Generation of Colored Subtitle Images Based on Emotional Information of Speech Utterances

1st Ryoichi Takashima
Graduate School of System Informatics
Kobe University
Kobe, Japan
rtakashima@port.kobe-u.ac.jp

2nd Fumiya Nakamura
Graduate School of System Informatics
Kobe University
Kobe, Japan

3rd Ryo Aihara
Information Technology R&D Center
Mitsubishi Electric Corporation
Kamakura, Japan

4th Tetsuya Takiguchi
Graduate School of System Informatics
Kobe University
Kobe, Japan

5th Yusuke Itani
Information Technology R&D Center
Mitsubishi Electric Corporation
Kamakura, Japan

***Abstract*—Conventional automatic subtitle generation systems based on speech recognition do not consider paralinguistic information such as emotions included in speech. In this paper, we propose a method to generate subtitles that visualize the speaker's emotions through the fonts and colors that are used in the subtitles. The proposed method uses a GAN-based image transformation model conditioned by the score of each emotion category to generate fonts that can represent complex emotions that are a mixture of multiple emotions. For colorizing the generated font images, we consider the emotion in the two-dimensional space of valence and activation based on the Russell's circumplex model and define a color map in the emotion space based on Plutchik's wheel of emotions. We confirmed the effectiveness of the proposed method by conducting subjective evaluation experiments on the emotions felt from the generated colored font images.**

***Index Terms*—subtitle image generation, speech emotion recognition, speech recognition, image style transfer**

I. Introduction

In recent years, the accuracy of automatic speech recognition (ASR) using neural networks has dramatically improved with the development of deep learning technology. Automatic subtitle generation for video content and video calls using speech recognition is especially useful for viewers from different languages and those with hearing difficulties, and it is already being put into practice. However, although we can understand what the speaker is saying by simply looking at a transcribed text, it is difficult to understand the speaker's emotions.

Humans use various types of information other than linguistic information in an integrated manner to understand the content of speech. In particular, emotional information inferred from a speaker's facial expression and tone of voice is known to play an important role in facilitating human communication [1]. In video content, such as TV shows and YouTube videos, subtitles in various fonts are often used to express the situation and the emotions of the performers. Therefore, automatically generating subtitles that change font and color according to the speaker's emotions will provide the viewer with richer information about the speech utterance.

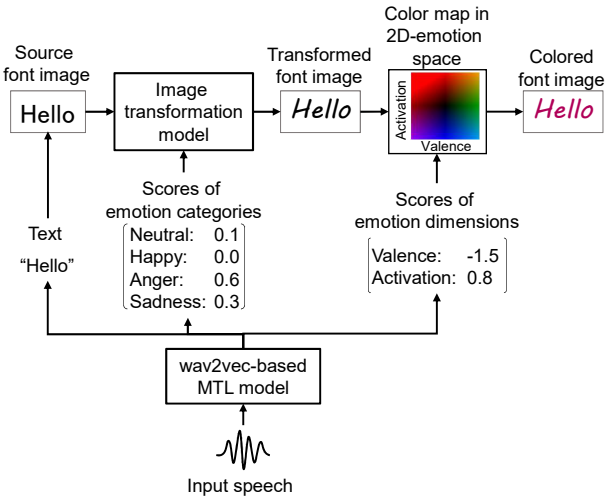

Fig. 1. Overview of the proposed colored subtitle image generation system.

One possible way to realize emotion-reflective subtitle generation is a rule-based method [2]. In this method, the font to be used is predefined for each emotion category (e.g. anger, sadness, happiness, etc). Then, speech emotion recognition (SER) is performed in addition to ASR, and the font is selected depending on the emotion category estimated by the SER. However, this method can only express predefined categorical emotions and cannot express complex emotions that are a mixture of multiple emotions (e.g. anger tinged with sadness).

