Speech Synthesis of Children's Reading Based on CycleGAN Model

Ning JIA*, Chunjun Zheng1,2, Wei Sun1
1School of Computer and Software, Dalian Neusoft University of Information, Liaoning Dalian
2Information Science and Technology College, Dalian Maritime University, Liaoning Dalian
jianing@neusoft.edu.cn

Abstract. The generation of emotional speech is a challenging and widely applied research topic in the field of speech processing. Because the design method of effective speech feature expression and generation model directly affects the accuracy of emotional speech generation, it is difficult to find a general solution of emotional speech synthesis. In this paper, the CycleGAN model is used as the starting point, and the improved convolution neural network (CNN) model and identity mapping loss scheme are used to achieve effective timing information capture. At the same time, we learn the positive mapping and the reverse mapping to find the best matching design scheme, and retain the speech information in this process, without relying on other audio data. Experiments show that the emotional speech can be accurately recognized by comparing the speech emotion before and after the improvement on the speech corpus of children's reading. By comparing with the common emotional speech generation model, the advantages of the model proposed in this paper are verified.

1. Introduction
With the increasing frequency of human-computer "dialogue", the technology of human-computer interface has stepped into the era of multimedia users from the era of mechanization. People want to have a way of communication with machines that is similar to people to people. Speech is an important tool for human communication. It is a long-term goal for people to make computers have the ability of "listening" and "speaking" like people. Speech synthesis technology is one of the key technologies to achieve this goal [1].

At present, speech synthesis technology has been greatly developed under the promotion of natural language processing, signal processing and random process processing. The intelligibility of synthetic speech has basically met the requirements, and the naturalness also has a good performance. How to further improve the expressive power of synthetic speech and give it rich mood and emotional color has become a new problem to be solved in the field of speech synthesis. This also gave birth to an important branch in the field of speech synthesis, that is, the development of emotional speech synthesis [2].

Affective speech synthesis is a combination of affective computing and speech synthesis. With the help of the concept of affective computing, this paper analyzes the relationship between emotion and speech from the speech signals of known emotional state, and applies these emotional features to the speech synthesis process, so as to obtain a natural and friendly synthetic voice with rich mood changes
and able to simulate human emotions [3]. Emotional speech synthesis can not only promote human-computer interaction, but also promote the development of artificial intelligence. It also has a good application prospect [4]. For example, it can be used to monitor the driver's mental state when driving and remind when the driver is tired, so as to avoid the occurrence of traffic accidents. It is also used in the remote network classroom to analyze the students' emotional state, adjust the teaching method and schedule in time, and improve the teaching quality, so as to achieve good teaching effect. It can be used to track the emotional state of patients with depression and provide basis for the diagnosis and treatment of the disease.

For emotional speech synthesis, it is necessary to study not only adults, but also children's. The research in this area is helpful to cultivate children's social emotional ability, and also to a great extent affects children's early development. Compared with adult speech emotion recognition, children's emotion speech database resources are scarce, and the development of children's vocal cords is not mature, which makes the research of this paper very challenging.

2. RELATED WORK

At present, speech synthesis technology usually refers to text to speech (TTS) technology, which mainly solves how to transform text information into audible voice information. A typical speech synthesis system consists of three modules: text analysis, prosodic generation and synthesizer.

The existing emotional speech synthesis system is usually based on the general TTS to add specific emotional information, so that it can participate in the control and adjustment of rhythm, and then generate the speech with the target emotion. There are also related systems that use specific emotion speech database to synthesize specific emotion speech. Emotional information is usually offline and directly set by human. Emotional speech synthesis system should be able to combine text content and scene factors for automatic emotion prediction [5].

The main problems of emotion speech synthesis are as follows: the current research in this field focuses on how to express emotion in speech, while the description of emotion state itself is not enough. Compared with adult speech emotion recognition, the focus of children's speech emotion expression is not clear enough. Moreover, there are few child corpora available for this kind of research.

