Speech2expression:
Change the Expression Behind a Voice

Xiaolin Deng, Data Science and AI for the Creative Industries

Project Rationale

The human voice contains more information than we can imagine, your age, height, weight, spirit, mood, even emotion, temperament, drive, creativity, and so on. And in the virtual world nowadays, the voice gives an additional personality label to a person, giving another imaginary form of expression to others. In this project plan, I would like to go through the classification of emotions extracted from speech and try to make the face of a face image change accordingly due to the speech under different emotions, which to explore the unknown effects of diverse voices on the generation of facial expressions.

Project Output

Speech2expression: A model first inputs any picture and speech and can then generate faces that change expressions because of the speech emotion, based on the emotion represented by the speech.

Datasets: MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation[1], a talking-face video corpus featuring 60 actors and actresses talking with eight different emotions at three different intensity levels. In order to generate more accurate results, for this project, I have selected only some of the datasets with the five more variable mood categories of "angry", "neutral", "fear", "happy" and "sad" as the generation and testing datasets. "happy", and "sad", which are five highly variable mood categories, were used as the generation and test dataset.

Methodologies

This project consists of two aspects: Voice recognition, Facial expression generation

- **Voice recognition**

  In this module, it is important to avoid the effects of personalized speech, e.g., when Person_A is angry and when Person_B is angry, they are different in timbre but have high similarity in pitch and degree of audio variation. I use Librosa's MFCC method to extract high-dimensional MFCC information from voice (Speech-Emotion-Analyzer[2])and use a Linear layer to realize audio classification features. This model structure limits the length of each voice input to 3s.

---

[1] wywu.github.io. (n.d.). MEAD: A Large-scale Audio-visual Dataset for Emotional Talking-face Generation. [online] Available at: https://wywu.github.io/projects/MEAD/MEAD.html.

[2] DimensionNXG. “Inference on Pre-Recorded Audio Samples.” GitHub,22Feb.2022,github.com/DimensionNXG/Speech-Emotion-Analyzer. Accessed 25 Mar. 2022.

At the same time, I have chosen the same words of being said, the same level of emotion, different emotional features to do generation and test work, making me only focus on tone of voice.

In the end, I implemented the voice in the input dataset and then generated the corresponding emotion labels.

- **Facial expression generation**

  In this module I have tried two approaches to facial emotion modification：

  1. pix2pixHD： (The detailed files are in the "pix2pix_approach" folder)

  a) First, all video frames were extracted as images and stored in folder corresponding to emotions.(propressed_dataset.py)

  b) Set up five corresponding signs based on the five emotions.(draw_sign_image.py)

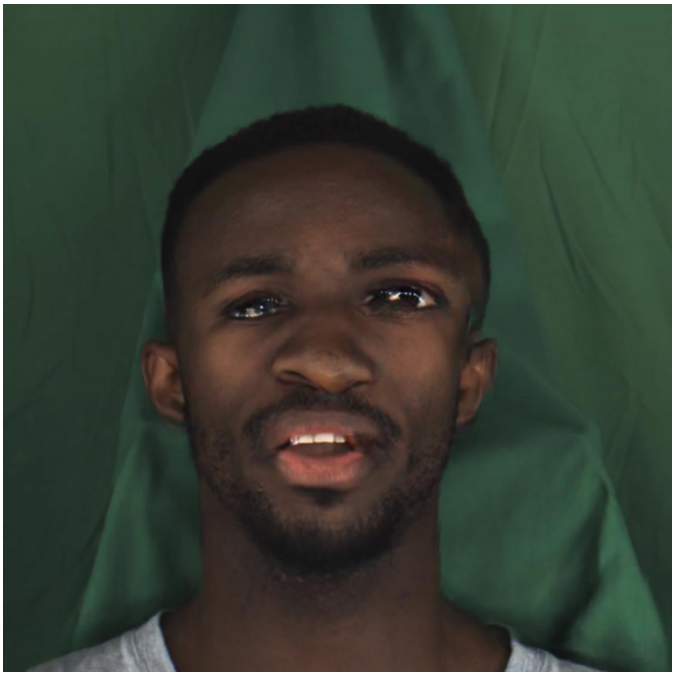
  c) The generated labels were matched to the signs according to "audio recognition" and drawn on the corresponding emotion images by CV.(utils.py&infer.py(pix2pix approach_version))

  d) Matched signed images with the original images to form a pix2pix dataset.

  e) Data augmentation of datasets (pix2pix_approach/pix2pixhd /draw_sign_image.py)

  f) Implementing pix2pix model to complete face generation

  But the result was very unsatisfactory：



Firgure 1: Generated images for the preliminary training of pix2pix model.