In this paper, we propose a method to generate subtitles that expresses the speaker's mixed complex emotions through the fonts and colors that are used in the subtitles. Fig. 1 shows the overview of our system. For generating fonts, instead of rule-based font selection, the proposed method uses a GAN-based image transformation model conditioned by the score of each emotion category. For colorizing the generated font images, we

consider the emotion in the two-dimensional space of valence and activation (we refer to them as *emotion dimensions*) based on the Russell's circumplex model [3] and define a color map in the emotion space based on Plutchik's wheel of emotions [4]. In order to estimate the text, the scores of the emotion categories, and the scores on valence–activation axes from a speech utterance, we apply multi-task learning (MTL) based on wav2vec 2.0 model [5]. We confirmed the effectiveness of the proposed method by conducting subjective evaluation experiments on the emotions felt from the generated colored font images.

II. Proposed method

A. Overview

In our proposed system, a colored subtitle is generated from a speech utterance according to the following flow:

1) Speech recognition, emotion category recognition, and emotion dimension estimation are performed using a wav2vec 2.0-based MTL model to obtain speech content text and emotion information.
2) Based on the text of the speech content, a subtitle image is generated in the source font.
3) The subtitle image in the source font and the score of each emotion category are input to an image transformation model, that generates a subtitle image in the font that reflects the emotion scores.
4) By using the estimated scores of emotion dimensions, the color of the subtitle image is changed according to the color map, which we define in the emotion dimension space based on Plutchik's wheel of emotions.

B. Multi-task learning of speech recognition, emotion category recognition, and emotion dimension estimation with wav2vec 2.0

Recent studies on automatic speech recognition have shown remarkable results by using self-supervised learning models such as wav2vec 2.0 [5] and HuBERT [6], which are pre-trained on a large amount of unlabeled speech, followed by fine-tuning with a small amount of labeled speech. The self-supervised learning models have also been shown to be effective in tasks other than speech recognition, such as speech emotion recognition, speaker identification, and spoken language understanding [7]–[9]. In the related study [10], when fine-tuning the wav2vec 2.0 model, a multi-task learning (MTL) of speech recognition and emotion category recognition was proposed. By using the MTL, the speech recognition and emotion category recognition can be performed using one model. In our system, we apply the MTL of the emotion dimension estimation in addition to the speech recognition and the emotion category recognition tasks.

The proposed MTL model is shown in Fig 2. It consists of the pre-trained wav2vec2.0 model with an additional fully-connected (FC) layer for speech recognition, a global average pooling layer in the time direction and a FC layer for emotion category recognition, and a global average pooling layer and a FC layer for emotion dimension estimation. The IEMOCAP [11] dataset that we use in the experiments contains ground truth of the speech content text, emotion category, and scores of emotion dimensions for each speech utterance. Following the related work [10], the CTC loss L_{CTC} and the cross-entropy (CE) loss L_{CE} are used for speech recognition and emotion category recognition, respectively. For emotion dimension estimation, we use the mean square error (MSE) L_{MSE} between the estimated scores of emotion dimensions and the ground truth of the scores as the loss function. Therefore, the overall model loss L_{MTL} is defined as the weighted sum of the three loss functions:

$$L_{\mathrm{MTL}} = L_{\mathrm{CE}} + \alpha L_{\mathrm{CTC}} + \beta L_{\mathrm{MSE}}, \tag{1}$$

where α and β are the weights of CTC loss and MSE loss in multi-task learning, respectively, and $\alpha = \beta = 0.1$ in this study.

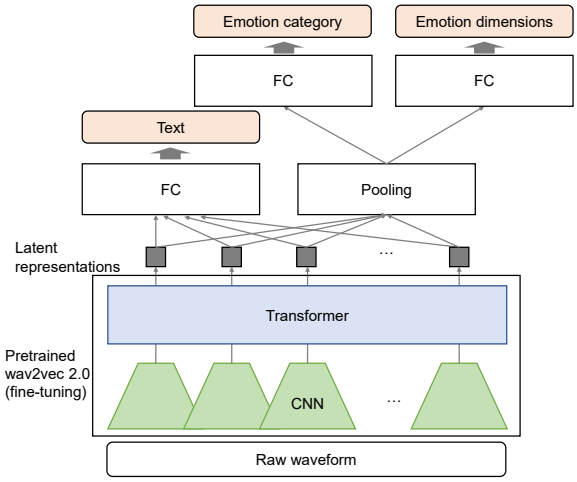

Fig. 2. Structure of multi-task learning model for speech recognition, emotion category recognition, and emotion dimensions estimation.