With the rise of neural network, a lot of researches on speech generation using neural network have emerged. The speech generation method based on deep learning can mine the potential correspondence from a large number of corpus and automatically learn the dependency from the source sequence to the target sequence. Many researchers have applied the "sequence to sequence" generation framework to speech tasks, and achieved some results. For example, research institutions such as MIT Artificial Intelligence Laboratory in the United States, Queen's College in Belfast in the United Kingdom, Department of electronics and computer of Keio University in Japan have begun to conduct in-depth research on the emotional elements contained in speech [6].

In China, Zhou et al. [7] added the influence of emotional factors to the framework of basic dialogue model, and introduced three mechanisms of emotion. Ghosh et al [8] proposed an affective text generation framework “Affect-LM”, which can generate more real affective sentences. Wang et al. proposed Senti-GAN, an emotional text generation framework, which can generate diverse, high-quality texts with different emotional categories. However, this framework can only discriminate and grade complete sequences.

In other countries, Chris et al. proposed to use WaveGAN framework to generate audio signal [9], and applied GAN to unsupervised synthesis of original audio waveform. The speech synthesized by DCGAN has great difference in waveform and frequency, so it is not suitable for spectrogram modeling. Therefore, Kaneko et al. [10] proposed a spectrum post filtering method based on GAN, which is used to filter the predicted spectrum on the basis of traditional speech synthesis.

In this paper, we set the research object as reading speech. Compared with natural spoken language, it is a more regular language form, and the expression of emotion is more stable and reproducible. In the background of reading aloud, the influence of the scene constitutes the weak sensitive environment...
factor, and the background music, environment, the role of the speaker and the object-oriented are relatively fixed.

Therefore, the research in this paper does not consider the influence of the scene temporarily, but focuses on the emotional synthesis of specified speech based on children's reading scenes and known emotional themes. According to the characteristics of children's speech, we choose cycle consistent adversarial networks (CycleGAN) [11] model to generate specific emotional speech.

3. Design of Speech Synthesis Model Based On CycleGAN

3.1. Cycle Consistent Adversarial Networks (CycleGAN)

Cycle Consistent Adversarial Networks was originally used in image domain to find a way to map an image from the source domain to the target domain without any additional information. It can turn gray image into color image and ordinary horse into zebra with black and white stripes. It can also turn an ordinary painting into a work with the style of masters such as Monet and Van Gogh.

The core idea of CycleGAN is a ring structure, mainly composed of converter \( F \) and \( G \) (a converter contains a set of generation model and discrimination model). As shown in the figure, \( X \) represents the original data of \( X \) domain and \( Y \) represents the original data of \( Y \) domain. The original data of \( X \) domain is first generated by converter \( G \), and then reconstructed by converter \( F \). In the same way, the original data of \( Y \) domain is first generated by converter \( F \), and then reconstructed back to the original data of \( Y \) domain. The discriminant models \( D_X \) and \( D_Y \) play the role of discriminant, promoting the data to complete the style migration.

Therefore, for a raw data \( x \) in domain \( X \), we expect the result of function \( G(F(x)) \) to be the same as that of \( x \). Similarly, for a raw data \( y \) in domain \( Y \), we expect the result of function \( F(G(y)) \) to be the same as that of \( Y \). Figure 1 depicts the transfer process from one domain to another.

The CycleGAN model has two loss functions:

1. Adversarial loss: It matches the distribution of the generated data and the allocation of the target domain.
2. Cycle-consistency loss: It is used to avoid conflicts between the learned converters \( G \) and \( F \).

