EG-GAN: Cross-Language Emotion Gain Synthesis based on Cycle-Consistent Adversarial Networks

Xiaoqi Jia, Jianwei Tai, Qingjia Huang, Yakai Li, Weijuan Zhang, Haichao Du
Key Laboratory of Network Assessment Technology,
Institute of Information Engineering, CAS, Beijing, China.
School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China.
Beijing Key Laboratory of Network Security and Protection Technology.
{jiaxiaoqi,taijianwei,huangqingjia,zhangweijuan,duhaichao}@iie.ac.cn

Abstract

Despite remarkable contributions from existing emotional speech synthesizers, we find that these methods are based on Text-to-Speech system or limited by aligned speech pairs, which suffered from pure emotion gain synthesis. Meanwhile, few studies have discussed the cross-language generalization ability of above methods to cope with the task of emotional speech synthesis in various languages. We propose a cross-language emotion gain synthesis method named EG-GAN which can learn a language-independent mapping from source emotion domain to target emotion domain in the absence of paired speech samples. EG-GAN is based on cycle-consistent generation adversarial network with a gradient penalty and an auxiliary speaker discriminator. The domain adaptation is introduced to implement the rapid migrating and sharing of emotional gains among different languages. The experiment results show that our method can efficiently synthesize high quality emotional speech from any source speech for given emotion categories, without the limitation of language differences and aligned speech pairs.

1 Introduction

Consciously or not, the speech is a crucial way for human beings to convey emotional information in communication. As the naturalness of artificial synthesis speech is closer to that of real speech, the need to compensate for the deficiency of emotional expression in synthesized speech is becoming more obvious. The emotional speech plays a vital role in the field of human-computer interaction and robotics. Chiba et al. [1] pointed out that even in a non-task-oriented dialogue system, emotional speech can significantly improve user experience of the system. Moreover, with the rapid development of brain-computer interface technology, synthesized emotional speech can help language impaired people express their emotions directly in communication. These motivations have inspired researchers to try to synthesize emotional speech for the given emotional categories in the last decade. Unfortunately, compared with other human emotional signals such as gestures, facial expressions and postures, the emotion of speech is more complex and hard to regularize. To a certain extent, there is still a gap between synthesize emotional speech and real speech on authenticity and accuracy. In addition, most of the relevant research focuses on a few mainstream languages such as English and Mandarin while ignoring more minority languages. However, there is a commonality of the emotion gain which is a universal emotional representation between different languages. Unlike emotional speech, the emotion gain is a purer emotional representation that can be attached to any speech regardless of language differences. To cope with the deepening trend of globalization, the method for synthesizing cross-language emotion gain is urgently needed.
According to Schröder’s summary [2], the foundation for synthesizing more authentic emotional speech are accurate extraction of emotional features in phonetics, together with rational allocation of emotional features to speech. The analysis of emotional speech mainly includes prosody features and spectrum features. Based on speech spectrum and prosody modeling, the research on accurate extraction of emotional features has been continued to express desired emotions in synthetic speech. By changing the prosody, the emotional representation of speech can be changed obviously, and the adjustment of speech spectrum can affect the intensity of emotional representation. The methods based on gaussian mixture model (GMM) [3] and hidden markov model (HMM) [4] were proposed and successfully synthesized emotional speech in experimental environment. Although above works have achieved prospective results, due to the complexity of speech emotion, it is still difficult to implement the high-level control of emotional speech only by analyzing phonetics features.

Thanks to the tremendous progress of deep learning, deep learning is utilized to learn the mapping function of emotional transformation in high-dimensional space, which has a more superior capability to regularize the prosody of emotional speech than previous methods. In recent work [5, 6], convolutional neural network (CNN) and recurrent neural network (RNN) have been proved to be effective for this assignment. However, most of extant methods are based on fine-tuning text-to-speech system (TTS) or limited by paired data. The method of fine-tuning TTS system does not have the universality and it is a costly work to collect paired data. There is a lack of methods based on deep learning for independently extracting and synthesizing emotional gains of speech without relying on aligned data pairs. It is noteworthy that deep learning models have potential cross-domain migration capabilities. Coutinho et al. [7] have demonstrated that shared acoustic models between speech and music have emotional commonality. Transfer learning provides a feasible way to alleviate the limitation of inadequate datasets and motivates the cross-language generalization of emotion gain synthesis model.

