MIST-Tacotron: End-to-End Emotional Speech Synthesis Using Mel-Spectrogram Image Style Transfer

Sungwoo Moon¹, Junghyun Kim¹, and Yong-Hoon Choi¹, Member, IEEE

¹School of Robotics, Kwangwoon University, 20, Kwangwoon-ro, Nowon-gu, Seoul 01897, Korea

Corresponding author: Yong-Hoon Choi (e-mail: yhchoi@kw.ac.kr).

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government Ministry of Science and ICT (No. 2021R1F1A1064080). The present Research has been conducted by the Research Grant of Kwangwoon University in 2022.

ABSTRACT With the development of voice synthesis technology using deep learning, voice synthesis research that expresses the characteristics and emotions of speakers is actively being conducted. Current technology does not satisfactorily express various emotions and characteristics for speakers with very low or high vocal ranges and for speakers with dialects. In this paper, we propose mel-spectrogram image transfer (MIST)-Tacotron, a Tacotron 2-based speech synthesis model that adds a reference encoder with an image style transfer module. The proposed method is a technique that adds image style transfer to the existing Tacotron 2 model and extracts the speaker’s feature from the reference mel-spectrogram using a pre-trained deep learning model. Through the extracted feature, the style such as pitch, tone, and duration of the speaker are trained to express the style and emotion of the speaker more clearly. To extract the speaker’s style independently from the speaker’s timbre and emotion, the ID value for the speaker and the ID value for the emotional state were used as inputs. Performance is evaluated by F0 voiced error (FVE), F0 gross pitch error (F0 GPE), mel-cepstral distortion (MCD), band aperiodicity distortion (BAPD), voiced/unvoiced error (VUVE), false positive rate (FPR), and false negative rate (FNR). The performance of the proposed model was observed to have lower error values than the existing models, GST (Global Style Token) Tacotron and VAE (Variational Autoencoder) Tacotron. As a result of measuring mean opinion score (MOS), the sound quality of the proposed model received the highest score in terms of emotional expression and speaker style reflection.

INDEX TERMS Image style transfer, speech synthesis, Tacotron, text-to-speech (TTS), variational autoencoder (VAE), emotion expression, multi-speaker

I. INTRODUCTION

Speech synthesis, more specifically known as text-to-speech (TTS), is a comprehensive technology that involves many disciplines such as acoustics, linguistics, digital signal processing. Recently, as artificial neural network technology is applied to audio and speech fields, speech synthesis technology is rapidly progressing. The process of creating synthesized sound from text consists of two steps. The first step is to express the text as an intermediate expression, such as mel-spectrogram (i.e., text2mel [1-2]), and the second step is to synthesize the intermediate expression into high-quality raw waveform audio (i.e., mel2wav or vocoder [3-8]).

Various end-to-end TTS models [9-17] that combine mel-spectrogram generation process and raw waveform audio generation process into one pipeline are emerging and are rapidly progressing. Tacotron [9] is the first end-to-end generative TTS model based on the sequence-to-sequence neural network model [18] with attention module. It adopts Griffin-Lim vocoder [19] which requires acoustic features with high dimensionality and synthesizes speech with low quality. Google released Tacotron 2 [10], a more sophisticated model with location-sensitive attention [20], changed the model structure of encoders and decoders, and improved sound quality by applying WaveNet vocoder [1] instead of Griffin-Lim. Tacotron 2 greatly improved the quality of synthesized speech, but the inference time is longer than that of the later models. FastSpeech 1, 2, and 2s [13, 14] and SpeedySpeech [15] are representative models that improve the inference speed. The former [13, 14] are non-autoregressive model, mainly using a convolution layer, and the latter [15]
uses a knowledge distillation technique to improve the inference speed. These models [9-17] have good performance when synthesizing emotionless narrative-style monotone voices, but the speaker’s speech style cannot be controlled and voice containing various emotions cannot be produced.

Recently, studies on speech synthesis that can express various emotions and imitate the speaker’s style have been actively conducted [21-25]. Emotional expressions are directly influenced by the speaker’s intentions, leading to voices in various emotional categories such as happiness, anger, sadness, and fear. There are two ways to generate a voice with style. First, reference-based style transfer has emerged as a good solution for TTS that can express emotions. For example, global style token (GST) [21] is learned to have similar characteristics to reference audio. By learning a style embedding space through expressive audio samples in an unsupervised learning, synthetic audio that matches the prosody of the reference audio can be generated. The second method is to explicitly find out what emotions the token values can express, and then give them as conditional inputs to the TTS model to generate speech. The second method is to use multiple token values to generate the desired style voice, but a specific prosody may not be learned in the token during the learning process.

