2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP) | 978-1-7281-6994-1/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISCSLP49672.2021.9362060

Towards Realizing Sign Language to Emotional Speech Conversion by Deep Learning

Weizhe WANG[2], *Hongwu YANG*[1,2,3*]

[1]School of Educational Technology, Northwest Normal University, Lanzhou 730070, China
[2]College of Physics and Electronic Engineering, Northwest Normal University, Lanzhou 730070, China
[3]National and Provincial Joint Engineering Laboratory of Learning Analysis Technology in Online Education, Lanzhou 730070, China

*yanghw@nwnu.edu.cn

Abstract

This paper proposes a framework of sign language to emotional speech conversion based on deep learning to solve communication disorders between people with language barriers and healthy people. We firstly trained a gesture recognition model and a facial expression recognition model by a deep convolutional generative adversarial network (DCGAN). Then we trained an emotional speech acoustic model with a hybrid long short-term memory (LSTM). We select the initials and the finals of Mandarin as the emotional speech synthesis units to train a speaker-independent average voice model (AVM). The speaker adaptation is applied to train a speaker-dependent hybrid LSTM model with one target speaker emotional corpus from AVM. Finally, we combine the gesture recognition model and facial expression recognition model with the emotional speech synthesis model to realize the sign language to emotional speech conversion. The experiments show that the recognition rate of gesture recognition is 93.96%, and the recognition rate of facial expression recognition in the CK+ database is 96.01%. The converted emotional speech not only has high quality but also can accurately express the facial expression.

Index Terms: gesture recognition, facial expression recognition, emotional speech synthesis, deep learning, sign language to speech conversion, people with language barriers

1. Introduction

Because speech communication is direct, fast, and convenient, it plays an essential role in human-computer interaction and human life nowadays. Unfortunately, in China, many people cannot produce voice due to language barriers. The communication between people with language barriers and other people is mainly completed by sign language. However, most healthy people do not understand sign language, which not only hinders the healthy life and development of people with language barriers but also brings certain pressure to society. Therefore, it is crucial for the life of people with language barriers to converting sign language into the corresponding speech [1] to ease their communication with healthy people.

In recent years, gesture recognition and facial expression recognition have achieved excellent results with the development of machine learning technology and big data technology [2][3]. The neural network-based methods have become the mainstream of gesture recognition and facial expression recognition because of their strong classification characteristics and anti-interference ability [4][5]. At the same time, the computer can also synthesize high-quality speech from the given text. In particular, a deep learning-based statistical parametric speech synthesis method is widely used in emotional speech synthesis. Thanks to its ability to learn the relationship between language features and acoustic features, deep learning-based methods can significantly improve the naturalness of the synthesized emotional speech when the number of model parameters is similar [6][7]. However, the current work mainly focuses on gesture recognition, facial expression recognition, and emotional speech synthesis, respectively. This situation not only ignores the critical role of speech information and facial expression information, but the single gesture also cannot accurately convey the emotional color of the content expressed by the people with language barriers. Therefore, this paper uses a deep learning-based method to realize sign language to emotional speech conversion to provide convenience for the learning and life of the people with language barriers and to promote the development of human-computer interaction.

2. Related works

2.1. Deep Convolutional Generative Adversarial Network

Just as the name suggests, Generative Adversarial Network (GAN) is a generative model based on the adversarial method, which can learn the distribution of real data and generate new datasets with high similarity [8]. The basic GAN consists of a generator (G) and a discriminator (D). The function of G is to generate samples as similar as real samples while the D distinguishes the generated fake samples from real samples. G gets the random noise z as input and maps it to $G(z)$, x is a real sample which satisfies $x \sim P(x)$, x and $G(z)$ are input into D together, and D judges x as true and $G(z)$ as false as far as possible. Generator and discriminator strive to improve their generating ability and discriminating ability through this process and finally achieve the Nash equilibrium between generator and discriminator. The basic GAN model can be optimized through the following equation:

$$\min_G \max_D E_{x\sim p(x)}[\log D(x)] + E_{x\sim p(z)}[\log(1 - D(G(z)))] \quad (1)$$

Alec Radford proposed the Deep Convolutional Generative Adversarial Network (DCGAN) [9] by introducing a convolution network into GAN structure to improve the effect of GAN, which has the following characteristics:

- The generator model uses four fractionally stridden convolution to realize the generation process from noise to image, and the discriminator model uses stridden convolutions to replace the pooling layer [10].