![Figure 1. the Transfer Process](image)

The complete CycleGAN objective function is shown in formula 1-2:

\[
G^*, F^* = \arg\min_{G,F} \max_{D_X,D_Y} L(G,F,D_X,D_Y)
\]

where

\[
L(G,F,D_X,D_Y) = L_{GAN}(G,D_Y,X,Y) + L_{GAN}(F,D_X,Y,X) + \lambda L_{cyc}(G,F)
\]

3.2. The Speech Synthesis Model of Children's Reading Based on CycleGAN

The original CycleGAN mainly deals with image tasks. It uses adversarial loss and cycle-consistency loss to learn both positive mapping and inverse mapping, so that the best pairing can be found from unmatched data. The goal of this paper is to achieve the effective mapping of \( X \) domain and \( Y \) domain through continuous learning for children's speech data. At the same time, effective emotional speech is generated, and it does not depend on other audio in the process. Therefore, we need to make some modifications to the CycleGAN architecture, that is, design convolutional neural networks (CNN) model and identity-mapping loss.
There is a significant feature of speech data, which is a set of units with time series. When designing, it is often combined with the context dependence to process. An effective way to learn this structure is to use Recurrent Neural Network (RNN). RNN needs to consider the influence of long-term and short-term memory, which leads to its high computational requirements and directly affects the learning effect. Therefore, we use the improved CNN model to configure CycleGAN, which not only allows parallel processing of sequential data, but also realizes the model construction of emotional speech.

Considering the variability of global feature expression in children's reading aloud, the direct use of global features for calculation would discard its temporal relationship, resulting in the decline of discrimination ability. At this time, children's pronunciation is in the development stage, at this time, the range of low-level description characteristics changes is large. Based on this, we can preprocess the features of speech reading and retain as many structural features as possible to ensure that the features of children's speech can be effectively acquired.

In this paper, all information is read in the form of data stream, and the default sampling rate is 16,000. All information read in this form is the first generation feature of speech. This feature representation covers all the data of current speech and covers a wide range. At this time, the speech contains a small amount of background and noise. At this time, it needs to separate the information of signal and background, and only keep the information related to the speaker's speech. This process can be understood as a simplified version of the denoising scheme, where the improved spectral subtraction is used. This method takes advantage of the fact that additive noise is not related to speech. Assuming that the noise is a stationary signal, the noise data is calculated by the gap without speech, and the information of noise during speech is estimated, which is subtracted from the noisy speech data to obtain the estimated value of speech spectrum.

On this basis, all data sources are calculated as follows to generate image forms suitable for CNN.

Suppose the length of the speech is \( N \), then \( \sqrt{N} \) is taken as the size of the generated image, and the image with the size of \( \sqrt{N} \times \sqrt{N} \) is to be generated. The generation rules are as follows:

\[
t_{i,j} = p_{i,\sqrt{N},j}, i < \sqrt{N}
\]

(3)

Where \( t_{i,j} \) is the matrix representation of the original data. After normalization, the pixel matrix \( t'_{i,j} \) is generated, which is used as the third dimension data to directly generate the image expression of audio.

On this basis, CNN model is designed, which is composed of input layer, 6-layer convolution layer, 6-layer pooling layer and full connection layer. Each feature map has equal weight, also known as weight sharing, which can greatly reduce the scale of parameters and accelerate the training speed.

Convolution layer is used for feature extraction. Through a set of convolution kernels, the neurons of the current layer are connected to several characteristic graphs of the previous layer for convolution operation, and then the characteristic graphs of the current layer are obtained by adding bias. Each convolution layer is 512*3, 1024*3, 256*3, 128*3, 32*3 and 2*3. The mapping relationship before and after convolution layer is as follows.

\[
x^m_j = f\left(\sum_{n=0}^{m-1} x^{n-1}_j * k^{n}_{i,j} + b^{n}_j\right)
\]

(4)

Where \( x^m_j \) represents the input of the \( j \) th characteristic graph of the \( m \) th convolution layer. \( k^{n}_{i,j} \) represents the convolution kernel, \( b^{n}_j \) represents bias, \(* \) represents the convolution operation, \( M_j \) represents the set of characteristic graphs, and \( f \) represents the activation function.

