1: *Speech autoencoder model with partially shared latent space. Speech with emotion i is decomposed into an emotion-specific space \mathcal{S}_i and a shared content space \mathcal{C} . Corresponding speech (x_1, x_2) are encoded to the same content code.*

Fig. 2 shows an overview of our nonparallel emotional speech conversion system. The features are extracted and recombined by WORLD [22] and converted separately. We modify F_0 by linear transform to match statistics of the fundamental frequencies in the target emotion domain. The conversion is performed by log Gaussian normalization

$$f_2 = \exp((\log f_1 - \mu_1) \cdot \frac{\sigma_2}{\sigma_1} + \mu_2) \quad (2)$$

where μ_i, σ_i are the mean and variance obtained from the source and target emotion set. Aperiodicity (AP) is mapped directly since it does not contain emotion-related information.

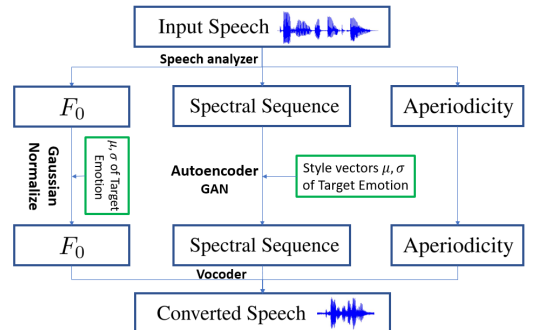


Figure 2: *Overview of nonparallel emotion conversion system*

For spectral sequence, we use low-dimensional representation in mel-cepstrum domain to reduce complexity. [28] shows

that 50 MCEP coefficients are enough to synthesize full-band speech without quality degeneration. Spectra conversion is learnt by the autoencoder model in Fig. 1. The encoders and decoders are implemented with gated CNN [23]. In addition, a GAN module is added to produce realistic spectral frames. Our model has 4 subnetworks E^c, E^s, G, D , in which D is the discriminator in GAN to distinguish between real samples and machine-generated samples.

3.3. Loss functions

We jointly train the encoders, decoders and GAN’s discriminators with multiple losses displayed in Fig. 3. To keep encoder and decoder as inverse operations, we apply reconstruction loss in the direction $x_i \rightarrow (c_i, s_i) \rightarrow x'_i$. The spectral sequence should not change after encoding and decoding.

$$L_{recon}^{x_i} = \mathbb{E}_{x_i}(\|x_i - x'_i\|_1), \quad x'_i = G_i(E_i^c(x_i), E_i^s(x_i)) \quad (3)$$

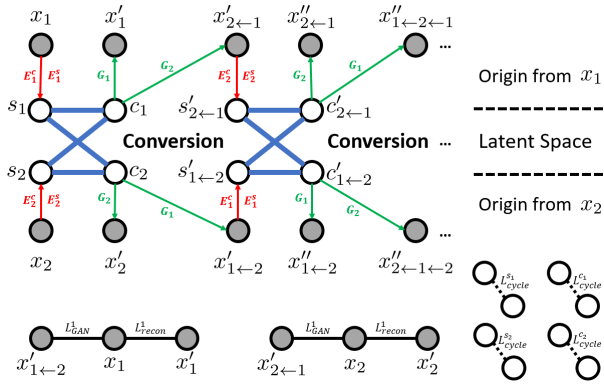


Figure 3: Train on multiple loss functions

In our model, the latent space is partially shared. Thus the cycle consistency constraint [21] is not preserved, i.e., $x'_{1 \leftarrow 2 \leftarrow 1} \neq x_1$. We apply a semi-cycle loss in the coding direction $c_1 \rightarrow x'_{2 \leftarrow 1} \rightarrow c'_{2 \leftarrow 1}$ and $s_2 \rightarrow x'_{2 \leftarrow 1} \rightarrow s'_{2 \leftarrow 1}$.

$$L_{cycle}^{c_1} = \mathbb{E}_{c_1, s_2}(\|c_1 - c'_{2 \leftarrow 1}\|_1), \quad c'_{2 \leftarrow 1} = E_2^c(x'_{2 \leftarrow 1}) \quad (4)$$

$$L_{cycle}^{s_2} = \mathbb{E}_{c_1, s_2}(\|s_2 - s'_{2 \leftarrow 1}\|_1), \quad s'_{2 \leftarrow 1} = E_2^s(x'_{2 \leftarrow 1})$$

Moreover, we add a GAN module to improve the speech quality. The converted samples should be indistinguishable from the real samples in the target emotion domain. GAN loss is computed between $x'_{i \leftarrow j}$ and x_i , ($i \neq j$).