- The generator and discriminator use batch normalization to solve the problem of initialization error while maintaining this gradient and propagating to each layer.
- The convolution layer is used to connect the output layer of the generator and the input layer of the discriminator.
- The generator uses the rectified linear unit (ReLU) activation function for all layers except the output layer which uses the activation function, and all layers of the discriminator use the leaky rectified linear unit activation function [11].

2.2. LSTM

The recurrent neural network (RNN) structure has the vanishing gradient problem to model long-term dependencies. The most effective way to solve the vanishing gradient problem is to use the RNN variant LSTM. Gates and memory cells are added to the LSTM structure. This structure allows information to be retained across many time steps. Therefore, LSTM can effectively capture long-term dependencies [12]. In mathematics, the neurons complete the storage and update of the information in the network through the following mathematical recursion about time t.

$$i_t = \boldsymbol{\sigma}(\boldsymbol{W}_{xi}x_t + \boldsymbol{W}_{hi}h_{t-1} + \boldsymbol{W}_{ci}c_{t-1} + b_i) \quad (2)$$

$$f_t = \boldsymbol{\sigma}(\boldsymbol{W}_{xf}x_t + \boldsymbol{W}_{hf}h_{t-1} + \boldsymbol{W}_{cf}c_{t-1} + b_f) \quad (3)$$

$$o_t = \boldsymbol{\sigma}(\boldsymbol{W}_{xo}x_t + \boldsymbol{W}_{ho}h_{t-1} + \boldsymbol{W}_{co}c_{t-1} + b_o) \quad (4)$$

$$c_t = f_t \cdot c_{t-1} + i_t \cdot tanh(\boldsymbol{W}_{xc}x_t + \boldsymbol{W}_{hc}h_{t-1} + b_c) \quad (5)$$

$$h_t = o_t \cdot tanh(c_t) \quad (6)$$

Where i_t, f_t, o_t represents the input gate, forgetting gate and output gate respectively. c_t represent memory cell. σ is the sigmoid activation function, $\boldsymbol{W}$ represents the weight matrix connecting different gates, $x$is an input vector, h_t is an hidden state vector. $tanh(.)$ is the hyperbolic tangent function, and b is the corresponding bias vector.

3. Framework of sign language to emotional speech conversion

The proposed framework of sign language to emotional speech conversion includes a gesture recognition model, a facial expression recognition model, and an emotional speech synthesis model, as shown in Figure 1. In the recognition stage, we transform gestures and facial expressions into corresponding text and emotion labels, respectively. In the emotional speech synthesis stage, we input the context-dependent information corresponding to gesture semantics and emotional labels corresponding to facial expressions into the trained emotional speech synthesis model to generate emotional speech.

3.1. Gesture recognition and facial expression recognition

The gesture recognition and facial expression recognition are modified based on the network structure of conditional DCGAN proposed by Tang [13]. Firstly, the gesture images and facial expression images are converted into gray images to eliminate the influence of color. All gesture images and facial expression images are resized to $64 * 64$ and normalized for reducing the neural network's training time. Then, we connect 100-dimensional noise and image labels data to the input generator and obtain $64 * 64 * 1$ image samples through convolution operation and

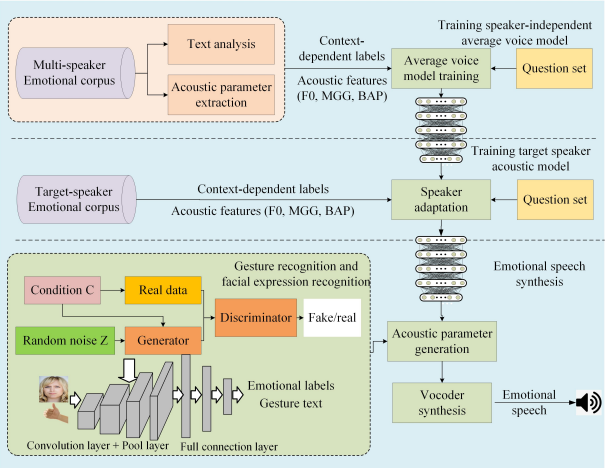

Figure 1: *The framework of sign language to emotional speech conversion.*

pooling operation. The image samples generated by the generator are input into the discriminator together with the real data. We get the optimal weights of each layer of the generator and discriminator through multiple pre-training. Finally, we fine-tune the discriminator parameters as the recognition model for gesture recognition and facial expression recognition to obtain the semantics corresponding to the gesture and the emotional labels corresponding to facial expression.