In this paper, we propose EG-GAN, a novel cross-language emotion gain synthesis method. EG-GAN is based on cycle-consistency generative adversarial networks which can learn high-dimensional features between emotions and effectively solve the problem of unpaired data to expand available datasets. As a widely used transfer learning method, domain adaptation is introduced to implement the cross-language generalization performance of the model. Furthermore, to avoid gradient explosion/disappearance, we use the loss of the Earth-Mover (EM) distance [8] to replace the original loss of the model and add a gradient penalty [9] term for a stable convergence. It is well known that the loss of the generative adversarial model has no guiding significance for observing the quality of the output samples. In the field of images, FID [10] is widely used to evaluate the quality of generated images, while there are few similar metrics except the traditional ones with insufficient correlation with human perception on speech. Inspired by the latest research by Kilgour et al. [11], we first introduced FAD with high perception correlation to evaluate the quality of synthesized emotional speech. Experiments show that the proposed method can effectively synthesize high-quality emotional speech with excellent cross-language generalization performance. Through above discussions, our contributions can be summarized as follows:

- We propose a novel parallel-data-free emotion gain synthesis method named EG-GAN;
- We implement cross-language generalization performance of EG-GAN to simplify the migration of emotional gains to any language by domain adaptation;
- By introducing EM distance and gradient penalty, we provide a more stable training for the model without the issue of gradient explosion/disappearance;
- For the first time, as far as we know, FAD as a speech evaluation metric, is used to evaluate the performance of synthesized emotional speech with the high relevance to human perception.

2 Related Work

In order to transform neutral speech into emotional speech, many researchers focus on analyzing the prosody and spectrum features of emotional speech. Kawanami et al. [3] proposed GMM-based emotional speech synthesis method to synthesize emotional speech under specific emotion categories. Tao et al. [12] proposed a pitch target model for describing the F0 contour of Mandarin, which evaluates the expressiveness of synthesized emotional speech by the deviation of perceptual
Theoretically, adversarial training have the ability to learn stochastic functions $G$ which is an advanced TTS system. In recent years, emotional speech synthesis based on deep learning only confirm that $G$ maps randomly arranged samples in the target domain, where any known mapping can induce the context information of $G$. When the capacity is large enough, the network can map the same set of samples in the source domain identically distributed as target domain $Y$. The context information of the function learned by the model should be circularly consistent, in detail, the minimization and rules for modification, Xue et al. [15] used fuzzy inference system (FIS) to correlate emotion and prosody dimensions (valence and activation) and semantic primitives, and used Fujisaki model and target prediction model to parameterize prosody-related features [16]. Nevertheless, regularization of emotional representation in speech is still a serious challenge.

Many research works [17, 18, 19, 20] also discuss that how to fine-tune the existing Text-to-Speech System (TTS) by appropriately allocating prosodic parameters for increasing emotional expression in the output speech. Emotional speech synthesis method proposed by Lee et al. [21] and Skerry-Ryan et al. [22] has achieved impressive success through attaching extension modules for Tacotron which is an advanced TTS system. In recent years, emotional speech synthesis based on deep learning has become an emergent area of interest. LSTM-based emotional statistical parametric speech synthesis method was proposed by An et al. [5], and Li et al. [23] further tried to synthesize questionable and exclamatory speech rather than specific emotional categories with a real time synthesis system. Choi et al. [6] combines speaker information and emotional expression based on CNN by encoding speakers and emotions. However, the method that was based on deep learning relies on a large number of labeled and paired data, which is a tough work to obtain. In addition, few studies have focused on the model’s cross-language generalization performance, although it is an obvious problem to be solved for a rapid development of software services.

3 Our Method

3.1 CycleGAN

In this subsection, we start by briefly reviewing the concept and formulation of CycleGAN [24]. CycleGAN learns a mapping function between two domains $X$ and $Y$ without a set of aligned image pairs [24]. CycleGAN includes two mappings $G : X \rightarrow Y$ and $G : Y \rightarrow X$ which are learned by using two different losses, namely the adversarial loss [25] and the cycle-consistency loss [26] with two different discriminators. An adversarial loss is used to measure how distinguishable converted data $G_{X \rightarrow Y} (x)$ are from target data $y$, where $x \in X$ and $y \in Y$. Therefore, the closer the distributions of converted data $P_{G_{X \rightarrow Y}} (x)$ and target data $P_{data} (y)$ become, the smaller the adversarial loss will be. This adversarial loss function is written as