$$L_{GAN}^i = \mathbb{E}_{c_j, s_i}[\log(1 - D_i(x'_{i \leftarrow j}))] + \mathbb{E}_{x_i}[\log D_i(x_i)] \quad (5)$$

The full loss is the weighted sum of L_{recon} , L_{cycle} , L_{GAN} .

$$\min_{E_1^c, E_1^s, E_2^c, E_2^s, G_1, G_2} \max_{D_1, D_2} L(E_1^c, E_1^s, E_2^c, E_2^s, G_1, G_2, D_1, D_2)$$

$$= \lambda_s(L_{cycle}^{s_1} + L_{cycle}^{s_2}) + \lambda_c(L_{cycle}^{c_1} + L_{cycle}^{c_2})$$

$$+ \lambda_x(L_{recon}^1 + L_{recon}^2) + \lambda_g(L_{GAN}^1 + L_{GAN}^2) \quad (6)$$

where $\lambda_s, \lambda_c, \lambda_x, \lambda_g$ control the weights of the components.

4. Experiments

Training emotional speech conversion models often suffers from lacking of data. Parallel datasets such as Emo-DB [24]

and RAVDESS [25] have limited sentence diversity and are difficult to build. Our end-to-end model is trained on raw audios of natural speech, and not rely on paired data or any manual operations. Training set can be collected from daily conversations in everyday life.

4.1. Experiment setup

We evaluated the proposed approach on the Interactive Emotional Dyadic Motion Capture database (IEMOCAP) [12]. It is organized in five sessions and contains 12 hours of audio-visual data. Each session records natural dialogues between a pair of speakers in scripted and improvised scenarios, in which the emotions are naturally elicited. In this paper, we only consider four emotional categories: 1) angry, 2) happy, 3) neutral, 4) sad. Since the model is not designed to change the speaker identity, experiments are conducted for each speaker independently. We only use the utterances with a clear majority vote regarding the ground truth labels. There are 2754 utterances shared amongst four emotional labels: ang (747), hap (675), neu (788), sad (544). Training and testing sets are non-overlapping utterances randomly selected from the same speaker (80% for training, 20% for test). For example in session 1, there are 420 training samples and 108 testing samples for the female speaker.

Training samples with fixed length of 128 are randomly selected from raw audio sequences. Energy-based voice-activity detection (VAD) is used to remove silent frames. We use WORLD [22] vocoder to extract fundamental frequencies (f_0 s), spectral sequences (sps) and aperiodicities (aps) from raw audio waveforms sampled at 16KHz. The frame length is 5ms. After coding, we take the first 24 Mel-cepstral coefficients (MCEPs) as feature vectors. Mean and variance of the entire training set are calculated for feature normalization. Testing samples can have arbitrary temporal length, and be converted in real-time.

4.2. Network architecture

The network architecture is illustrated in Fig. 4, with details listed in Table 1. The autoencoders take 24-dimensional MCEPs as input and learn disentangled representations of content and style. In the content encoder, instance normalization (IN) [26] removes the original feature mean and variance that represent emotional style information. In the style encoder, the emotional characteristics are encoded by a 3-layer MLP that outputs channel-wise mean and variance $\mu(s), \sigma(s)$. Then they are fed into the decoder to reconstruct MCEP features. The desired emotion is added through an adaptive instance normalization (AdaIN) [28] layer before activation. This mechanism is similar to the conversion model of F_0 in eq. (2).

$$\text{AdaIN}(c, s) = \sigma(s) \left(\frac{c - \mu(c)}{\sigma(c)} \right) + \mu(s) \quad (7)$$

The encoders and decoders are implemented with 1D-CNNs to capture the temporal dependencies, while the GAN discriminators are implemented with 2D-CNNs to capture the spectral-temporal patterns. Higher resolution data is generated by the pixel shuffler layer in upsample blocks. All networks use gated linear units (GLU) [23] to keep track of sequential information.

Training details: We use Adam optimizer with $\beta_1 = 0.5$. The learning rate is initialized as 0.0001 for D and 0.0002 for E^c, E^s, G ; it begins to linearly decay after 150K iterations. Weights are chosen as $\lambda_s = \lambda_c = \lambda_g = 1, \lambda_x = 10$. E^c, E^s, G are trained 2 iterations for each D ’s iteration in the first 100K iterations; after that they are trained equally.