3.2. Emotional speech synthesis

Emotional speech synthesis is divided into the training stage and the synthesis stage. In the training stage, we extracted the fundamental frequency (F0), the generalized Mel-generalized Cepstral (MGC) coefficient, and the band a periodical (BAP) coefficients from each emotional speech in a multi-speaker emotional corpus. We trained a prosodic model for each emotion by obtaining prosodic labels for the text through text normalization, grammatical analysis, and prosodic analysis. Furthermore, context-dependent labels of different emotional sentences are obtained. We then used context-dependent labels information to train the average voice model (AVM) for different emotional speech. Next, a speaker-dependent target emotional acoustic model is re-trained using a small target emotional speaker corpus based on the speaker-independent AVM. We first used the emotion labels obtained from facial expression recognition to select the corresponding emotion prosodic model and the speaker-dependent emotional acoustic model in the speech synthesis stage. The input text from gesture recognition is labeled by the corresponding emotional prosodic model to generate the context-dependent labels. The context-dependent labels are then fed into the speaker-dependent emotional acoustic model to generate the acoustic parameters. Finally, the WORLD vocoder is used to generate the emotional speech waveform from the acoustic parameters.

4. Experimental data

4.1. Gesture data

Ten students were invited to record 30 different gestures, and each gesture was recorded 20 times. All the collected gesture images are saved in JPG format with the file name correspond-

ing to the gesture. In the experiment, 5000 gesture images were randomly selected as the training set and the rest of the data as the testing set.

4.2. Facial expression data

We use the CK+[14] and Jaffe[15] expression database as the facial expression database. We extracted 800 facial expressions including happiness, sadness, surprise and neutral. At the same time, we enhance all the data by rotation to obtain more 3200 facial expression data. We randomly select 80% of the data on each facial expression for the training, and the remaining images are used for testing.

4.3. Emotional speech data

Firstly, we designed a text corpus selected from the daily conversation in Mandarin for four emotions in which each emotion has 2000 sentences. Then, we collect each emotion's corresponding video to create a specific scene to stimulate the speaker's emotion for emotional speech recording. We invited ten female Mandarin speakers to record according to the designed text corpus, and each speaker recorded 200 utterances of each emotional speech. In the experiment, all speech data are sampled at 16KHz, quantized at 16 bits, stored in mono WAV format. 7000 utterances are used for model training, and the rest is used for model testing.

5. Experiments set-up and experiments results

5.1. Experiments set-up

The generator network consists of two fully connected layers, one dropout layer, and eight deconvolutional layers. The dropout layer is between two fully connected layers, and the dropout rate is 0.05. The ReLU activation function was applied in each layer, except for the last layer. The last layer using a linear activation function. The structure of discriminator and classifier is similar. The classifier consists of 12 deep neural layers, in which the first nine layers are convolutional layers, and the last three layers are fully connected layers. The classifier uses Leaky ReLU as the activation function in each layer, except the output layer. The output layer using a sigmoid activation function. Each layer of generator and classifier uses batch normalization to batch normalize the input of the hidden layer. Adam is used to optimizing the loss function. The learning rate is 0.0002.

In the experiment of emotional speech synthesis, the hybrid LSTM-based framework consists of a single LSTM layer with 256 memory blocks and four hidden DNN layers with 512 units. The LSTM hidden layer follows the last hidden layer of DNN. We use the backpropagation through time algorithm to initialize the model parameters and optimize them. We use the mini-batch stochastic gradient descent algorithm to train the emotional speech acoustic model, and the mini-batch size is 256. The ReLU activation function was applied in each layer, except for the output layer. The output layer uses a linear activation function with mean square error (MSE) as the loss function. If the validation MSE did not improve within ten epochs, the training stopped. The learning rate was 0.004 for the first 20 epochs and then halved for the remaining epochs.

Table 1: *Gesture recognition results under different methods.*

Methods	**Recognition rate (%)**
DNN	90.57
CNN	92.82
Our method	93.96

Table 2: *Confusion matrix for the mixed dataset (%).*

	Happiness	**Sadness**	**Surprise**	**Neutral**
Happiness	96.85	0.30	1.28	0.91
Sadness	1.37	95.27	0.84	2.52
Surprise	1.26	0.50	97.39	0.85
Neutral	1.73	0.83	1.39	96.05

5.2. Experiments results

5.2.1. Gesture recognition and facial expression recognition

We use different experiments to evaluate the proposed method's effect on gesture recognition and facial expression recognition. In gesture recognition, the proposed method is compared with the DNN-based and CNN-based methods. The results of gesture recognition under different methods are shown in Table 1. We can see that because the DCGAN method is better able to extract gesture features, it is more suitable for gesture classification than the DNN-based method and CNN-based method.