Variational Autoencoder (VAE) Tacotron [22] is also one of the TTS models that express the speaker’s style and various types of emotions. It learns the latent representation of speaking styles in an unsupervised manner. Style transfer is achieved by first inferring style representation through the recognition network of VAE, then feeding it into TTS network to guide the style in synthesizing speech. Although the VAE Tacotron reflects the speaker’s characteristics, the generated high- or low-pitched sound tends to be unclear because the inference results are concentrated on the average value. Li et al. [23] proposed a controllable emotion speech synthesis approach based on emotion embedding space learned from references. This method only supports a single speaker, but the synthesized sound is of good quality. Jia et al. [24] proposed a multi-speaker speech synthesis method through speaker embedding. Although this speech model has the advantage of being able to generate the voices of multiple speakers, it cannot express emotions clearly. Mellotron [25] is a model that learns only voices without emotions to create emotional voices and songs. This model does not require a dataset with emotion expressions, so it can be easily collected and trained. However, the emotional expression is not clear, and the inference speed is very slow. Representative multi-speaker TTS services currently commercialized in South Korea are Naver’s Clova Dubbing and Minds Lab’s AVA. To date, these two commercial services allow only 2-3 emotional expressions and can only express the voices of some pre-designated speakers.

In this paper, we propose mel-spectrogram image style transfer (MIST)-Tacotron, which can more clearly express the speaker’s characteristics and express various emotions than the existing TTS models. MIST-Tacotron is an end-to-end TTS model that combines Tacotron and image style transfer scheme. The image style transfer technique is a method of style conversion of the mel-spectrogram of the reference audio and using the features of the reference audio for training. The pre-trained visual geometry group (VGG)-19 [26] model was used to extract styles from the mel-spectrogram, and the speaker’s style was extracted through the mel-spectrogram reflecting the style. It has a robust model structure with fewer errors, such as skipping and repeating by accurately synthesizing the voices of various speakers given as reference audio and desired emotions.

To evaluate the performance of the proposed model, we perform speech synthesis for both single and multiple speakers. In the single speaker experiment, it is checked whether the style of the reference audio is reflected in the generated voice, and whether it can contain 7 different emotions in the voice. In the multi-speaker experiment, it is checked that the style and 4 different emotions for the reference audio were well expressed and the speaker’s voice was similar. Objective and subjective assessment of TTS voice were conducted to verify the proposed model. F0 voiced error (FVE) [27], F0 gross pitch error (F0 GPE) [28], mel-cepstral distortion (MCD) [29], band aperiodicity distortion (BAPD) [29], voiced/unvoiced error (VUVE) [30], false positive rate (FPR) [30], and false negative rate (FNR) [30] are used as the objective evaluation metrics of the proposed model. The subjective evaluation was conducted using the mean opinion score (MOS) for single speaker and multi-speakers, respectively. As a result of the single speaker evaluation, it was confirmed that the emotions to be expressed can be easily controlled. In the multi-speaker experiment, it was confirmed that the proposed model generates not only the four emotional expressions but also the reference speaker’s voice regardless of the speaker’s voice in the dataset used for training. Through experiments, we observed that MIST-Tacotron achieved higher MOS scores than the state-of-the-art publicly available models, GST and VAE Tacotron. Also, to compare under fair conditions, the speaker ID and emotion ID were learned by supervised learning in GST and VAE Tacotron and compared with the proposed MIST Tacotron. Our audio samples are publicly available on the demo website [31], and models and code are released for reproducibility and future development [32].

The structure of this paper is as follows. Section II describes the proposed model, and Section III describes the training method and performance evaluation method of the proposed model.
model. Section IV describes the experimental results. Finally, in Section V, conclusions and further studies are described.

II. THE PROPOSED MODEL

A. MEL-SPECTROGRAM IMAGE STYLE TRANSFER

Image style transfer is the task of changing the style of an image in one domain to the style of an image in another domain. Gatys et al. [33] introduced a neural algorithm that renders a content image in the style of another image, achieving so-called style transfer. Image style transfer is a technique used to take and blend two images, a content image and a style reference image, so that the output image looks like a content image but is drawn with the style of the style reference image. Fig. 1 shows examples of image style transfer.

There have been several studies [34], [35] that improved the inference speed in the field of style image transfer, and it was observed that the neural style transfer model [33] was superior according to the quantitative evaluation criteria. In this paper, the purpose is to obtain speaking style rather than enhancing processing speed, so the style information was extracted from the mel-spectrogram using the technique of [33] with VGG-19 deep neural networks for a pre-trained image classification model. Unlike the existing style transfer method, only style information of the mel-spectrogram is needed, so the style was extracted using only the style reconstruction module of [33].

Image style transfer is a technique that was applied to the image and video fields, and no cases of voice application have been observed yet. A method of extracting a style mel-spectrogram from audio is as follows. Audio is converted into mel-spectrogram using short-time Fourier transform (STFT). As the converted mel-spectrogram has a large value, so min-max normalization is applied to have a value between -1 and 0, and then data is preprocessed to match the input dimension of the pre-trained model. In this paper, since the pre-trained VGG-19 model was used, two-dimensional mel-spectrogram is converted into four-dimensional image. Since the length of each audio clip in the speech dataset is different, the size of the mel-spectrogram image was also created in proportion to the length of the audio clip. The generated mel-spectrogram image passes through five layers of an image classification model with a pretrained convolutional neural network (CNN) structure, and Gram matrices are obtained for features in each layer. Gram matrix \( G^k = [G_{ij}^k] \) is the inner product between the vectorized feature maps \( i \) and \( j \) in layer \( k \). We get another Gram matrix \( M^k = [M_{ij}^k] \) from mel-spectrogram. The contribution of layer \( k \) to the total loss is

\[
E_k = \sum_{ij} (G_{ij}^k - M_{ij}^k)^2
\]

and the total style loss is the square of the Frobenius norm of the difference between Gram matrices

\[
L_{style} = \sum_{k=0}^{L} (w_k E_k),
\]

where \( w_k \) are weighting factors of the contribution of each layer to the total loss. In this paper, the same weight 1 was used as the \( w_k \) value in all layers.