In practice, pix2pix has a massive amount of data preparation and a long data preparation process. It is difficult for pix2pix to have as many paired datasets for training, and its robustness does not meet my current needs. Therefore it cannot be generalised to arbitrary characters.

So, in subsequent work, I referred to methods in the IALS framework to generate.

2. IALS： (The detailed files are in the "IALS-main" folder)

This framework (IALS) performs Instance- Aware Latent-Space Search to find semantic directions for disentangled attribute editing.[3] It can search for semantic directions inside the parametric hidden space of the GAN and can transform face characteristics by adjusting the features of each part of the face through the input.

a) Processed the input face image and aligned it to FFHQ face format (folder: static_files)
b) In this approach, there is no need to extract frame images from the video; instead, convert the .mp4 video in the original dataset to .wav format and then execute the speech recognition module to generate emotion labels (Xiaolin Deng-AI-Speech2expression/extract_audio.py）
c) Imported the values of the parameters corresponding to the generated sentiment labels into the IALS framework, matched smile and young condition attributes for primal attribute editing.(Xiaolin Deng-AI-Speech2expression/infer.py)
d) Repeatedly adjust the weight parameters to fit and generate facial expression images.

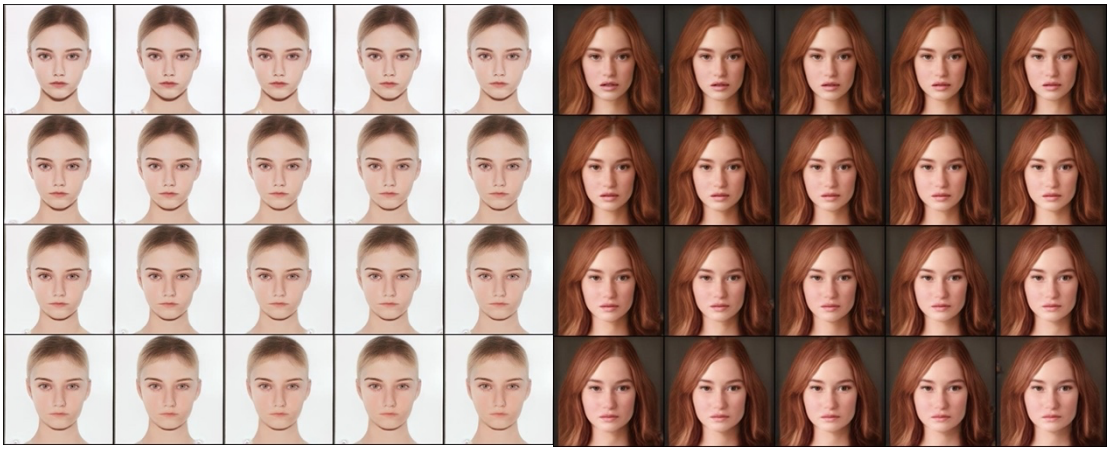
Discussion

I ended up choosing 11 photos of faces. Same voice content, with highest emotional level, five emotional features, and different people's voices, generates facial images that change expressions according to the voice.Finally, it is stored in the "IALS-main/image_output" path and categorised by the five emotions.

Firgure 2: The "happy" expression facial image was generated by inputting audio with a "happy" emotion label.

[3] Han, Yuxuan, et al. Disentangled Face Attribute Editing via Instance-Aware Latent Space Search. May 27AD.

As shown in Figure 2, The model has been able to apply the "happy" weight parameter to the input face image well, with continuous attribute variation towards the target label. The final expression of "Happy" is relatively natural, but there is also variation in skin tone.



Firgure 3&4: The "happy" expression facial image was generated by inputting audio with a "happy" emotion label.

However, In Figure 3,We can see in the picture that the face complexion has changed from fair to rosy, but the expression remains cold, and the effect of happiness is not obvious. Also, in Figure 4, the woman's mouth had shifted from open to closed, which is more like a shift in emotion from surprise to calm, without a hint of happiness.

If we speculate from this, it is possible that the model's learning of the various emotions of the face is a variety of facial features, including the mouth, eyes, and face colour; the effect is also related to the initial emotion of the input face.