$$L_{adv} (G_{X \rightarrow Y}, D_Y, X, Y) = \mathbb{E}_{y \sim P_{data}(y)} [\log D_Y (y)] + \mathbb{E}_{x \sim P_{data}(x)} [\log (1 - D_Y (G_{X \rightarrow Y} (x)))]$$

where generator $G_{X \rightarrow Y}$ attempts to minimize this loss to generate speech samples indistinguishable from target data $y$ by the discriminator $D_Y$, while $D_Y$ aims to maximize this loss to distinguish between translated speech samples $G_{X \rightarrow Y}$ and real samples $y \in Y$. As described above, the formula can be expressed as $\min_{G_{X \rightarrow Y}} \max_{D_Y} L_{GAN} (G_{X \rightarrow Y}, D_Y, X, Y)$. In CycleGAN, there is a similar adversarial loss for the mapping function $G : Y \rightarrow X$ with its discriminator $D_X$, and the objective is shown as $\min_{G_Y} \max_{D_X} L_{GAN} (G_{Y \rightarrow X}, D_X, Y, X) [24]$. Theoretically, adversarial training have the ability to learn stochastic functions $G$ that produce outputs identically distributed as target domain $Y$ and source domain $X$ respectively [27]. Unfortunately, when the capacity is large enough, the network can map the same set of samples in the source domain to randomly arranged samples in the target domain, where any known mapping can induce an output distribution that matches the target distribution [24]. The context information of $X$ and $G_{X \rightarrow Y}$ is not necessarily consistent only by optimizing the adversarial loss. Because the adversarial loss only confirm that $G_{X \rightarrow Y} (x)$ follows the target data distribution, and do not help save the context information of $x$. In order to further reduce the possible mapping function space, the mapping function learned by the model should be circularly consistent, in detail, $G_{Y \rightarrow X} (G_{X \rightarrow Y} (x)) \approx x$.
and $G_{X \rightarrow Y} (G_{Y \rightarrow X}(y)) \approx y$. In CycleGAN, two additional items are introduced to address this problem. One is an adversarial loss $L_{GAN} (G_{Y \rightarrow X}, D_X, Y, X)$ for an inverse mapping $G_{Y \rightarrow X}$ and the other is a cycle-consistency loss, written as

$$L_{cyc} (G_{X \rightarrow Y}, G_{Y \rightarrow X}) = E_{x \sim P_{data}(x)} \left[ \| G_{Y \rightarrow X} (G_{X \rightarrow Y}(x)) - x \|_1 \right] + E_{y \sim P_{data}(y)} \left[ \| G_{X \rightarrow Y} (G_{Y \rightarrow X}(y)) - y \|_1 \right].$$  

(2)

The cycle-consistency loss make $G_{X \rightarrow Y}$ and $G_{Y \rightarrow X}$ to seek out $(x, y)$ pairs which have the same context information as far as possible. With weighted parameter $\lambda_{cyc}$, the full objective of CycleGAN [24] can be showed as

$$L (G_{X \rightarrow Y}, G_{Y \rightarrow X}) = L_{adv} (G_{X \rightarrow Y}, D_Y, X, Y) + L_{adv} (G_{Y \rightarrow X}, D_X, Y, X) + L_{cyc} (G_{X \rightarrow Y}, G_{Y \rightarrow X}).$$  

(3)

3.2 EG-GAN

We first describe our proposed model named EG-GAN, which is an emotion gain generation framework independent of parallel datasets. Then, we discuss how EG-GAN quickly extend emotional gains to different languages through transfer learning so that input speech can be efficiently converted into emotional target speech in different languages according to given emotion labels.

Our goal is to learn a language-independent mapping function between different emotional domains which not relies on aligned speech pairs. To achieve this goal, we implement the EG-GAN architecture based on CycleGAN [24]. In the EG-GAN architecture, we mainly made three modifications to the CycleGAN architecture: linguistic-information loss, speaker-verifying loss and gradient penalty [28]. Figure 1 illustrates the architecture of our proposed model.

