IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 14, 2020, accepted August 31, 2020, date of publication September 4, 2020, date of current version September 16, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3021758

Effective Emotion Transplantation in an End-to-End Text-to-Speech System

YOUNG-SUN JOO[1,2], (Member, IEEE), HANBIN BAE[1], YOUNG-IK KIM[1], HOON-YOUNG CHO[1], AND HONG-GOO KANG[2], (Member, IEEE)

[1]Speech AI Lab., NCSOFT Corporation, Gyeonggi-do 13494, South Korea
[2]Department of Electrical and Electronics Engineering, Yonsei University, Seoul 03722, South Korea

Corresponding author: Hong-Goo Kang (hgkang@yonsei.ac.kr)

ABSTRACT In this paper, we propose an effective technique to transplant a source speaker's emotional expression to a new target speaker's voice within an end-to-end text-to-speech (TTS) framework. We modify an expressive TTS model pre-trained using a source speaker's emotional speech database to reflect the voice characteristics of a target speaker for which only a neutral speech database is available. We set two adaptation criteria to achieve this. One criterion is to minimize the reconstruction loss between the target speaker's recorded and synthesized speech, such that the synthesized speech has the target speaker's voice characteristics. The other criterion is to minimize the emotion loss between the emotion embedding vectors extracted from the reference expressive speech and the target speaker's synthesized expressive speech, which is essential to preserve expressiveness. Since the two criteria are applied alternately in the adaptation process, we are able to avoid the kind of bias issues frequently encountered in similar tasks. The proposed adaptation technique demonstrates more effective performance compared to conventional approaches in both quantitative and qualitative evaluations.

INDEX TERMS End-to-end text-to-speech, expressive TTS, adaptation.

I. INTRODUCTION

The task of generating natural speech from the input text, i.e., text-to-speech (TTS), is becoming increasingly important, as it is a key module in building human-computer interaction systems. Thanks to the powerful modeling capabilities of deep learning technologies, the sound quality and naturalness of synthesized speech have substantially improved in recent years [1]–[6]. In particular, end-to-end framework-based TTS models that infer acoustic features directly from input character sequences without laborious feature-engineering tasks have shown great success [6]–[8].

Because of the success of end-to-end text-to-speech (E2E-TTS) models, researchers have been trying to expand this framework to synthesize more expressive speech [9]–[13]. Unlike emotionally neutral speech (narrative speech) which has monotonic prosody, expressive speech has many variations in prosody. Thus, a key challenge in synthesizing expressive speech lies in determining distinctive characteristics of different expressions and representing them using condition vectors to control the expressive TTS model.

Condition vectors can either be handcrafted or learned in the TTS model's training stage. An E2E-TTS framework mainly uses learned vectors, so-called embedding vectors or latent variables. These are jointly trained with the weights of the TTS model using backpropagation [14]. For example, [10] and [11] trained embedding vectors in a supervised manner using emotion labels. Recently, several studies have adopted an unsupervised method in which embedding vectors are trained in a deep learning framework, but without annotated labels [12], [13], [15]. This method is useful when it is difficult to obtain labeled data or when the speech data contains ambiguous styles that are difficult to classify. In [12] and [13], networks were trained to directly extract embedding vectors from a reference speech waveform during the overall training process, and the style of synthesized speech was controlled using the embedding vectors.

Although E2E-TTS models with condition vectors are very effective, it is often difficult to deploy them in real-world applications because of database issues. In such applications, high-quality expressive speech databases with

The associate editor coordinating the review of this manuscript and approving it for publication was Inês Domingues.

multiple voice identities are required. This means that expressive speech databases recorded by many professional voice actors and actresses are required. However, it is expensive and time-consuming to construct these kinds of expressive speech databases. In addition, it is difficult to utter expressive speech while maintaining consistent expressiveness.

An effective way to solve this problem is to use a technique like speaker adaptation [16]–[20], in which a baseline model is trained using a large database, then adjusted to a target speaker using only a small amount of data. This approach can similarly be applied to expressiveness tasks through emotion transplantation, i.e. training an expressive TTS model using available other speaker's expressive speech database and adjusting the pre-trained model to the target speaker's voice [21]–[24]. Even when there is only a small amount of expressive speech data of the target speaker, a target speaker's expressive TTS model can be obtained fairly easily by adapting the pre-trained model to minimize reconstruction loss, which is an error between recorded and synthesized expressive-speech of the target speaker.