The method of maximum pooling is adopted to make it robust to small changes of input. The activation function of gating linear units (GLUs) is used to build a deep network model to learn more representative features by continuously overlapping convolution layer and pooling layer. There are 512 neurons in the full connection layer, which map the learned distributed feature representation to the
speech signal space. The dropout operation is used to prevent overfitting. The structure of CNN model is shown in Figure 2.

(b) Identity-mapping Loss

The cycle consistency loss mentioned above provides constraints for the structure, but it is not enough to ensure that the original mapping always retains the historical speech information. Therefore, in order to keep as much effective speech information as possible without relying on other models and data, we add the identity mapping loss. In Formula 1, \( L_{\text{cy}} \) is modified as follows. This way of generating \( \langle x, y \rangle \rightarrow \text{cy} \) helps the generator find a valid mapping between input and output.

\[
L_{\text{cy}}(G, F) = E_{y_D} \left[ \|G(F(x)) + \| \right] + E_{y_D} \left[ \|F(G(y)) - y \right]
\]

Through the effective combination of CNN model and identity-mapping loss, children's emotional speech can be effectively transferred to another emotional expression, so as to achieve the generation of emotional speech.

4. Experiments and Relevant Results

In this paper, we test the effect of the affective speech generation model by using the data collected from the speech corpus of children's reading aloud. The researchers recruited a total of 400 volunteers (5-12 years old, average age 9 years old, 50% male and 50% female respectively). After reading the designated text, each volunteer read the content aloud in a quiet environment, and their multiple speech signals were recorded in time. At the same time, six broadcasting experts were invited to rate the volunteers after listening to their speeches. The score result is an integer within the range of [0, 5]. When 70% of the experts have the same score, the score information of this sample shall be kept. Otherwise, the experts are required to repeat the score of this data in the second round and discard the sample data directly in the third round.

At present, there are more than 8000 effective speeches in this language database, including four kinds of emotions: happy, angry, neutral and sad. The amount of each kind of emotional data is more balanced. For this data set, a five-fold cross validation method is used. 80% of the data is used to train the deep neural network, and the remaining data is used to verify and test the accuracy. When preprocessing the speech, the size of the standard window is 25ms and the offset is 10ms. Features are normalized to zero mean.
In the CNN model, the batch size is 100, and the epochs is 100000. The learning rate is set to 0.001. Dropout is 0.5. Glus is used as an activation function, Adam is used as an optimizer, and mean square error is used as a loss function. In order to test the effectiveness of the generated speech emotion, we build a speech emotion recognition model, and use the newly generated speech to test, and select the external genetic minimal acoustic parameter set (eGeMAPS)\cite{12} as the original feature set. Among them, Bidirectional Long Short-Term Memory (Bi-LSTM) model is used, the batch, epochs, learning rate, dropout and optimizer are the same as CNN model. ReLU is used as the loss function. Weighted accuracy (WA) and Unweighted accuracy (UA) are used as indicators to test the generated emotion and verify the effectiveness of the generated model.

In the experiment, the original speech is used as the baseline, and the following comparison models are designed. Model 1 uses HMM model, model 2 uses conditional GAN model, and model 3 uses current model, CycleGAN. The similarity between the generated speech and the target emotion in the three models is verified. Table 1 shows the accuracy of the generated speech emotion recognition after experimental verification.

| Model     | Model details            | WA  | UA  |
|-----------|--------------------------|-----|-----|
| BaseLine  | the original speech +LSTM| 75% | 73% |
| Model 1   | HMM Model +LSTM          | 65% | 64% |
| Model 2   | conditional GAN+LSTM     | 72% | 71% |
| Model 3   | CycleGAN+LSTM            | 76% | 74% |

From table 1, we can see that the emotional classification performance of the model proposed in this paper is the best among all models, with the optimal accuracy, and the recognition rate of the current model is closest to the baseline. The second is conditional GAN model, and the accuracy of HMM model is the lowest. It can be concluded that the speech generated by this model can effectively express the specified emotion.