**Linguistic-information loss** Although the loss of circular consistency provides constraints on the structure of learned mappings, such constraints are insufficient to ensure that learned mappings always retain linguistic information. Inspired by the loss of identity mapping [29], we introduce the linguistic-information loss to retain linguistic information of learned mappings without bringing an additional module into our model. Linguistic-information loss can encourage generators to learn mappings that preserve combinations between inputs and outputs. According to the definition of identity mapping, Linguistic-information loss can be written as

$$L_{li} (G_{X \rightarrow Y}, G_{Y \rightarrow X}) = E_{x \sim P_{data}(x)} \left[ \| G_{X \rightarrow Y}(x) - x \|_1 \right] + E_{y \sim P_{data}(y)} \left[ \| G_{Y \rightarrow X}(y) - y \|_1 \right].$$  

(4)

**Speaker-verifying loss** Recognition of a person by voice is an important human feature, and recognition the speaker’s identity is an important prerequisite for continuing communication [28]. There-
fore, a familiar speaker’s voice will make people more accustomed to listening to the speech. We aim to generate more pure emotional gains and to change the speaker’s phonological features as little as possible. In other words, adding emotional gains to speech through our model will not change the speaker verification features. In detail, we introduce a speaker feature extraction module with a discriminator to encode and verify the speaker. The speaker-verifying loss constrains the range of speaker features to ensure that the generated target speech has the same speaker attributes as the original speech. Inspired by VGGVox [30], we implement feature extraction and verification of speakers based on this model. Since speaker identification is considered as normal classification task [31], the output of the feature extraction module is fed into a softmax to produce a distribution over different speakers. The prediction of speaker identification is represented as $p_i$, and the value of real sample is $t_i$. In our method, the speaker-verifying loss can be simply written as

$$L_{sv} (G_{X→Y}, G_{Y→X}) = -\sum t_i \log (p_i).$$

(5)

**Gradient penalty** During GAN training process, the distribution of generated data is encouraged to approximate the distribution of real data. The objective function of GAN consists of all F divergences and exotic combinations. The traditional GAN always faces the problem of training instability, so we need to pay attention to the structure of the model, and the training level of generator and discriminator must be carefully coordinated. To address this problem, the Earth-Mover (EM) distance is introduced to replace the original loss measurement method while the weight clipping method is proposed to implement Lipschitz constraint indirectly [8]. In practical training, the discriminator loss is to maximize the difference between real and fake samples. However, weight clipping limits the range of values of each network parameter independently, so the optimal strategy will make all parameters go to extremes as far as possible. Therefore gradient penalty [9] is proposed to implement Lipschitz constraint as an optimization scheme for weight clipping to avoid undesired behavior. Due to the fact that Lipschitz constraint requires the gradient of discriminator not to exceed constant K, we introduce gradient penalty into EG-GAN to implement Lipschitz constraint with an additional loss term which establishes the relationship between the gradient and K. Gradient penalty will encourage more stable convergence of EG-GAN and circumvent gradient explosion/disappearance issues. Assuming that the random sample is represented as $P_{data}(\tilde{x})$. EG-GAN is based on cycle-consistent adversarial networks with the gradient penalty. The adversarial loss function of EG-GAN with the weight parameter $\lambda_{gradient}$ is written as

$$L_{adv} (G_{X→Y}, D_{Y}, X, Y) = \mathbb{E}_{x \sim P_{data}(x)} [D_{Y} (G_{X→Y} (x))] - \mathbb{E}_{y \sim P_{data}(y)} [D_{Y} (y)] + \lambda_{gradient} \mathbb{E}_{\tilde{x} \sim P_{data}(\tilde{x})} \left[ \left\| \nabla_{\tilde{x}} D_{Y} (\tilde{x}) \right\|_{2} - 1 \right]^{2}. \tag{6}$$

EG-GAN is constrained by three additional loss functions: linguistic-information loss, speaker-verifying loss and gradient penalty loss. Hence, our full objective with weight parameters $\lambda_{cyc} = 10$, $\lambda_{li} = 5$, $\lambda_{sv} = 1$ is

$$L_{full} (G_{X→Y}, G_{Y→X}) = L_{adv} (G_{X→Y}, D_{Y}, X, Y) + L_{adv} (G_{Y→X}, D_{X}, Y, X) + \lambda_{cyc} L_{cyc} (G_{X→Y}, G_{Y→X}) + \lambda_{li} L_{li} (G_{X→Y}, G_{Y→X}) + \lambda_{sv} L_{sv} (G_{X→Y}, G_{Y→X}). \tag{7}$$