C. Font image transformation model

Approaches that use generative adversarial network (GAN), such as zi2zi [12] and FontGAN [13], have been studied for font image generation. zi2zi is an extension of pix2pix [14], an image transformation model based on GAN, applied to font image generation. The model structure of zi2zi is shown in Fig. 3. Unlike pix2pix, zi2zi handles the one-to-many task of generating multiple target fonts from a single source font. In addition to the encoder-decoder model of pix2pix's U-net structure [15], a category embedding obtained by passing the font label (ID) through an embedding layer is combined with the encoder output and is input to the decoder. In the generator, constant loss [16] L_{const} is added so that the encoder learns to map the same character in different fonts to a close vector. In addition, the discriminator adds a FC layer for classifying which font an image is, and introduces a category loss [17] L_{category} to prevent the generation of images that differ from any of the fonts used in training.

zi2zi can generate a font that interpolates two fonts by linearly interpolating the category embedding of each learned

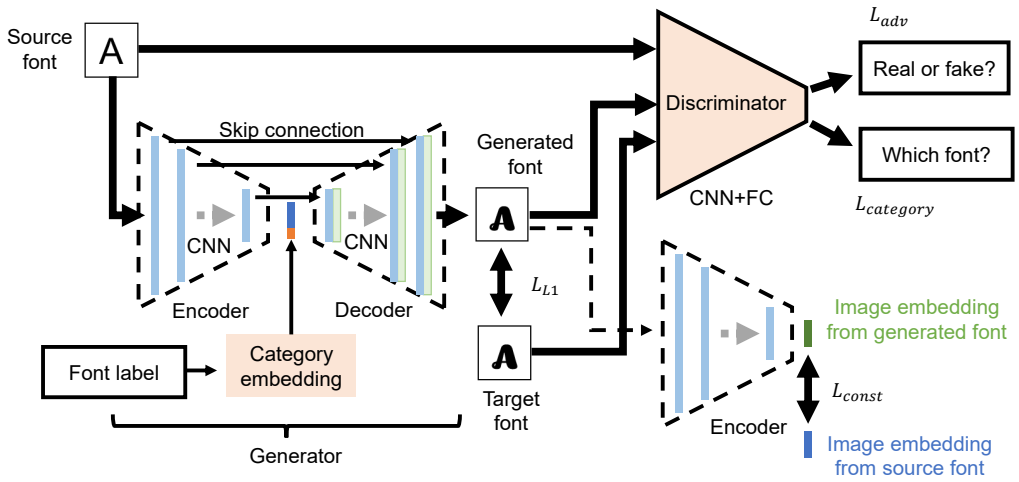

Fig. 3. Model structure of zi2zi

font and combining it with the encoder output. In our font transformation model, we define the font for each emotion, and the font of emotions that are a mixture of multiple emotions is generated by linear interpolation of the fonts corresponding to the mixed emotions. In this work, we use four fonts for the four emotions. We conducted a preliminary subjective questionnaire in which 14 subjects selected the most appropriate font for each emotion from four to six candidate fonts. Based on the results of the questionnaire, we adopted "Birdmathon" for "happiness", "Reggae One" for "anger", "Shoukakari Sareri (line shape)" for "sadness", and "Genshin Gothic Light" for "neutral", respectively. Then, the linear interpolation of the four fonts is performed as:

$$\hat{z} = \sum_{i=1}^{4} p_i z_i, \tag{2}$$

where z_i $(i = 1, 2, 3, 4)$ and $\hat{z}$ are the category embedding of the font for the i-th emotion and the interpolated embedding, respectively, and p_i $(\sum_{i=1}^{4} p_i = 1)$ is the probability of the i-th emotion obtained from the soft-max layer for emotion category recognition.

D. Coloring the font image

1) Russell's circumplex model: In general, emotion models are designed based on either of the following two theories: the basic emotion theory, which assumes that emotions are divided into several basic emotions, such as "anger" and "sadness," and the dimensional theory, which assumes that emotions change continuously on a dimensional space. Russell [3] proposed a circumplex model from the viewpoint of dimensional theory, in which emotions can be expressed on a two-dimensional space of unpleasant–pleasant (valence axis) and deactivation–activation (activation axis). In Russell's model, for example, "anger" is located at the point where valence is small and activation is large.