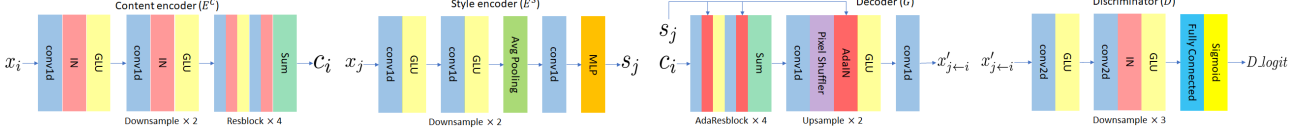


Figure 4: The network structure of content encoder, style encoder, decoder, and GAN discriminator.

Table 1: Network Architecture. *C-F-K-S-X* indicates convolution layer with filters *F*, kernel size *K*, strides *S*, and shuffle *X*. *IN* is instance normalization; all modules use *GLU* activation.

Content Encoder	
Conv1d, IN, GLU	C-128-15-1
Downsample1d $\times 2$	C-256-5-2, C-512-5-2
Resblock1d $\times 4$	C-512-3-1, content code
Style Encoder	
Conv1d, GLU	C-128-15-1
Downsample1d without IN $\times 2$	C-256-5-2, C-512-5-2
Downsample1d without IN $\times 2$	C-512-3-2
Adaptive average pooling	
Conv1d	C-16-1-1
MLP: linear $\times 2$	flatten, dense output
Decoder	
Adaptive Resblock1d $\times 3$	C-512-3-1
Upsample1d $\times 2$	C-512-5-1-2, C-256-5-1-2
Conv1d	C-24-15-1, MCEPs output
Discriminator	
Conv2d	C-128-(3,3)-(1,2)
Downsample2d	C-256-(3,3)-(2,2)
Downsample2d	C-512-(3,3)-(2,2)
Downsample2d	C-1024-(6,3)-(1,2)
Dense layer	sigmoid output (real/fake)

4.3. Experiment results

We evaluate the generated speech on three metrics: voice quality, speaker similarity, and the emotion conversion ability.

Subjective evaluation We perform perception tests on Amazon Mechanical Turk¹. Each utterance was listened by 5 random human workers, and each worker can answer at most 5 hits in a single experiment. To evaluate the voice quality and speaker similarity, the listeners were asked to give a 5-scale opinion score (5 for the best, 1 for the worst). The mean opinion score (MOS) is shown in Fig. 5. To annotate the emotion state, each listener was asked to choose a label from the source and target emotions. For example in the trial "ang2neu", utterances with label "ang" in IEMOCAP were converted, and the generated speech was labelled by the majority vote of human annotators. We compute percentage change from the source emotion to the target emotion. Higher value indicates stronger ability of emotional conversion. We choose four emotion pairs with significant differences [24]. The baseline models are a simple linear F0 conversion system [27], and a neutral network model VC-StarGAN [28]. Results are displayed in Fig. 6. More details and some converted speech samples are provided at².

Results and Discussion Note that not all utterances can be successfully converted, because some emotions are delivered

by linguistic information, a non-changeable part in our setting. Our model is slightly better than VC-StarGAN regarding the emotion conversion ability (average 48% vs 44%) and speaker similarity (average 3.55 vs 3.05). One reason is that VC-StarGAN is designed for voice conversion among different speakers, while our model learns the disentangled representations that can decompose the emotional characteristic and speaker identity. Moreover, VC-StarGAN has poor voice quality in the direction of sad2ang (1.71) and sad2hap (1.81). In [28], all emotions are trained together, thus it's unfair to the sad domain since it has lower signal to noise ratio, and may amplify the noise when converted to more energetic emotions.

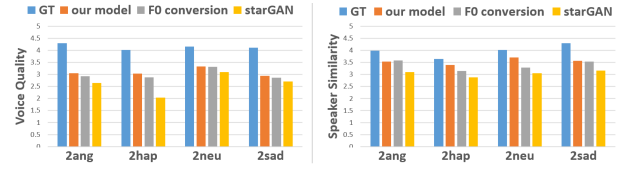


Figure 5: MOS for voice quality and speaker similarity. left: voice quality. right: speaker similarity, 2ang means the target emotion is Angry, and compared with originally Angry speech.

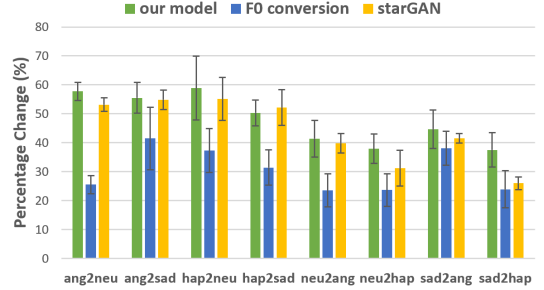


Figure 6: Comparison of the emotion conversion ability of our model and the baseline systems: (1) F0 conversion, (2) VC-starGAN [28]. ang2neu is conversion from Angry to Neutral.