**3.3 Domain Adaptation**

Domain adaptation is an important part of transfer learning, which was introduced by Daumé et al. in 2006. In our proposed scenario, it is assumed that each domain $D$ consists of feature space $Z$ and edge probability distribution $P(S)$, where the sample $S = s_1, s_2, ..., s_n \in Z$. Given a domain $D = \{Z, P(X)\}$, a task $T$ is composed of the emotion label space $E$ and the mapping function $f$. Because emotions are common in different languages, the emotional gains extracted from one language can be transferred to another language only through a small amount of training and data theoretically. In the task of generating emotional gains, the task domain $T_a, T_b$ based on different languages $A$ and $B$ can be expressed as $D_a$ and $D_b$. $P_a (E|S = s) = P_b (E|S = s)$ holds for $\forall s \in Z$, and $P_a (S) \neq P_b (S)$. The purpose of domain adaptation is to map data from different domains (e.g. $D_a, D_b$) into a emotion space $E$, and make it as close as possible in the space. Hence, the mapping function trained by the source domain $D_a$ in emotion space $E$ can then be migrated.
to the target domain $D_b$ to improve the accuracy of the mapping function in the target domain $D_b$. By introducing domain adaptation, EG-GAN can share the knowledge of emotional gains in different languages by sharing a part of model parameters, thus implement an excellent cross-domain generalisation performance with and without feature representation transfer. Furthermore, domain adaptation can alleviate the lack of labeled emotional speech.

4 Experiments

4.1 Datasets

We conducted our experiments to evaluate EG-GAN on emotion gain synthesis task without aligned speech pairs. Three different datasets are used in our experiments, which are interactive emotional dyadic motion capture database (IEMOCAP) [33] and Chinese emotional speech datasets (CEmoSD) provided by our research team. CEmoSD was recorded in a laboratory environment by four professional Chinese speakers (2 males and 2 females) according to four emotion categories (Angry, Happy, Neutral, Sad) and the corresponding sentences. To maintain uniformity in emotion categories, a subset was selected from IEMOCAP in our experiments: Angry, Happy, Neutral and Sad. We utilized CEmoSD and IEMOCAP to validate the performance of EG-GAN, while we used IEMOCAP to evaluate the cross-language generalization performance of EG-GAN with domain adaptation. In the preprocess, all speech inputs are sampled at 16000Hz, and 24-dimension Mel-ccepstral coefficients (MCEPs) are extracted by using the WORLD analysis system [34] afterwards. EG-GAN learned mappings between different domains which contained MCEPs of speech inputs.

4.2 Network Architectures

EG-GAN adopts the network architecture from CycleGAN-VC [35, 36] which have success in the voice conversion task. To capture the wide-range temporal structure while preserving the information of sample structure, the generator networks of EG-GAN is designed as a 1-dimension(1D) CNN [37] with several residual blocks [38], including 2 downsample blocks, 6 residual blocks and 2 upscale blocks. A gated CNN [39] is used in the generator to extract the acoustic features from time sequences and hierarchical structures. Due to outstanding results of high-resolution image generation [40], we use pixel shuffler in upsampling blocks. For the discriminator networks, we use a 2-dimension (2D) CNN to handle a 2D spectral texture [41]. To make a more stable convergence of the generator networks, the Earth-Mover (EM) distance is used in the discriminator which can distinguish real samples from fake samples to measure the discriminative loss. And we add a gradient penalty loss into the discriminator to prevent gradient explosion/disappearance issues by implementing the Lipschitz constrain. It’s worth mentioning that instance normalization [42] is used in both the generator and the discriminator to normalize a single sample in a batch. An auxiliary classifier is designed based on VGGnet [43] to extract, encode and classify speaker features, which retains unchanged speaker features during the process of emotion domain transformation.

4.3 Training

In our experiment, we used two Nvidia RTX2070 graphics cards with Ubuntu 16.04 (64 bits). In the preprocessing, MCEPs of all speech samples were extracted and normalized to unit variance and zero mean. We randomly cut a 128-frame segment from speech samples which have different lengths to ensure the randomness of the training process. The network is trained through the Adam optimizer ($\beta_1 = 0.5, \beta_2 = 0.9$). We set the generator’s learning rate to 0.0002, the discriminator’s learning rate to 0.0001, and the auxiliary speaker classifier’s learning rate to 0.0001. The weight parameters are set as $\lambda_{cycle} = 10, \lambda_1 = 5, \lambda_{ent} = 1$ and $\lambda_{gradient} = 5$. For migration learning on datasets of different languages, except the classification layer, all layers of the model are migrated to maximize the shared features of emotional gains by using domain adaptation. TTS samples provided by the Iflytek Open Platform [44] are used as test sets in all experiments. The Chinese TTS sample set (200 sentences) were generated through 50 sentences containing 4 speakers (2 males and 2 females), and the English TTS sample set (201 sentences) was generated through 67 sentences containing 3 speakers (2 males and 1 females). Note that no additional process is used to align speech samples, which ensures that the experimental results are not based on a training set with aligned speech pairs.
4.4 Objective Evaluation