B. MIST-TACOTRON

In our model, Tacotron 2, which has a sequence-to-sequence structure, was used as the basic framework for generating a mel-spectrogram. The overall structure of the proposed model is shown in Fig. 2. If the input information is insufficient and modeled in one-to-many mapping according to L1 loss [36], the mel-spectrogram prediction may become over-smoothed. Over-smoothing effect usually makes synthetic speech sound muffled compared to concatenative speech synthesis or natural speech [37], [38]. Therefore, in order to improve the expression of the synthesized voice, it is necessary to provide as much information as possible as input and to have a model structure to learn it. Some multi-speaker TTS systems [39-42] explicitly model speaker representations through speaker table or speaker encoder.
In this paper, for the robust model structure, text, emotion ID, and reference mel-spectrogram were used for single speakers as inputs, and speaker ID was additionally used for multi-speakers. To extract the speaker's style, features containing the speaker's style are extracted using an image style transfer technique from the reference mel-spectrogram. A Style Encoder is used to encode the time series characteristics and key style characteristics of the data. The structure of the Style Encoder is shown in Fig. 3. It is composed of 6 convolution layers, and batch normalization and rectified linear unit (ReLU) activation function are applied to each layer. Finally, the feature values considering the temporal order are extracted by passing through the gated recurrent unit (GRU). This method enables prosody transfer to learn about the pronunciation of other speakers.

In the case of Korean, synthesized sounds are produced differently depending on the preprocessing method of the input text. The syllable-by-syllable input is seamlessly connected to each sound, resulting in a natural synthesized sound. However, the pheme-based input tends to be more articulate and the speaker's tone more vivid than the syllable-based input for each sound. We choose the phoneme-based input method. Phonemes are changed to integers by applying one-hot encoding, and are input to the encoder through text embedding. The speaker ID is a value that distinguishes speakers, and the emotion ID is a value that distinguishes emotions. Two ID values are embedded and used for training. After concatenating the text for TTS with the feature values passed through the Style Encoder, the embedded speaker ID, and the embedded emotion ID, the concatenated vector is applied as an input of the attention module. Decoder generates a mel-spectrogram to be used as an input of a vocoder by using the value obtained through the Attention Module. Through this method, the speaker’s voice and emotion are independently learned from the reference audio, and only feature values such as pitch and tone of the reference audio are extracted.

### III. TRAINING AND EVALUATION

Three experiments were performed to validate MIST-Tacotron model. The first experiment is an experiment for a single speaker, and we compared with other models whether the proposed model can generate voices containing 7 different emotions reflecting the speaker’s speaking style. The second experiment is an experiment for multi-speakers, and we compared with other models whether the proposed model can generate voices containing 4 different emotions reflecting another speaker’s speaking style. It also evaluated whether the proposed model accurately represents the speaker's voice. The third experiment evaluated the difference in synthesis speech according to the style loss value. The proposed model was compared with the existing models original GST-Tacotron, original VAE-Tacotron, GST-Tacotron with the robust structure, and VAE-Tacotron with the robust structure. Unlike the original models, the robust GST and VAE Tacotron models use the speaker ID value and the emotion ID value as inputs for supervised learning. As a subjective evaluation, a total of 182 sound sources were evaluated for 76 Koreans. The evaluation score is given by item from the lowest 1 point to the highest 5 points, and the higher the score, the better the sound quality.

#### A. DATASET

The single speaker voice data used in this experiment consists of 3,000 sentences (approximately 21,000 utterances) each for seven emotions (i.e., neutral, anger, fear, disgust, happiness, sadness, and surprise) by one professional female voice actor, and the recorded volume is about 30 hours. Of a total of 21,000 utterances, 100 utterances are used for model validation and 100 utterances are used for testing. The test data consisted of 7 emotions evenly mixed. Except for 200 utterances, the rest of the dataset are used for training. The multi-speaker dataset is provided by Pitchtron [43]. It consists of about 24,000 utterances and the recording time is about 34 hours. Of about 24,000 utterances, 100 utterances are used for model validation and 100 utterances for testing. The test data consisted of mixed emotions and several speakers. All data except 200 utterances are used as training data.