Firgure 5: The "angry" expression facial image was generated by inputting audio with a "angry" emotion label.

As can be seen in figure 5, Apart from the change in facial expressions, it is clear that the picture's overall tone has also changed. It shows that the model learns hue change as well.

Future Evaluation

In future work, if the model is to be evaluated further, I plan to use the machine index and manual evaluation index to evaluate the results comprehensively:

1. Using a facial expression recognition classification model for evaluation, refer to Expression-Recognition[4]. For example, after generating a face image with a changed expression through certain emotion voice, the resulting face image is passed through this emotion classification model to identify whether the image also carries the same emotion.
2. Testers are invited to infer emotions by listening to speech unknowingly and evaluate the veracity of the generated faces to form experimental data to verify the feasibility of Speech2expression in practical applications.

Reference List

1. DimensionNXG. "Inference on Pre-Recorded Audio Samples." *GitHub*, 22 Feb. 2022, github.com/DimensionNXG/Speech-Emotion-Analyzer. Accessed 25 Mar. 2022.
2. Duarte, Amanda, et al. "SPEECH-CONDITIONED FACE GENERATION USING GENERATIVE ADVERSARIAL NETWORKS." *GitHub*, 1 Jan. 2022, github.com/imatge-upc/wav2pix.
3. Eskimez, Sefik Emre, et al. "Speech Driven Talking Face Generation from a Single Image and an Emotion Condition." *Arxiv.org*, 8 Aug. 2020, arxiv.org/abs/2008.03592, 10.48550/arXiv.2008.03592. Accessed 25 Mar. 2022.
4. genforce. "InterFaceGAN - Interpreting the Latent Space of GANs for Semantic Face Editing." *GitHub*, 24 Mar. 2022, github.com/genforce/interfacegan. Accessed 25 Mar. 2022.
5. Han, Yuxuan, et al. *Disentangled Face Attribute Editing via Instance-Aware Latent Space Search*. May 27AD.
6. junleen. "Toward Fine-Grained Facial Expression Manipulation (ECCV 2020, Paper)." *GitHub*, 17 Mar. 2022, github.com/junleen/Expression-manipulator. Accessed 25 Mar. 2022.

[4] WuJie. "Facial-Expression-Recognition.Pytorch." GitHub, 25 Mar. 2022, github.com/WuJie1010/Facial-Expression-Recognition.Pytorch. Accessed 25 Mar. 2022.

7. Ling, Jun, et al. “Papers with Code - toward Fine-Grained Facial Expression Manipulation.” *Paperswithcode.com*, 4 Dec. 2020, paperswithcode.com/paper/toward-fine-grained-facial-expression. Accessed 25 Mar. 2022.
8. seeprettyface.com. “人脸属性编辑器（赠品）.” *GitHub*, 24 Mar. 2022, github.com/a312863063/seeprettyface-face_editor. Accessed 25 Mar. 2022.
9. Shankar, Shrivu. “Inverse Style GAN.” *GitHub*, 11 Mar. 2022, github.com/sshh12/Inverse-Style-GAN. Accessed 25 Mar. 2022.
10. Shen, Yujun, et al. *Interpreting the Latent Space of GANs for Semantic Face Editing Original Pose Age Gender Eyeglasses*. Mar. 31AD.
11. Su, Bo-Hao, and Chi-Chun Lee. “A Conditional Cycle Emotion Gan for Cross Corpus Speech Emotion Recognition.” *2021 IEEE Spoken Language Technology Workshop (SLT)*, 19 Jan. 2021, biic.ee.nthu.edu.tw/archive/doc/research/A%20Conditional%20Cycle%20Emotion%20Gan%20for%20Cross%20Corpus%20Speech%20Emotion%20Recognition.pdf, 10.1109/slt48900.2021.9383512. Accessed 25 Mar. 2022.
12. Tang, Hao, et al. “Expression Conditional GAN for Facial Expression-To-Expression Translation.” *Www.arxiv-Vanity.com*, 14 May 2019, www.arxiv-vanity.com/papers/1905.05416/. Accessed 25 Mar. 2022.
13. WuJie. “Facial-Expression-Recognition.Pytorch.” *GitHub*, 25 Mar. 2022, github.com/WuJie1010/Facial-Expression-Recognition.Pytorch. Accessed 25 Mar. 2022.