As Kaneko et al. discussed in previous work [41], it is fairly intricate to use a single metric to comprehensively evaluate the quality of MCEPS after conversion. Therefore, the Mel-cepstral distortion (MCD) is used to evaluate the difference of global structure, and the modulation spectra distance (MSD) is used to evaluate the difference of local structure in the follow-up work of Kaneko et al. [36]. Although using two metrics to measure the quality of MCEPS is a feasible solution, in voice generation tasks, we want to directly evaluate the quality of synthesized speech rather than MCEPS as an intermediate result. On the other hand, the objective metric is expected to have a higher correlation with human perception, which makes the metric more instructive to model training. Thanks to FAD [11] proposed by Kilgour et al., it is possible to solve above problems. Compared with traditional evaluation metrics (distortion ratio, cosine distance and magnitude L2 distance), FAD can accurately evaluate the effect of synthesized speech and has a higher correlation with human perception as a novel, reference-free evaluation metric [11]. FAD has been innovatively introduced into our experiments for the first time to evaluate the quality of synthesized emotional speech. For FAD metrics, smaller values represent that the synthesized emotional speech is more similar to the real target emotional speech. We choose three advanced voice conversion models as the baseline methods in our experiments: StarGAN-VC [45], CycleGAN-VC [35], CycleGAN-VC2 [36].

Table 1 shows the FAD of synthesized emotional speech in four targeted emotion domains over dataset IEMOCAP and CEmoSD. Compared with the baseline methods, Table 1 proves that EG-GAN can synthesize higher quality emotional speech under limited iterations of the generator, besides it also compares the effects of domain adaptation on the quality of synthesized emotional speech on English datasets IEMOCAP [33]. Under limited iterations, using domain adaptive technology can significantly improve the quality of synthesized emotional speech. This advantage proves that the emotion gain learned by EG-GAN is common and transferable between different languages. In view of a reasonable inference, the proposed method can be extended to any other languages to implement cross-language emotion gain synthesis rapidly and efficiently.

4.5 Subjective Evaluation

We take a mean opinion score (MOS) as a subjective evaluation metric. A naturalness MOS test (terrible: 0 to outstanding: 5) is used to measure the naturalness of synthesized emotional speech. To verify the emotion correctness of synthesized emotional speech, we use an XYR MOS test, where “X” and “Y” are the emotional speech synthesized from the baseline methods and the proposed method respectively, while “R” is the real emotional speech in target domain. Synthesized emotional speech are randomly selected and played to the listener in random order to avoid the selection from sequential inertia. The listener was asked to choose one (“X” or “Y”) that was more similar to the emotion of the real voice R after playing “X” and “Y”. “Draw” can be chosen when the listener is too tough to make a clear choice. Similar to this approach, we use XYR MOS test to measure the speaker similarity between synthetic emotional speech and original speech, where “R” represents the original speech. In the naturalness MOS test and XYR MOS test, 20 sentences for each emotion category are
randomly selected from synthesized emotional speech. 20 listeners (10 males and 10 females) with good language proficiency in both Chinese and English were invited to participate in our experiment. For all MOS metrics, higher MOS values represent better performance of synthesized emotional speech in corresponding aspects.

Table 1 shows the naturalness MOS test results of emotional speech synthesized by different methods in the 200th generator iteration. It is obvious that the emotional speech synthesized by EG-GAN is better than other baseline methods in naturalness. Emotional speech synthesized were tested by XYR MOS for emotion correctness and speaker similarity, which are shown in Figure 2 and Figure 3. Compared with other baseline methods, Figure 2 illustrates the subjective preference of listeners for the emotional speech synthesized by EG-GAN. And Figure 3 proves that EG-GAN have a slightly superior performance over the task of maintaining the speaker representation. The experimental results demonstrate that EG-GAN has better performance in emotion correctness and speaker similarity with a high level control of emotional gains.