#### B. MODEL SETUP

The initial parameters of Text Encoder, Attention Module, and Decoder were set to the same values as the original Tacotron 2 model [10]. The number of six convolution layer output channels of Style Encoder is set to 32, 32, 64, 64, 128, 128, the same as the reference encoder used in GST-Tacotron. The size of the filter is set to 3 × 3, and the size of the stride is set to 2× 2. The tensor that passed through the convolution layer is converted into three dimensions and passed through a 256-sized unidirectional GRU with a fixed length. The mel-
A total of 7 metrics are used: FVE, F0 GPE, MCD, BAPD, VUVE, FPR, and FNR. Voice features are extracted using a WORLD vocoder [46] to obtain mel-cepstrum and band aperiodicity. The test is performed using 100 corpora not used for training and validation. MCD and BAPD are performance indicators that express the difference between the mel-cepstrum of the original voice and the synthetic sound, and the unit is [dB]. MCD is given by

$$\text{MCD}(x, \hat{x}) = \frac{10}{\ln(10)} \sum_{i=1}^{N} \sqrt{2 \sum_{j=1}^{4} (x_{ij} - \hat{x}_{ij})^2},$$  

where $x$ and $\hat{x}$ are mel-cepstrum of original and synthetic speeches, respectively, and $N$ is the total number of frames of synthetic speeches. BAPD is given by

$$\text{BAPD}(y, \hat{y}) = \frac{10}{\ln(10)} \sum_{i=1}^{N} \sqrt{2(y_{ij} - \hat{y}_{ij})^2},$$

where $y$ and $\hat{y}$ are band aperiodicity of the original voice and the synthetic speech, respectively.

### Table I: Objective Performance Evaluation of the Multi-Speaker Model

| Model          | FVE (±) | F0 GPE (±) | BAPD (±) | MCD (±) | VUVE (±) | FPR (±) | FNR (±) |
|----------------|---------|------------|----------|---------|----------|---------|---------|
| Original GST   | 64.24 (±6.25) | 54.53 (±5.50) | 6.82 (±0.34) | 6.71 (±0.16) | 33.34 (±1.31) | 65.90 (±2.24) | 23.21 (±1.32) |
| Original VAE   | 54.75 (±5.12) | 44.79 (±3.72) | 6.81 (±0.32) | 6.78 (±0.16) | 31.98 (±1.20) | 64.47 (±2.25) | 22.84 (±1.19) |
| Robust GST     | 53.98 (±4.14) | 41.61 (±3.88) | 5.84 (±0.34) | 6.59 (±0.13) | 32.33 (±1.21) | 63.68 (±2.22) | 22.33 (±1.27) |
| Robust VAE     | 53.48 (±4.41) | 44.85 (±2.66) | 5.95 (±0.32) | 6.62 (±0.20) | 32.85 (±1.31) | 64.74 (±2.04) | 22.99 (±1.28) |
| MIST 100       | 50.36 (±3.92) | 37.96 (±2.53) | 5.32 (±0.29) | 6.57 (±0.15) | 31.43 (±1.19) | 62.47 (±2.03) | 22.04 (±1.23) |
| MIST 10        | 48.99 (±3.87) | 39.70 (±2.42) | 5.31 (±0.31) | 6.51 (±0.14) | 31.53 (±1.16) | 62.60 (±2.27) | 21.79 (±1.23) |

### Table II: Objective Performance Evaluation of the Single Speaker Model

| Model          | FVE (±) | F0 GPE (±) | BAPD (±) | MCD (±) | VUVE (±) | FPR (±) | FNR (±) |
|----------------|---------|------------|----------|---------|----------|---------|---------|
| Original GST   | 68.68 (±8.86) | 41.77 (±5.15) | 3.62 (±0.40) | 5.76 (±0.15) | 44.09 (±1.37) | 53.95 (±2.56) | 37.33 (±2.43) |
| Original VAE   | 84.00 (±8.66) | 43.97 (±4.85) | 3.57 (±0.39) | 5.50 (±0.10) | 42.91 (±1.42) | 53.04 (±2.79) | 37.79 (±2.37) |
| Robust GST     | 60.84 (±7.95) | 34.71 (±4.35) | 3.58 (±0.40) | 5.38 (±0.07) | 41.71 (±1.33) | 52.47 (±2.71) | 37.05 (±2.23) |
| Robust VAE     | 62.45 (±7.54) | 31.56 (±3.98) | 3.57 (±0.39) | 5.42 (±0.07) | 41.03 (±1.34) | 51.17 (±2.70) | 36.40 (±2.25) |
| MIST 100       | 59.69 (±7.37) | 31.60 (±3.69) | 3.55 (±0.39) | 5.32 (±0.07) | 40.05 (±1.32) | 50.21 (±2.52) | 36.19 (±2.31) |
| MIST 10        | 61.74 (±7.34) | 31.23 (±3.85) | 3.55 (±0.39) | 5.35 (±0.07) | 40.55 (±1.30) | 51.24 (±2.68) | 36.75 (±2.19) |

The computer operating system is Ubuntu 18.04 LTS and Docker. Two RTX TITAN GPUs are used for training MIST-Tacotron, GST-Tacotron, and VAE-Tacotron. Vocoder is trained using two Quadro RTX 6000 GPUs.

### C. TRAINING AND INFERENCE

Existing models, GST-Tacotron and VAE-Tacotron, conducted training in the existing method without additional input information. GST-Tacotron and VAE-Tacotron, which applied robust model structures, used emotion ID values for single speakers and emotion ID and speaker ID values for multi-speakers for training. MIST-Tacotron also used emotion ID and speaker ID values for training for robust model structure.