5. Conclusion

We proposed a nonparallel emotional speech conversion approach based on style transfer autoencoders. Since our model does not require any paired data, transcripts or time alignment, it is easy to be applied in real-world situations. To our knowledge, this is the first work of nonparallel emotion conversion using style transfer. Future work is to include phonetic duration conversion and develop a general model for unseen speakers.

6. Acknowledgements

This research was supported by Signify Research and U.S. Air Force under grant FA9550-17-1-0259.

¹<https://www.mturk.com>

²<https://www.jian-gao.org/emovc>

7. References

- [1] S. Mohammadi and A. Kain, "An overview of voice conversion systems," *Speech Communication*, vol. 88, pp. 65–82, 2017.
- [2] G. Zhao, S. Souda, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent conversion using phonetic postergograms," in *ICASSP*. IEEE, 2018, pp. 5314–5318.
- [3] M. Wang, M. Wen, K. Hirose, and N. Minematsu, "Emotional voice conversion for mandarin using tone nucleus model–small corpus and high efficiency," in *Speech Prosody 2012*, 2012.
- [4] Z. Wang and Y. Yu, "Multi-level prosody and spectrum conversion for emotional speech synthesis," in *Signal Processing (ICSP)*. IEEE, 2014, pp. 588–593.
- [5] Y. Xue, Y. Hamada, and M. Akagi, "Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space," *Speech Communication*, vol. 102, pp. 54–67, 2018.
- [6] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds1," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.
- [7] H. Fujisaki and K. Hirose, "Analysis of voice fundamental frequency contours for declarative sentences of japanese," *ASJ's Japan (E)*, vol. 5, no. 4, pp. 233–242, 1984.
- [8] Y. Xue and M. Akagi, "A study on applying target prediction model to parameterize power envelope of emotional speech," in *RISP workshop NCSP'16*. , 2016.
- [9] L. Gatys, A. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *CVPR*. IEEE, 2016, pp. 2414–2423.
- [10] X. Huang, M. Liu, S. Belongie, and J. Kautz, "Multimodal unsupervised image-to-image translation," in *The European Conference on Computer Vision (ECCV)*, September 2018.
- [11] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NIPS*, 2014, pp. 2672–2680.
- [12] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.
- [13] S. Sahu, R. Gupta, and C. Y. Espy-Wilson, "On enhancing speech emotion recognition using generative adversarial networks," in *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018.*, 2018, pp. 3693–3697.
- [14] H. Kawanami, Y. Iwami, T. Toda, H. Saruwatari, and K. Shikano, "Gmm-based voice conversion applied to emotional speech synthesis," in *Eurospeech*, 2003.
- [15] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," in *Advances in Neural Information Processing Systems (NIPS)*. Curran Associates, Inc., 2017, pp. 700–708.
- [16] 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," in *Proc. Interspeech 2017*, 2017, pp. 3364–3368.
- [17] T. Kaneko, H. Kameoka, K. Hiramatsu, and K. Kashino, "Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks," in *Proc. Interspeech 2017*, 2017, pp. 1283–1287.
- [18] F. Fang, J. Yamagishi, I. Echizen, and J. Lorenzo-Trueba, "High-quality nonparallel voice conversion based on cycle-consistent adversarial network," *arXiv preprint arXiv:1804.00425*, 2018.
- [19] A. Van Den 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." in *SSW*, 2016, p. 125.
- [20] C. Huang and M. Akagi, "A three-layered model for expressive speech perception," *Speech Communication*, vol. 50, no. 10, pp. 810–828, 2008.
- [21] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE ICCV*, Oct 2017.
- [22] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Trans. on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.
- [23] Y. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *ICML*, 2017, pp. 933–941.
- [24] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *INTER-SPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005*, 2005, pp. 1517–1520.
- [25] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, 05 2018.
- [26] D. Ulyanov, A. Vedaldi, and V. S. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *CoRR*, vol. abs/1607.08022, 2016.
- [27] J. Tao, Y. Kang, and A. Li, "Prosody conversion from neutral speech to emotional speech," *IEEE Trans. Audio, Speech & Language Processing*, vol. 14, no. 4, pp. 1145–1154, 2006.
- [28] H. Kameoka, T. Kaneko, K. Tanaka, and N. Hojo, "Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks," in *2018 IEEE Spoken Language Technology Workshop, SLT 2018, Athens, Greece, December 18-21, 2018*, 2018, pp. 266–273.