In order to measure the transfer effect of emotion gain in different languages intuitively, we compared the effects of direct training model and model trained from pre-training model of Chinese emotion gain by using domain adaptation technique. With the 200th generator iterations, Table 1 shows the naturalness MOS of the synthesized emotional speech, which proves that the performance of the model using domain adaptation technique is better than that of the model not used. Moreover, to some extent, Table 1 verified the correlation between FAD and human subjective perception.

5 Discussion and Conclusion

We propose a novel parallel-data-free cross-language emotion gain synthesis method, called EG-GAN, which can synthesize emotional speech from any input speech for given emotion categories without fine-tuning TTS system. In order to synthesize purer emotion gain by constraining the change of speaker characteristics, EG-GAN learns the mapping function from the input speech to the target speech without aligning the time sequences and speech samples. Meanwhile, due to emotional commonality in different languages, the emotion gain learned from one language can be quickly applied into another language through domain adaptation. The objective evaluation and the subjective evaluation shows that the emotional speech synthesized by EG-GAN has a more superior performance on emotion correctness and speaker retention than the existing methods [45, 35, 36]. Both subjective and objective evaluation proved that the use of domain adaptive technology can...
effectively achieve the migration of emotion gain between different languages for the task of cross-language emotion gain synthesis. However, compared with the original speech, the synthesized emotional speech still has low distortion; the synthesized emotional speech has a tiny gap with the real emotional speech in emotion correctness. In future work, we plan to improve the accuracy and authenticity of emotion gain synthesis by optimizing preprocessing methods, model structure and training strategies. It is predictable that the proposed model can be applied not only to the task scenarios presented in this paper, but also to other interest directions [46, 47, 48, 37, 45].

References

[1] Yuya Chiba, Takashi Nose, Taketo Kase, Mai Yamanaka, and Akinori Ito. An analysis of the effect of emotional speech synthesis on non-task-oriented dialogue system. In SIGDIAL Conference, pages 371–375, 2018.

[2] Marc Schröder. Emotional speech synthesis: A review. In Proc. EUROSPEECH Scandinavia, pages 561–564, 2001.

[3] Hiromichi Kawanami, Yohei Iwami, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Gmm-based voice conversion applied to emotional speech synthesis. In INTERSPEECH, 2003.

[4] Muntaz Begum Mustafa and Raja Noor Aion. Emotional speech acoustic model for malay: Iterative versus isolated unit training. Journal of the Acoustical Society of America, 134(4):3057–3066, 2013.

[5] Shumin An, Zhenhua Ling, and Lirong Dai. Emotional statistical parametric speech synthesis using lstm-rnns. In 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pages 1613–1616, 2017.

[6] Heejin Choi, Sangjun Park, Jinuk Park, and Minsoo Hahn. Multi-speaker emotional acoustic modeling for cnn-based speech synthesis. In ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[7] Eduardo Coutinho and Björn Schuller. Shared acoustic codes underlie emotional communication in music and speech—evidence from deep transfer learning. PLOS ONE, 12(6), 2017.

[8] Martín Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In International Conference on Machine Learning, pages 214–223, 2017.

[9] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of wasserstein gans. neural information processing systems, pages 5767–5777, 2017.

[10] Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. neural information processing systems, pages 700–709, 2018.

[11] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A metric for evaluating music enhancement algorithms. arXiv preprint arXiv:1812.08466, 2018.

[12] Jianhua Tao, Yongguo Kang, and Aijun Li. Prosody conversion from neutral speech to emotional speech. IEEE Transactions on Audio, Speech, and Language Processing, 14(4):1145–1154, 2006.

[13] Miaomiao Wen, Miaomiao Wang, Keikichi Hirose, and Nobuaki Minematsu. Prosody conversion for emotional mandarin speech synthesis using the tone nucleus model. In INTERSPEECH, pages 2797–2800, 2011.

[14] Ryo Aihara, Ryochi Takashima, Tetsuya Takiguchi, and Yasuo Ariki. Gmm-based emotional voice conversion using spectrum and prosody features. American Journal of Signal Processing, 2(5):134–138, 2012.

[15] Yawen Xue, Yasuhiro Hamada, and Masato Akagi. Emotional speech synthesis system based on a three-layered model using a dimensional approach. In 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pages 505–514, 2015.

[16] Yawen Xue, Yasuhiro Hamada, and Masato Akagi. Voice conversion for emotional speech: Rule-based synthesis with degree of emotion controllable in dimensional space. Speech Communication, 102:54–67, 2018.