To assess the performance of the model, both objective and subjective evaluations are performed. For objective evaluation,
FVE and F0 GPE are indicators representing the quality of voice using fundamental frequency F0, and the unit of both indicators is [%]. FVE is given by

\[ \text{FVE}(f, \hat{f}) = \sqrt{\frac{\sum_{i=1}^{N} (f_i - \hat{f}_i)^2}{N}}, \]  

where \( f \) and \( \hat{f} \) are F0 values of original and synthesis speech, respectively, and \( N \) is the total number of F0’s in the original speech. F0 GPE is given by

\[ \text{F0 GPE} = \frac{N_{p_{0e}}}{N_{v}} \times 100, \]  

where \( N_{p_{0e}} \) is the number of voiced frames when the estimated F0 values are incorrect, and \( N_{v} \) is the total number of voiced frames. The incorrect F0 value is defined as a value that falls outside \( \delta \) (typically \( \pm 20\% \)) of the correct F0 value.

VUV E, FPR, and FNR are indicators that determine whether the time at which the synthetic voice is uttered is similar to that of the original voice, and the unit is [%]. VUV E can be obtained as in (7) as a method of calculating the error ratio between the talk spurt and the silent period of the original speech and the synthesized speech.

\[ \text{VUV E} = \frac{1}{N} \sum_{i=1}^{N} S_i \times 100, \]  

where

\[ S_i = \begin{cases} 0, & v_i = \hat{v}_i \\ 1, & v_i \neq \hat{v}_i \end{cases}, \]  

where \( v_i \) and \( \hat{v}_i \) are voiced and unvoiced periods of original voice and synthesized voice, respectively. FPR and RNR are false positives (unvoiced frames predicted as voiced) rate and false negatives (voiced frames predicted as unvoiced) rate, respectively. For all evaluation indicators, the lower the score, the better the performance.

IV. RESULT

A. OBJECTIVE PERFORMANCE EVALUATION

Table I and Table II are the results of comparing the existing method and the proposed method. Original GST and Original VAE are models that only contain text and reference audio as input of the model as before. Robust GST and Robust VAE are models with robust structures using speaker ID and emotion ID values as inputs in addition to the input of existing models. MIST is a model proposed in this paper and is a model to which a robust model structure and a mel-spectrogram image style transfer technique are applied. The number next to MIST is the style loss value.

As a result of the experiment, the proposed model, MIST-Tacotron, shows the lowest error rates, so it can be confirmed that the performance is excellent. Comparing the original and robust models, it can be seen that the model’s performance improves when it has a robust model structure rather than using the original model. Additional input information on...
reference audio seems to have improved the performance of the model.

We checked the difference in performance according to the style loss value. MIST-Tacotron 100 has low VUVE, FPR, and F0 GPE values, indicating that the speaker's speaking speed is similarly followed. In addition, it can be seen that the overall performance of MIST-Tacotron 100 is better when it is for a single speaker than when it is for multi-speakers. MIST-Tacotron 10 has relatively the lowest BAPD, MCD, and FVE values, so it can be seen that the speaker's breathing and pronunciation characteristics are better learned.

Fig. 4 visualizes the F0 value for each model for one of the 100 test sound sources. In Fig. 4, the solid blue line is F0 of the reference audio and the dotted orange line is F0 of the synthetic sound. In Fig. 4(a), which is the existing GST-Tacotron, the value of F0, which is the synthetic audio, is generally lower than the value of the reference audio. Therefore, it can be seen that the FVE and F0 GPE values in Table I are higher than those of other models. In addition, since the synthetic sound was uttered faster than the reference audio, it shows a high value of VUVE as can be seen in Tables I and II. The F0 values of the existing VAE-Tacotron are shown in Fig. 4(b). As shown in Fig. 4(b), it can be seen that the overall value of F0 of the synthesized sound is similar to that of the reference audio, but the synthesized sound does not match the value of F0 of the reference audio. Robust GST-Tacotron (see Fig. 4(c)) and Robust VAE-Tacotron (see Fig. 4(d)) show that the F0 value of the synthetic sound has become similar to the reference audio overall, but the synthetic voice still speaks quickly.

The MIST-Tacotron model proposed in this paper shows that the value of F0 is very similar to the reference audio, as shown in Fig. 4(e) and Fig. 4(f). Therefore, as shown in Table I and Table II, the FVE, MCD, and BAPD values are relatively lower than those of other models, which can produce good synthetic sound. Observing Fig. 4(f), it can be seen that the F0 values of the reference audio and the synthetic voice are consistent compared to other models, indicating that the timing of the synthetic voice is very similar to the reference sound. Therefore, as shown in Table I and Table II, it can be confirmed that the VUVE and F0 GPE values of the MIST-Tacotron 10 model are the lowest.

**B. SUBJECTIVE PERFORMANCE EVALUATION**

Table III shows Style and Emotion MOS results for a single speaker and multi-speakers. MOS evaluation of style and emotion was conducted to evaluate whether the synthesized voice included the speaker's style and whether emotions were properly expressed. Style MOS is an evaluation of whether the synthesized sound represents the speaker's style well and the overall quality of the synthesized sound quality is good. Emotion MOS is a score on whether the synthesized voice expresses the same emotion as the reference sound given as input.