In the facial expression recognition experiment, we randomly selected 20 expressions from the test set for testing, and calculate the confusion matrix of recognition corresponding to each expression, as shown in Table 2. Also, four facial expressions are recognized on CK+ database and compared the recognition results with other methods using the same database, as shown in Table 3.

5.2.2. Emotional speech synthesis

To evaluate the quality of the synthesized emotional speech, we use the emotional corpus of 9 female speakers to train the average voice model and the emotional corpus of 1 female speaker to speaker adaptation. Each emotion uses 800 emotion speech to train the average voice model, and 200 emotion speech to train the target speaker adaptation, of which 160 are used as the training set, and 40 are used as the testing set.

Objective evaluations: In the objective evaluation, we evaluate the quality of the synthesized emotional speech by calculating the distortions between the original speeches' acoustic parameters and the the synthesized speeches' acoustic parameters. The evaluated acoustic parameters include RMSE of F0, MCD, BAPD and V/UV. The objective evaluation results are shown in Table 4.

Subjective evaluations: We invited 30 native Mandarin subjects to randomly select 20 sentences from the test set for subjective evaluation. We use the emotional mean opinion score (EMOS) test, and degradation mean opinion score (EDMOS) test to evaluate the quality of synthesized emotional speech. In the EMOS test, subjects were asked to use a 5-point scale score to rate the synthesized emotional speech's naturalness. A 5-points indicates that synthesized emotional speech's naturalness is perfect, while 1-point indicates that the naturalness of synthesized emotional speech is very poor. The EMOS of the synthesized emotional speech is shown in Figure 2.

Table 3: *The recognition rate (%) of different methods on CK+.*

Methods	Recognition rate (%)
CNN[16]	95.94
LeNet-5[17]	82.62
GCNET[18]	97.93
WGAN[19]	96.00
Our method	96.01

Table 4: *Objective evaluation results of synthesized emotional speech.*

Type	MCD (dB)	BAP (dB)	F0 RMSE(Hz)	V/UV (%)
Happiness	6.754	0.197	23.791	8.001
Sadness	6.992	0.209	25.376	8.227
Surprise	7.261	0.189	25.187	8.971
Neutral	6.857	0.209	26.783	8.201

In the EDMOS test, the synthesized emotional speech and their corresponding original recording formed a speech file pair. We randomly played each pair of speech files to the subjects according to the original recording order before the synthesized emotional speech. We asked the subjects to compare these two speeches carefully and evaluate the degree of similarity between the synthesized emotional speech and the original recording within a 5-point scale score. A 1-point indicates that the synthesized emotional speech is quite different from the original recording, while 5-point indicates that the synthesized emotional speech is very close to the original recording. The EDMOS of emotional speech is shown in Figure 3.

5.2.3. Evaluation of sign language to emotional speech conversion

Based on the predefined gestures and facial emotions, this paper designs emotional speech corresponding to 60 words and sentences. Then, the emotional speech was played to 20 Mandarin reviewers. The evaluators are asked to listen carefully to the synthesized emotional speech, and select the corresponding text and emotion from the four options A, B, C, and D according to the emotional speech heard. Finally, the evaluation results are compared with the standard answer to calculate the correct rate of sign language to emotional speech conversion. The evaluation results are shown in Table 5. We can see from Table 5 that the correct conversion rate of sentences is higher than the words. Because we evaluate the emotional speech, it is hard to judge the speaker's emotion from a single word. The sentence can express complete emotion so that we can easily judge the speaker's emotion from the speech.

Table 5: *Evaluation results of sign language to emotional speech conversion.*

Type of evaluation	Words	Sentences
Correct rate (%)	76.82	93.91

Figure 2: *The average EMOS score of synthesized emotional speech under 95% confidence intervals.*

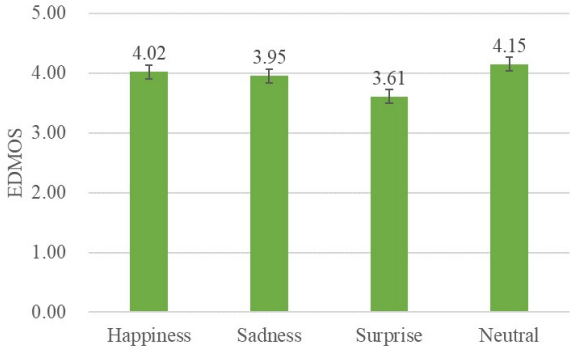

Figure 3: *The average EDMOS score of synthesized emotional speech under 95% confidence intervals.*