[17] Mostafa Al Masum Shaikh, Antonio Ferreira Rebordao, and Keikichi Hirose. Improving tts synthesis for emotional expressivity by a prosodic parameterization of affect based on linguistic analysis. 5th International Conference on Speech Prosody (Speech Prosody 2010), 2010.
[18] Masato Akagi, Xiao Han, Reda Elbarougy, Yasuhiro Hamada, and Junfeng Li. Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*, pages 1–10, 2014.

[19] Blaise Potard, Matthew P. Aylett, and David A. Baude. Cross modal evaluation of high quality emotional speech synthesis with the virtual human toolkit. *Intelligent Virtual Agents*, pages 190–197, 2016.

[20] Jayanta Kalita and Nabamita Deb. Emotional text to speech synthesis: A review. *International Journal of Advanced Research in Computer and Communication Engineering*, 6(4):428–430, 2017.

[21] Younghun Lee, Azam Rabiee, and Soo-Young Lee. Emotional end-to-end neural speech synthesizer. *arXiv preprint arXiv:1711.05447*, 2017.

[22] Rj Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J. Weiss, Rob Clark, and Rif A. Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. *International Conference on Machine Learning*, pages 4693–4702, 2018.

[23] Hao Li, Yongguo Kang, and Zhenyu Wang. Emphasis: An emotional phoneme-based acoustic model for speech synthesis system. In *Interspeech 2018*, pages 3077–3081, 2018.

[24] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017.

[25] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[26] Tinghui Zhou, Philipp Krähenbühl, Mathieu Aubry, Qixing Huang, and Alexei A. Efros. Learning dense correspondence via 3d-guided cycle consistency. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 117–126, 2016.

[27] Ian J. Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[28] John H. L. Hansen and Taufiq Hasan. Speaker recognition by machines and humans: A tutorial review. *IEEE Signal Processing Magazine*, 32(6):74–99, 2015.

[29] Yaniv Taigman, Adam Polyak, and Lior Wolf. Unsupervised cross-domain image generation. *International Conference on Learning Representations*, 2017.

[30] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech 2018*, pages 1086–1090, 2018.

[31] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: A large-scale speaker identification dataset. In *Interspeech 2017*, pages 2616–2620, 2017.

[32] Hal Daumé and Daniel Marcu. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26(1):101–126, 2006.

[33] Carlos Busso, Murtaza Bulut, Chi Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4):335–359, 2008.

[34] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. World: A vocoder-based high-quality speech synthesis system for real-time applications. *IEICE Transactions on Information and Systems*, 99(7):1877–1884, 2016.

[35] Takuhiro Kaneko and Hirokazu Kameoka. CycleGAN-VC: Non-parallel voice conversion using cycle-consistent adversarial networks. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2100–2104, 2018.

[36] Takuhiro Kaneko, Hirokazu Kameoka, Kou Tanaka, and Nobukatsu Hojo. CycleGAN-VC2: Improved cycleGAN-based non-parallel voice conversion. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019.

[37] Takuhiro Kaneko, Hirokazu Kameoka, Kaoru Hiramatsu, and Kunio Kashino. Sequence-to-sequence voice conversion with similarity metric learned using generative adversarial networks. In *Interspeech 2017*, pages 1283–1287, 2017.
[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[39] Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. *International Conference on Machine Learning*, pages 933–941, 2016.

[40] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1874–1883, 2016.

[41] Takuhiro Kaneko, Hirokazu Kameoka, Nobukatsu Hojo, Yusuke Iijima, Kaoru Hiramatsu, and Kunio Kashino. Generative adversarial network-based postfilter for statistical parametric speech synthesis. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4910–4914, 2017.

[42] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations*, 2015.

[44] iflytek.com. Iflytek open platform. [https://www.xfyun.cn/services/online_tts](https://www.xfyun.cn/services/online_tts).

[45] Hirokazu Kameoka, Takuhiro Kaneko, Kou Tanaka, and Nobukatsu Hojo. Stargan-vc: non-parallel many-to-many voice conversion using star generative adversarial networks. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 266–273, 2018.

[46] Alexander B. Kain, John Paul Hosom, Xiaochuan Niu, Jan P H van Santen, Melanie Fried-Oken, and Janice Staehely. Improving the intelligibility of dysarthric speech. *Speech Communication*, 49(9):743–759, 2007.

[47] Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano. Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech. *Speech Communication*, 54(1):134–146, 2012.

[48] Tomoki Toda, Mikihiro Nakagiri, and Kiyohiro Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing*, 20(9):2505–2517, 2012.