In the case of a single speaker, synthetic voices produced by original GST and original VAE tend to produce synthetic sounds of different emotions, even though the reference voice was neutral. In particular, VAE-Tacotron tends to have a blurred synthetic voice, so overall MOS scores were low. Both Robust GST-Tacotron and Robust VAE-Tacotron showed higher MOS values than original models. As mentioned earlier, the reason why the MOS score is higher than that of the two original models is that supervised learning about emotions was performed during training. In the case of MIST-Tacotron proposed in this paper, supervised learning of emotions was performed, and it was confirmed that there were few issues that deteriorated sound quality such as skipping and repeating. In addition, it can be estimated that the mel-spectrogram image style transfer technique extracts the speaker's style characteristics well, so that the Style MOS scores are high. When checking the MOS scores according to style loss, it can be seen that the synthetic sound quality of MIST Tacotron 10 is slightly higher than that of MIST Tacotron 100 by a slight difference.

In the case of multi-speakers, it can be seen that the MOS scores of the proposed model are the highest as in the case of a single speaker, and the Robust GST-Tacotron and Robust VAE-Tacotron models are higher than the original GST-Tacotron and original VAE-Tacotron models. In the case of the proposed MIST-Tacotron, since the emotion ID value and the speaker ID value are additionally used, it is possible to

| Model name | Single speaker | Multi-speaker |
|------------|----------------|---------------|
|            | Style MOS      | Emotion MOS   | Style MOS    | Emotion MOS   |
| Ground Truth | 4.84 (±0.05) | 4.53 (±0.09) | 4.79 (±0.04) | 4.70 (±0.05) |
| Original GST | 3.94 (±0.12) | 3.62 (±0.11) | 3.77 (±0.12) | 3.70 (±0.11) |
| Original VAE | 3.84 (±0.12) | 3.44 (±0.11) | 4.04 (±0.11) | 3.93 (±0.10) |
| Robust GST | 4.13 (±0.10) | 4.13 (±0.09) | 4.11 (±0.08) | 3.96 (±0.09) |
| Robust VAE | 4.08 (±0.10) | 4.20 (±0.09) | 4.10 (±0.10) | 4.05 (±0.09) |
| MIST 100 | 4.39 (±0.09) | 4.34 (±0.08) | 4.40 (±0.08) | 4.25 (±0.10) |
| MIST 10 | **4.41 (±0.10)** | **4.39 (±0.08)** | **4.45 (±0.08)** | **4.32 (±0.08)** |
accurately express emotions and to generate a synthesized sound with the voice of the desired speaker, so high MOS scores seem to have been obtained. In the case of MIST-Tacotron, the expression of high and low tones is clear, and the MOS scores of the proposed model are evaluated to be the highest, especially in the case of a voice with a wide range of pitch fluctuations, such as in the southeastern dialect of South Korea. GST-Tacotron and VAEE-Tacotron do not use CNN-series modules when extracting features from reference audio but use neural networks consisting of attention modules and fully connected layers. CNN has excellent performance in extracting the characteristics of data, so it is observed that it has better performance than models that do not use CNN series.

V. CONCLUSIONS
In this paper, we proposed MIST-Tacotron, a Tacotron 2-based speech synthesis model that adds a reference encoder with an image style transfer module. The proposed method is a technique for adding image style transfer to the existing Tacotron 2 model and extracting speaker features from a reference mel-spectrogram using a pre-trained deep learning model. Through the extracted feature, the style such as pitch, tone, and duration of the speaker are trained to express the style and emotion of the speaker more clearly. The performance evaluation for FVE, F0 FPE, MCD, BAPD, VUVE, FPR, and FNR shows that the performance of the proposed model has the lowest error value compared to the existing models, GST-Tacotron and VAEE-Tacotron. The MOS measurement result also confirmed that the proposed model had the highest score. Therefore, it was confirmed that the proposed model structure and technique can produce a voice that reflects the speaker's characteristics and natural emotion expression is possible.

It is expected that more lively services will be provided by using the proposed model for producing audio content that requires various speakers and various emotional expressions. The image style transfer technique continues to develop in the field of image processing, so if a new image style transfer model is developed, it will be applied to the proposed Style Encoder of reference mel-spectrogram. In addition, we are applying recently developed HiFi-GAN and LightSpeech vocoders to MIST-Tacotron to further improve the quality of synthetic speech. So far, the proposed model has only been applied to Korean voice synthesis, but it plans to expand it to various languages.

ACKNOWLEDGMENT
The experiments in this paper were performed with GPU resources provided by the Korea NIPA (National IT Industry Promotion Agency). In addition, the GPU resources purchased through the following research project were partially used. The research project is “Convergence and Open Sharing System” Project, supported by the Ministry of Education and National Research Foundation of Korea.