2) Colormap settings in emotional space: Following the Russell's circumplex model, the proposed method estimates scores of valence and activation (emotion dimensions) from a speech utterance, and change the font color depending on the estimated scores. To do this, we define a color map on the valence–activation emotional 2D space.

Plutchik's wheel of emotions [4] is one of the most famous studies on the relationship between emotions and colors. Plutchik assigned colors to the eight basic emotions. For example, yellow, red, and blue are assigned to joy, anger, and sadness, respectively.

Here, we focus on the fact that the positional relationship between "anger", "sadness", and "joy (happiness)" in Russell's model is similar to that in the Plutchik's model. Therefore, we set the color map to reproduce the colors of those three emotions defined by Plutchik in Russell's emotional 2D space as shown in Fig. 4. Specifically, the RGB values $c_{\mathrm{R}}, c_{\mathrm{G}}, c_{\mathrm{B}}$ were set as follows:

$$\begin{aligned} c_{\mathrm{R}} &= \min\left(\frac{\max\left(v_{\mathrm{act}}, 0\right)}{d_{\mathrm{act}}} \times 255 + \frac{\max(-v_{\mathrm{val}}, 0)}{d_{\mathrm{val}}} \times 160, 255\right) \\ c_{\mathrm{G}} &= \frac{\max(v_{\mathrm{val}}, 0)}{d_{\mathrm{val}}} \times 160 \\ c_{\mathrm{B}} &= \frac{\max(-v_{\mathrm{act}}, 0)}{d_{\mathrm{act}}} \times 255, \end{aligned} \tag{3}$$

where $v_{\mathrm{val}}, v_{\mathrm{act}}$ are values of valence and activation, respectively, and their range are defined as $[-d_{\mathrm{val}}, d_{\mathrm{val}}], [-d_{\mathrm{act}}, d_{\mathrm{act}}]$, respectively.

III. Experiments

A. Evaluation of speech recognition, emotion category recognition, and emotion dimension estimation

1) Experimental setup: We used the IEMOCAP [11] dataset, which consists of approximately 12 hours of spoken English by 10 speakers. Each utterance has a speech content text, an emotion category label and a three-dimensional emotion dimension label of valence, activation and dominance.

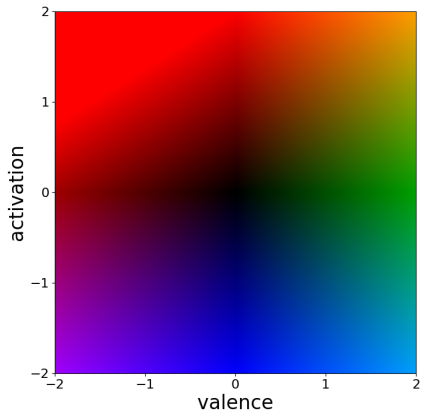

Fig. 4. Colormap on valence-activation space ($d_{\mathrm{val}} = d_{\mathrm{act}} = 2$).

We used 5,500 utterances in the dataset that were labeled with the five emotion categories of "happiness", "anger", "natural", "sadness", "excited". Then, regarding the "happiness" and "excited" as one emotion of the happiness emotion category, we treat the 5,500 utterances as belonging to four emotion categories.

We used the wav2vec2.0 base model[1] trained by LibriSpeech (approximately 960 hours). The speech recognition output is defined by 32 tokens, consisting of 26 alphabetic characters plus apostrophes and other symbols, while the emotion category recognition output is defined by the four emotions mentioned above. The emotion dimension estimation output is the scores of valence and activation. AdamW was used as the optimizer, and the model was evaluated at 100 epochs of training.

2) Results: Speech recognition was evaluated using word error rate (WER), and emotion category recognition was evaluated using Unweighted Accuracy (UA). The accuracy of valence and activation estimation was evaluated using the concordance correlation coefficient (CCC) between the value of the correct label and the estimated value, where CCC is calculated as follows:

$$CCC = \frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 + \sigma_y^2 + (\mu_x - \mu_y)^2} \tag{4}$$

where x and y are the correct value and the estimated value, respectively, μ_x and μ_y are their mean, and σ_x and σ_y are their standard deviation. Also, ρ is the Pearson's correlation coefficient between x and y.