In this paper, we deal with the case in which the target speaker has only neutral speech data. We experimentally found that the style of synthesized speech becomes ambiguous as the model adaptation progresses; eventually, the output synthesized speech does not faithfully present the expressiveness style. Because the model is adapted to reconstruct the target speaker's voice using neutral speech, the model's capability for generating expressive speech is impaired. To deal with this, we propose an effective emotion transplantation technique that guides the pre-trained model to preserve expressiveness characteristics during the adaptation process. The model update procedure has two alternating steps: (1) modifying the voice characteristics of the pre-trained model to match those for the target speaker and (2) preserving its capability for generating expressive speech. More specifically, when adapting the TTS model to minimize the reconstruction loss for the target speaker's neutral-style speech, the proposed technique synthesizes expressive speech with the target speaker's voice from the adapted TTS model and updates the model to minimize a metric we call the emotion loss. The emotion loss is the distance between the expressive condition vector extracted from the target speaker's synthesized expressive-speech and the input expressive condition vector used to synthesize speech expressively. That is, the model is updated so that the emotional style of synthesized expressive-speech matches the emotional style included in the input expressiveness condition vector. Here, the condition vector is extracted from the source speaker's expressive speech because the target speaker does not have expressive speech data. For the same reason, the emotion loss function compares the expressiveness condition vectors instead of the target speaker's expressive speech data. The condition vectors extracted from the synthesized target speaker's expressive speech should have an emotional style identical to that of the input condition vector. These two steps are repeated alternately until the model converges.

The remainder of this paper is organized as follows. In Section II, we describe the end-to-end expressive TTS model architecture to understand our proposed approach. In Section III, we explain the proposed effective emotion transplantation approach. Section IV provides objective and subjective experimental results, and Section V summarizes and concludes the paper.

II. MODEL ARCHITECTURE

The end-to-end expressive TTS model used in this paper consists of two components: (1) an emotion encoder which outputs an expressiveness condition vector based on a reference expressive speech input, and (2) an E2E-TTS model which synthesizes expressive speech using text input and expressiveness condition vectors.

A. EMOTION ENCODER

To obtain the expressiveness condition vector, we adopt the global style token (GST) approach [13], which derives the expressiveness condition vector on the fly by passing reference speech. Figure 1 illustrates the GST architecture. The GST, which consists of a reference encoder [12] and a style token layer [13], is jointly trained while training a TTS model as follows. The reference encoder outputs a prosody embedding vector based on a reference speech input, typically having a mel-scale spectrogram (mel-spectrogram) format. The prosody embedding vector is used as the input to the style token layer, comprising an attention module and a set of token embeddings [13]. The style of the given reference speech is then presented as the weighted sum of each token, where the weights are the contributions of each token to the prosody embedding vector. The *style* can be defined in various meanings such as voice characteristics, speaking style, emotions, etc. For this study, we focus on emotions and use the term *emotion embedding vector* rather than expressiveness condition vector for the remainder of this paper.

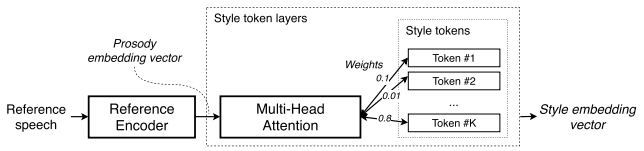

FIGURE 1. Global style token (GST) model diagram. The mel-spectrogram of the reference speech is fed to the reference encoder followed by a style token layer.

B. END-TO-END TTS MODEL

Among various E2E-TTS frameworks, we use the deep convolutional TTS (DCTTS) framework because of its fast training speed and stable alignment [8]. The DCTTS framework consists of a text-to-mel-spectrogram network (Text2Mel) and a spectrogram super-resolution network (SSRN). The Text2Mel module predicts a coarse mel-spectrogram, that is, a down-sampled mel-spectrogram in the time axis, based on input texts and the predicted mel-spectrogram at the previous

time step, in an auto regressive manner. The SSRN module predicts a linear-scale spectrogram (spectrogram), that is up-sampled in time and frequency axis, from the coarse mel-spectrogram.

More specifically, text embedding and audio embedding sequences are extracted from text and audio encoders, respectively. Here, the mel-spectrogram predicted at the previous time step (i.e., the audio decoder output) enters the audio encoder input. Attention values between these are multiplied to the text embedding vectors, after which the attended text embedding vectors are concatenated to audio embedding vectors. In this paper, the emotion embedding vector inferred by the emotion encoder is also concatenated. The audio decoder then autoregressively infers the mel-spectrogram from these combined embeddings. Finally, the spectrogram predicted by the SSRN module is converted into a time-domain speech waveform using either the Griffin-Lim algorithm [25] or any type of generative model such as WaveNet [5], [26].