Based on the emotion category judgment of the generated emotion speech, the confusion matrix of each emotion category is calculated, and the results are shown in Table 2. The results show that the accuracy of generating model is higher for emotions with higher arousal, such as happy and angry, while the accuracy of generating model is lower for those with lower arousal, such as neutral and sad.

| Accuracy | angry | happy | sad  | neutral |
|----------|-------|-------|------|---------|
| angry    | 78%   | 11%   | 5%   | 6%      |
| happy    | 11%   | 77%   | 5%   | 7%      |
| sad      | 5%    | 5%    | 72%  | 18%     |
| neutral  | 6%    | 7%    | 18%  | 69%     |

5. Conclusion

This paper introduces the principle of CycleGAN and trains CycleGAN to synthesize emotional children's speech. Through the improved CNN model and the effective combination of identity-mapping loss, children's reading speech can be effectively transferred to another emotional expression, and the generated speech is compared with the original voice signal. Using Tensorflow as the learning framework, a large number of experiments show that the emotional speech generation scheme proposed in this paper is generally better than the ordinary speech generation model.

In the future work, the author will continue to optimize and improve the emotional speech generation model, set up a large number of multi-age emotional database, and provide a reasonable human-computer interaction mode to improve the accuracy of human-computer dialogue and user's experience.
Acknowledgment
This paper was funded by the "Dalian Key Laboratory for the Application of Big Data and Data Science" and the Natural Science Foundation of Liaoning Province, Research on Emotional Analysis and Evaluation Model of Speech Reading Based on Machine Learning (20180551068).

References
[1] Andrew J. Hunt, Alan W. Black. UNIT SELECTION IN A CONCATENATIVE SPEECH SYNTHESIS SYSTEM USING A LARGE SPEECH DATABASE[C]// 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings. IEEE, 2002.
[2] Hieu-Thi Luong, Junichi Yamagishi. Multimodal Speech Synthesis Architecture for Unsupervised Speaker Adaptation[C]// Interspeech 2018. 2018.
[3] Jing-Xuan Zhang, Zhen-Hua Ling, Li-Rong Dai. FORWARD ATTENTION IN SEQUENCE-TO-SEQUENCE ACOUSTIC MODELING FOR SPEECH SYNTHESIS[C]// ICASSP 2018 - 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018.
[4] R. Pradeep, M. Kiran Reddy, K. Sreenivasa Rao. LSTM-Based Robust Voicing Decision Applied to DNN-Based Speech Synthesis[J]. Automatic Control and Computer Sciences, 2019, 53(4):328-332.
[5] Lin Jiang, Ping Tan, Junfeng Yang. Speech emotion recognition using emotion perception spectral feature[J]. Concurrency and Computation Practice and Experience, 2019(11):e5427.
[6] Kalantari M, Gooran S.R, Kanan H.R. Improved embedded pre-whitening subspace approach for enhancing speech contaminated by colored noise [J]. Speech Communication, 2018, 99:12-26.
[7] Zhou H, Huang M, Zhang T,et al.Emotional chatting machine:Emotional conversation generation with internal and external memory[C]//Thirty-Second AAAI Conference on Artificial Intelligence.2018.
[8] Ghosh S,Chollet M,Laksana E,et al.Affect-lm:A neural language model for customizable affective text generation[J]. arXiv preprint arXiv:1704. 06851, 2017.
[9] Donahue C, Meauley J, Puckette M. Adversarial Audio Synthesis[J]. 2018.
[10] Kaneko T, Kameoka H, Hojo N, et al. Generative adversarial network-based postfilter for statistical parametric speech synthesis[C]// ICASSP 2017-2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2017.
[11] Kaneko T, Kameoka H. Parallel-Data-Free Voice Conversion Using Cycle-Consistent Adversarial Networks[J]. 2017.
[12] Eyben F, Scherer K R, Truong K P, et al. The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing[J]. IEEE Transactions on Affective Computing, 2017, 7(2):190-202.