6. Conclusions

This paper further our previous work to realize a deep learning-based sign language to emotional speech conversion. The gesture recognition and facial expression recognition are realized by using the DCGAN. We also adopted the hybrid LSTM to improve the speech quality of the synthesized emotional speech. Finally, we combine these two networks to achieve the conversion from sign language to emotional speech. Experimental results show that the proposed recognition framework achieves good recognition in gesture recognition and facial expression recognition. The objective and subjective evaluation demonstrated that synthesized emotional speech has high quality. Besides, the conversion of sign language to emotional speech not only has high accuracy but also can accurately convey the facial expression. Therefore, our research can improve people's quality of life with language barriers and promote the development of people with language barriers. Future work focuses on studying the dynamic sign language to emotional speech conversion to improve the real-time performance and expressiveness of sign language to emotional speech conversion.

7. Acknowledgements

The research leading to these results was partly funded by the National Natural Science Foundation of China (Grant No.11664036 and No.31860285). Additionally, part of this work was performed in the High School Science and Technology Innovation Team Project of Gansu (Grant No.2017C-03).

8. References

[1] E. A. Kalsh, and N. S. Garewal, "Sign language recognition system," *International Journal of Computational Engineering Research*, vol. 3, no. 6, pp. 15–21, 2013.

[2] D. Rabiner, L. P. Gou, and P. J. Kindermans, et al, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1-1.

[3] M. D. Ding, and L. Lin, "CNN and HOG Dual-Path Feature Fusion for Face Expression Recognition," *Information and control*, vol. 49, no. 1, pp. 47–54, 2020.

[4] C. C. D. Santos, J. L. A. Vassallo and R. F. Vassallo, "Dynamic Gesture Recognition by Using CNNs and Star RGB: a Temporal Information Condensation," *Neurocomputing*, vol. 400, pp. 238–254, 2020.

[5] H. Zheng, R. L. Wang and W. T. Ji et al, "Discriminative deep multi-task learning for facial expression recognition," *Information Sciences*, vol. 533, pp. 60–71, 2020.

[6] S. M. An, Z. H. Ling and L. R. Dai, " Emotional statistical parametric speech synthesis using LSTM-RNNs," *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017.

[7] T. J. Lorenzo, G. E. Henter and S. Takaki et al, "Investigating different representations for modeling and controlling multiple emotions in dnn-based speech synthesis," *Speech Communication*, vol. 99, pp. 135–143, 2018.

[8] I. J. Goodfellow, J. Pouget-Abadie and M. Mirza, et al, "Generative Adversarial Networks," *Advances in Neural Information Processing Systems*, vol. 3, pp. 2672–2680, 2014.

[9] A. Radford, L. Metz and S. Chintala, "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks," *Computer Ence*, 2015.

[10] X. L. Tang, Y. M. Du and Y. W. Liu, et al, "Image Recognition With Conditional Deep Convolutional Generative Adversarial Networks," *Acta Automatica Sinica*, vol. 44, no. 5, pp. 855–864, 2018.

[11] M. Simon, E. Rodner and J. Denzler, "ImageNet pre-trained models with batch normalization," 2016.

[12] S. Hochreiter, J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] S. W. Liu, C. W. Liu and A. L. Zhang, "Real-time facial expression and gender recognition based on depthwise separable convolutional neural network," *Journal of Computer Applications*, vol. 40, no. 4, pp. 990–995, 2020.

[14] P. Lucey, J. F. Cohn and T. Kanade, et al, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops. IEEE*, pp. 94–101, 2010.

[15] M. J. Lyons, "Facial Expression Database : Japanese Female Facial Expression (JAFFE) Database,"

[16] Z. B. Meng, P. Liu and J. Cai, ea al, "Identity-Aware Convolutional Neural Network for Facial Expression Recognition," *IEEE Computer Society*, 2017.

[17] Y. Lin, X. Z. Lin and M. Y. Jiang, "Facial Expression Recognition with Cross-connect LeNet-5 Network," *Acta Automatica Sinica*, vol. 44, no. 1, pp. 176–182, 2018.

[18] Y. Kim, B. I. Yoo and Y. Kwak, et al, "Deep generative-contrastive networks for facial expression recognition," 2017.

[19] N. M. Yao, Q. P. Guo and F. C. Qiao, et al, "Robust Facial Expression Recognition with Generative Adversarial Networks," *Acta Automatica Sinica*, vol. 44, no. 5, pp. 865–877, 2016.

[20] X. M. Li, X. L. Fu and G. F. Deng, "Preliminary application of the abbreviated PAD emotion scale to Chinese undergraduates," *Chinese Mental Health Journal*, vol. 22, no. 5, pp. 327–329, 2008.