REFERENCES
[1] Aaron van den Oord, Sander Dieleman, HeigaZen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu, “Wavenet: A generative model for raw audio,” arXiv preprint arXiv:1609.03499, 2016.
[2] Aaron Oord, Yazhe Li, Igor Babuschkin, Karen Simonyan, Oriol Vinyals, Koray Kavukcuoglu, George Driessche, Edward Lockhart, Luis Cobo, and Florian Stimberg, “Parallel Wavenet: Fast High-Fidelity Speech Synthesis,” In proceedings of International Conference on Machine Learning, pages 3918–3926, PMLR, 2018.
[3] R. Prenger, R. Valle and B. Catanzaro, “Waveglow: A Flow-based Generative Network for Speech Synthesis,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 3617-3621, doi: 10.1109/ICASSP.2019.8683143.
[4] R. Yamamoto, E. Song and J.-M. Kim, “Parallel WaveGAN: A Fast Waveform Generation Model Based on Generative Adversarial Networks with Multi-Resolution Spectrogram,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6199-6203, doi: 10.1109/ICASSP40776.2020.9057975.
[5] Kundan Kumar, Rithesh Kumar, Thibault de Boissiere, Lucas Gestein, Wei Zhen Treol, Jose Sotelo, Alexandre de Brebisson, Yoshua Bengio, Aaron Courville, “MelGAN: Generative Adversarial Networks for Conditional Waveform Synthesis,” Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, Canada.
[6] Nal Kalchbrenner, Erich Elsen, Karen Simonyan, Seb Noury, Norman Casagrande, Edward Lockhart, Florian Stimberg, Aaro n Oord, Sander Dieleman, Koray Kavukcuoglu, “Efficient Neural Audio Synthesis,” Proceedings of the 35th International Conference on Machine Learning, PMLR 80:2410-2419, 2018.
[7] Kong, Jungil, et al. “HiFi-GAN: Generative Adversarial Networks for Efficient and High-Fidelity Speech Synthesis.” Advances in Neural Information Processing Systems 33 (NeurIPS), 2020.
[8] R. Luo et al., “Lightspeech: Lightweight and Fast Text to Speech with Neural Architecture Search,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 5699-5703, doi: 10.1109/ICASSP39728.2021.9414403.
[9] Yuxuan Wang RJ Skerry-Ryan Daisy Stanton Yonghui Wu Ron J. Weiss Nduevoep Jaitly Zongheng Yang Ying Xiao Zhiyong Chen Samy Bengio Quoc Le Yannis Agiomyrgiannakis Rob Clark Rif A. Saurous, "Tacotron: Towards End-to-End Speech Synthesis," Interspeech (2017), Stockholm, Sweden, 2017.
[10] J. Shen et al., “Natural TTS Synthesis by Conditioning Wavenet on Mel Spectrogram Predictions,” 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018, pp. 4779-4783, doi: 10.1109/ICASSP.2018.8461368.
[11] Sercan Ö. Arık, Mike Chrzanowski, Adam Coates, Gregory Diamos, Andrew Gibiansky, Yongguo Kang, Xian Li, John Miller, Andrew Ng, Jonathan Raiman, Shubho Sengupta, Mohammad Shoeybi, “Deep Voice: Real-time Neural Text-to-Speech,” Proceedings of the 34th International Conference on Machine Learning, PMLR 70:195-204, 2017.
[12] Li, Naihan et al. “Neural speech synthesis with transformer network,” Proceedings of the Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence. Vol. 33. No. 01. 2019.
[13] Yi Ren et al. “FastSpeech: Fast, Robust and Controllable Text to Speech,” Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019.
[14] Yi Ren et al. “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech,” International Conference on Learning Representations. 2020.
[15] Jan Vainer and Ondr’ej Dušek, “SpeedySpeech: Efficient Neural Speech Synthesis,” InterSpeech 2020, 3575-3579, DOI: 10.21437/Interspeech.2020-2867, 2020.
[16] Rafael Valle, Kevin Shih, Ryan Prenger, and Bryan Catanzaro, “Flowtron: An Autoregressive Flow-based Generative Network for Text-to-Speech Synthesis,” NVIDIA ADLR, May 2020.
[17] A. LaFGuci, “Fastpitch: Parallel Text-to-Speech with Pitch Prediction,” ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2021, pp. 6588-6592, doi: 10.1109/ICASSP39728.2021.9413889.
[18] I. Sutskever, O. Vinyals and Q.V. Le, “Sequence to sequence learning with neural networks,” Proc. Int. Conf. Neural Information Processing Systems (NeurIPS), pp. 3104–3112, 2014.

[19] D. Griffin and Jae Lim, “Signal estimation from modified short-time Fourier transform,” ICASSP ’83. IEEE International Conference on Acoustics, Speech, and Signal Processing, 1983, pp. 804-807, doi: 10.1109/ICASSP.1983.1172092.

[20] Vaswani, Ashish, et al. “Attention Is All You Need,” Advances in neural information processing systems, 2017, pp. 5998-6008.

[21] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in Proc. ICML, 2018.

[22] Y. Zhang, S. Pan, L. He and Z. Ling, “Learning Latent Representations for Style Control and Transfer in End-to-End Speech Synthesis,” ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 6945-6949, doi: 10.1109/ICASSP.2019.8683623.