Table I shows the results of 10- and 5-fold cross-validation (CV). The results of 5-fold CV are worse than that of 10-fold CV because the number of training data on 5-fold CV is less than 10-fold CV. Compared to recent work on the IEMOCAO dataset [10], [18], we found that our model performs as well as the compared methods, but our method can estimate three kind of information by using only one model.

[1]https://huggingface.co/facebook/wav2vec2-base

TABLE I
RESULTS OF SPEECH RECOGNITION, EMOTION CATEGORY RECOGNITION, AND EMOTION DIMENSION ESTIMATION.

method		WER [%]↓	UA [%]↑	CCC_{val}↑	CCC_{act}↑
[10]	10-fold	19.29	78.15	–	–
[18]	5-fold	–	74.4	0.660	0.717
ours	10-fold	19.82	77.29	0.7460	0.7332
ours	5-fold	22.12	73.78	0.7104	0.7291

TABLE II
RESULTS OF THE SUBJECTIVE EVALUATION WITH AND WITHOUT COLORING.

	ρ_{val}↑	ρ_{act}↑	CCC_{val}↑	CCC_{act}↑	category ACC[%]↑
w/o color	0.487	0.519	0.448	0.516	0.675
w/ color	0.502	0.674	0.467	0.663	0.632

B. Subjective evaluation of generated subtitle images

1) Experimental setup: Subjective evaluation experiments were conducted to evaluate the generated subtitle images. In these experiments, we used the ground truth of the emotion category and emotion dimensions to generate the font image instead of outputs of the wav2vec 2.0 because we evaluated the performance of the image generation methods purely independent of the performance of wav2vec 2.0. The participants in the experiment evaluated the following three items by looking at the image of a single lowercase letter of the alphabet for each of the colored and uncolored (black) font groups:

- How much positive emotion can you feel from this image? (scale of 1 to 5)
- How much active emotion can you feel from this image? (scale of 1 to 5)
- Which of the following emotions best describes what you feel as you look at the image: anger, sadness, happiness, or neutral?

2) Results: The results of the subjective evaluation by nine participants are shown in Table II. ρ_{val} and ρ_{act} are the correlation coefficient between the mean value of all the participants' evaluation values and the correct value for valence and activation, respectively. CCC_{val} and CCC_{act} are the CCC between the mean value of all the participants' evaluation values and the correct value, and category ACC is the rate (percentage) at which the correct emotion category was selected by the participants.

As shown in this table, the font image with color showed better performance on ρ_{val}, ρ_{act}, CCC_{val}, and CCC_{act} than without color. This result indicates that our defined color map successfully visualizes the valence and activation of the emotion. On the other hand, coloring the font image degraded the category ACC.

The confusion matrix of the emotion categories with and without coloring is shown in Fig. 5. From the figure, it can be seen that in the case of coloring, the generated image for "happiness" is more likely to be misidentified as "neutral" than in the case of no coloring. This may be because that in the proposed color map, the emotion with high valence and activation near 0, corresponding to the "happiness" category,

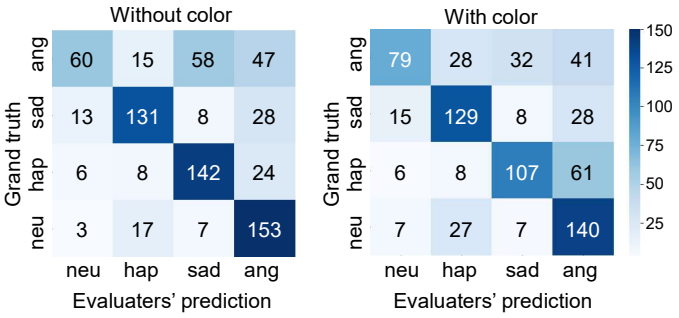

Fig. 5. Confusion matrices without and with coloring (neu: neutral, hap: happiness, sad: sadness, ang: anger).

is mapped in dark green even though we expected it to be mapped in yellow.