III. PROPOSED EMOTION TRANSPLANTATION APPROACH

In this section, we describe an effective adaptation method that successfully transplants the emotional expressiveness of a pre-trained model to the target speaker's voice.

Assume that we have an expressive TTS model pre-trained with a source speaker's expressive speech database, but only a neutral speech database is available for the target speaker. A simple idea is to adapt the pre-trained source speaker's expressive TTS model with the target speaker's neutral speech database. However, in a preliminary experiment, we found that this simple adaptation approach could not maintain the TTS model's ability to express appropriate emotions in the synthesized speech. Therefore, we designed the proposed approach by considering two aspects: generating the target speaker's voice characteristics and expressing appropriate emotions. During the process of adapting the TTS model based on a target speaker's neutral speech, we synthesize expressive speech with the target speaker's voice by inputting an emotion embedding vector extracted from the source speaker's expressive speech to the adapted TTS model. We then update the model by minimizing the emotion loss between the input emotion embedding vector and the emotion embedding vector extracted from the synthesized expressive-speech. The detailed procedure is described in Algorithm 1.

A. BUILDING A SOURCE SPEAKER'S EXPRESSIVE TTS MODEL

Figure 2 and Phase 1 of Algorithm 1 describe the procedure for training a baseline expressive TTS model using a source speaker's expressive speech database.

Training the expressive TTS model requires input text $\mathbf{X}$, output expressive speech $\mathbf{S}^{\text{src,emo}}$, and input reference expressive speech $\mathbf{S}_{\text{ref}}^{\text{src,emo}}$ from a source speaker's expressive speech database $D_{\text{src,emo}}$. In this work, we chose the reference speech to be non-parallel with the input text so that the emotion encoder is robust to the reference speech signal content; this maximizes expressive information and minimizes the influence of the reference speech signal content.

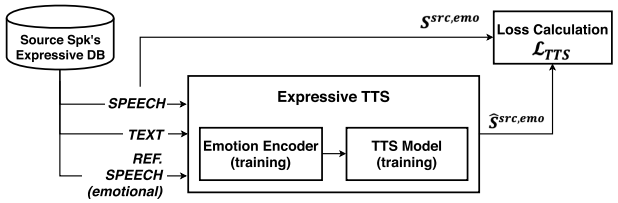

FIGURE 2. Baseline expressive TTS model training step.

The emotion encoder $\mathcal{M}_{\text{emo}}$ and the TTS model $\mathcal{M}_{\text{TTS}}$ are trained jointly. First, $\mathcal{M}_{\text{emo}}$ outputs the emotion embedding vector $\mathbf{e}$ from $\mathbf{S}_{\text{ref}}^{\text{src,emo}}$. Then, $\mathbf{e}$ is passed as input to $\mathcal{M}_{\text{TTS}}$ with $\mathbf{X}$. The training process is identical to conventional TTS models in that it minimizes a reconstruction loss between recorded and synthesized expressive-speech, namely $\mathbf{S}^{\text{src,emo}}$ and $\hat{\mathbf{S}}^{\text{src,emo}}$. We define the loss function of $\mathcal{M}_{\text{TTS}}$, $\mathcal{L}_{\text{TTS}}$, as the sum of L1 loss L_1 and a binary divergence function $\mathcal{D}_{\text{bd}}$ following the DCTTS system [8], namely,

$$\mathcal{L}_{\text{TTS}}(S, \hat{S}) = L_1(S, \hat{S}) + \mathcal{D}_{\text{bd}}(\hat{S}|S), \tag{1}$$

where S and $\hat{S}$ are recorded and synthesized speech, respectively, in mel-spectrogram format. In this study, we do not use the guided attention loss introduced in [8]. This guided attention loss prompts the attention matrix to be nearly diagonal, but it is not effective when there are many variations in the attention pattern caused by various speaking speeds dependent on the emotion classes.

B. ADAPTING VOICE CHARACTERISTICS WHILE MAINTAINING EXPRESSIVENESS CHARACTERISTICS

Figure 3 and Phase 2 of Algorithm 1 describe the emotion transplantation procedure. The pre-trained source-speaker's TTS model is modified to have the target speaker's voice characteristics while maintaining its own expressiveness characteristics. The model is updated in two steps. In the first step, only $\mathcal{M}_{\text{TTS}}$, excluding $\mathcal{M}_{\text{emo}}$, is adapted for an epoch of the target speaker's neutral speech database $D_{\text{tgt,neu}}$.