[23] T. Li, S. Yang, L. Xue and L. Xie, “Controllable Emotion Transfer for End-to-End Speech Synthesis,” 2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP), 2021, pp. 1-5, doi: 10.1109/ISCSLP49672.2021.9362069.

[24] Ye Jia, Yu Zhang, Ron J Weiss, Quan Wang, Jonathan Shen, Fei Ren, Zhifeng Chen, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno et al., “Transfer learning from speaker verification to multi-speaker text-to-speech synthesis,” Conference on Neural Information Processing Systems (NeuIPS), 2018.

[25] R. Valle, J. Li, R. Prenger and B. Catanzaro, “Mellotron: Multi-speaker Expressive Voice Synthesis by Conditioning on Rhythm, Pitch and Global Style Tokens,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 6189-6193, doi: 10.1109/ICASSP40776.2020.9054556.

[26] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-scale Image Recognition,” in Proc. Int. Conf. Learn. Representations, 2015.

[27] Juheon Lee, Hyeong-Seok Choi, Chang-Bin Jeon, Junghyun Koo, and Kyong Soo Lee, “Adversarially Trained End-to-end Korean Singing Voice Synthesis System,” Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech), 2019, pp. 2588–2592.

[28] Tomohiro Nakatani, Shigeaki Amano, Toshio Inizo, Kentaro Ishizuka, and Tadahisa Kondo, “A Method for Fundamental Frequency Estimation and Voicing Decision: Application to Infant Utterances Recorded in Real Acoustical Environments,” Speech Communication, vol. 50, no. 3, pp. 203–214, Mar. 2008.

[29] J. Kim, H. Choi, J. Park, M. Hahn, S. Kim, and J. Kim, “Korean Singing Voice Synthesis System Based on an LSTM Recurrent Neural Network,” INTERSPEECH, September 2018, pp. 1551–1555.

[30] M. Blauw and J. Bonada, “A Neural Parametric Singing Synthesizer Modeling Timbre and Expression from Natural Songs,” Natural Sciences, 2017, vol. 7, no. 12.

[31] https://restarbreaker.tistory.com/5

[32] https://github.com/Moon-sung-woo/MIST_Tacotron

[33] L. A. Gatys, A. S. Ecker and M. Bethge, “Image Style Transfer Using Convolutional Neural Networks,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2414-2423, doi: 10.1109/CVPR.2016.265.

[34] X. Huang and S. Belongie, “Arbitrary Style Transfer in Real-Time with Adaptive Instance Normalization,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1510-1519, doi: 10.1109/ICCV.2017.167.

[35] D. Ulyanov, A. Vedaldi and V. Lempitsky, “Improved Texture Networks: Maximizing Quality and Diversity in Feed-Forward Stylization and Texture Synthesis,” 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 4105-4113, doi: 10.1109/CVPR.2017.437.

[36] S. Gazor and Wei Zhang, “Speech Probability Distribution,” IEEE Signal Processing Letters, vol. 10, no. 7, pp. 204-207, July 2003, doi: 10.1109/LSP.2003.813679.

[37] S. Takamichi, T. Toda, A. W. Black, G. Neubig, S. Sakti and S. Nakamura, “Postfilters to Modify the Modulation Spectrum for Statistical Parametric Speech Synthesis,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 24, no. 4, pp. 755-767, April 2016, doi: 10.1109/TASL.2016.2522655.

[38] T. Toda, and K. Tokuda, “A Speech Parameter Generation Algorithm Considering Global Variance for HMM-based Speech Synthesis,” IEICE Transactions on Information and Systems, 90(5), 2007, pp.816-824.

[39] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, “Deep Voice 2: Multi-speaker Neural Text-to-Speech,” Advances in Neural Information Processing Systems 30, 2017, pp. 2962–2970.

[40] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, “S. Narang, J. Raiman, and J. Miller,” Deep voice 3: 2000- Speaker Neural Text-to-Speech,” CoRR, vol. abs/1710.07654, 2017.

[41] H. B. Moss, V. Aggarwal, N. Prateek, J. González and R. Barra-Chicote, “BOFFIN TTS: Few-Shot Speaker Adaptation by Bayesian Optimization,” ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020, pp. 7639-7643, doi: 10.1109/ICASSP40776.2020.9054301.

[42] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, “Multi-speech: Multi-speaker Text to Speech with Transformer,” arXiv pre-print arXiv:2006.04664, 2020.

[43] Sunghee Jung and Hee rin Kim, “Pitchtron: Towards Audiobook Generation from Ordinary People’s Voices,” arXiv pre-print arXiv:2005.10456, 2020.

[44] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto, “librosa: Audio and Music Signal Analysis in Python,” Proceedings of the 14th Python in Science Conference, 2015.

[45] D. P. Kingma and J. L. Ba, “Adam: A Method for Stochastic Optimization,” Proc. Int. Conf. Learn. Represent., 2015, pp. 1–15.

[46] M. Morise, F. Yokomori, and K. Ozawa, “WORLD: a Vocoder Based High-quality Speech Synthesis System for Real-time Aplications,” IEICE Trans. Inf. Syst., vol. E99-D, no. 7, 2016, pp. 1877–1884.