C. Subtitle generation from real speech utterances

Figure 6 shows an example of subtitle images generated by using the outputs of wav2vec 2.0 instead of the ground truth values[2]. Below each subtitle image, the estimated and correct values of the emotion category and emotion dimension of each utterance are shown.

In the top three images, the model's estimation results are generally correct; for example, in the top row, a font subtitle similar to the "happiness" font (Birdmathon) was generated in orange. In the bottom row, the speech of "sadness" was misrecognized as "neutral" and a font that is highly similar to the "natural" font (Genshin Gothic Light) was generated. However, this font was colored blue, corresponding to "sadness", due to the use of the emotion dimension estimation results. Thus, by independently determining the font from the emotion category recognition and the color from the emotion dimension estimation, it may be possible to correctly read the emotion from the other result even if a recognition error occurs in one of them.

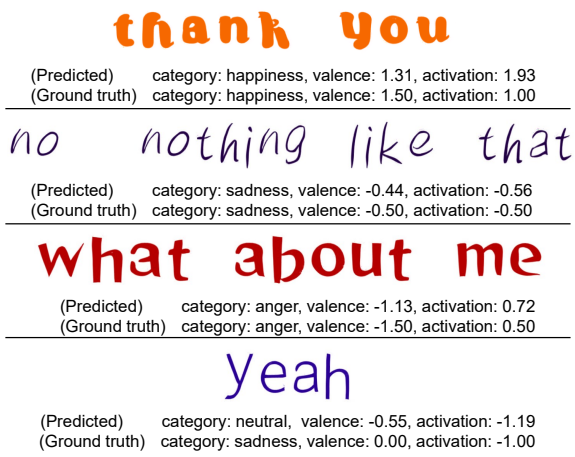

Fig. 6. Examples of generated subtitles by our proposed method.

IV. Conclusion

This paper proposed a method to generate subtitles that visualize the speaker's emotion using the fonts and colors that are used in the subtitles. Future work includes determining the appropriate font for each emotion and exploring color map settings that are more in line with human sensibilities.

References

[1] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] S.-B. Lim, Y.-S. Ji, B. Ahn, J. Park, and Y. Song, "Implementing and evaluating a font recommendation system through emotion-based content-font mapping," *Applied Sciences*, vol. 14, p. 1123, 01 2024.

[3] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[4] R. Plutchik, "A general psychoevolutionary theory of emotion," in *Theories of emotion*. Elsevier, 1980, pp. 3–33.

[2]Since the emotion dimension labels in IEMOCAP are between 1 and 5, we scaled them to a maximum of 2 and a minimum of −2.

[5] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[6] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[7] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[8] L. Pepino, P. Riera, and L. Ferrer, "Emotion recognition from speech using wav2vec 2.0 embeddings," *arXiv preprint arXiv:2104.03502*, 2021.

[9] J. Wagner, A. Triantafyllopoulos, H. Wierstorf, M. Schmitt, F. Burkhardt, F. Eyben, and B. W. Schuller, "Dawn of the transformer era in speech emotion recognition: closing the valence gap," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[10] X. Cai, J. Yuan, R. Zheng, L. Huang, and K. Church, "Speech Emotion Recognition with Multi-Task Learning." in *Interspeech*, 2021, pp. 4508–4512.

[11] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[12] "zi2zi: Master Chinese Calligraphy with Conditional Adversarial Networks." [Online]. Available: https://kaonashi-tyc.github.io/2017/04/06/zi2zi.html

[13] X. Liu, G. Meng, S. Xiang, and C. Pan, "FontGAN: a unified generative framework for Chinese character stylization and de-stylization," *arXiv preprint arXiv:1910.12604*, 2019.

[14] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[16] Y. Taigman, A. Polyak, and L. Wolf, "Unsupervised cross-domain image generation," *arXiv preprint arXiv:1611.02200*, 2016.

[17] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *International conference on machine learning*. PMLR, 2017, pp. 2642–2651.

[18] R. Sharma, H. Dhamyal, B. Raj, and R. Singh, "Unifying the Discrete and Continuous Emotion labels for Speech Emotion Recognition," *arXiv preprint arXiv:2210.16642*, 2022.