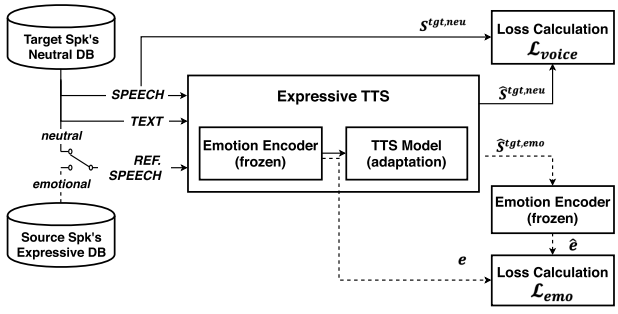

FIGURE 3. Proposed emotion transplantation approach, showing the pre-trained expressive TTS model adaptation step (solid line: flows related to neutral speech; dashed line: flows related to expressive speech).

Algorithm 1 The Emotion Transplantation Approach

Phase 1 - Training of a baseline expressive TTS model

Require:
Training data $(\mathbf{X}, \mathbf{S}^{\text{src,emo}}, \mathbf{S}^{\text{src,emo}}_{\text{ref}}) \in D_{\text{src,emo}}$

while $\mathcal{M}_{\text{emo}}$, $\mathcal{M}_{\text{TTS}}$ are not converged **do**
  $\mathbf{e} \leftarrow \mathcal{M}_{\text{emo}}(\mathbf{S}^{\text{src,emo}}_{\text{ref}})$
  $\hat{\mathbf{S}}^{\text{src,emo}} \leftarrow \mathcal{M}_{\text{TTS}}(\mathbf{X}, \mathbf{e})$
  $\mathcal{L}_{\text{TTS}} \leftarrow \mathcal{L}_{\text{TTS}}(\mathbf{S}^{\text{src,emo}}, \hat{\mathbf{S}}^{\text{src,emo}})$
  Update $\mathcal{M}_{\text{emo}}$, $\mathcal{M}_{\text{TTS}}$ jointly with regard to $\mathcal{L}_{\text{TTS}}$
end while

Phase 2 - Emotion transplantation

Require:
Training data $(\mathbf{X}, \mathbf{S}^{\text{tgt,neu}}, \mathbf{S}^{\text{tgt,neu}}_{\text{ref}}) \in D_{\text{tgt,neu}}$
Training data $(\mathbf{X}, \mathbf{S}^{\text{tgt,neu}}) \in D_{\text{tgt,neu}}$ and $\mathbf{S}^{\text{src,emo}}_{\text{ref}} \in D_{\text{src,emo}}$
Pre-trained
models $\mathcal{M}_{\text{emo}}$ and $\mathcal{M}_{\text{TTS}}$

while $\mathcal{M}_{\text{TTS}}$ is not converged **do**
  for $i \leftarrow 1, N_{\text{tgt}}$ **do**
    $\mathbf{e} \leftarrow \mathcal{M}_{\text{emo}}(\mathbf{S}^{\text{tgt,neu}}_{\text{ref}})$
    $\hat{\mathbf{S}}^{\text{tgt,neu}} \leftarrow \mathcal{M}_{\text{TTS}}(\mathbf{X}, \mathbf{e})$
    $\mathcal{L}_{\text{voice}} \leftarrow \mathcal{L}_{\text{TTS}}(\mathbf{S}^{\text{tgt,neu}}, \hat{\mathbf{S}}^{\text{tgt,neu}})$
    Update $\mathcal{M}_{\text{TTS}}$ with regard to $\mathcal{L}_{\text{voice}}$
  end for

  for $i \leftarrow 1, N_{\text{src}}$ **do**
    $\mathbf{e} \leftarrow \mathcal{M}_{\text{emo}}(\mathbf{S}^{\text{src,emo}}_{\text{ref}})$
    $\hat{\mathbf{S}}^{\text{tgt,emo}} \leftarrow \mathcal{M}_{\text{TTS}}(\mathbf{X}, \mathbf{e})$
    $\hat{\mathbf{e}} \leftarrow \mathcal{M}_{\text{emo}}(\hat{\mathbf{S}}^{\text{tgt,emo}})$
    $\mathcal{L}_{\text{emo}} \leftarrow \text{Dist}(\mathbf{e}, \hat{\mathbf{e}})$
    Update $\mathcal{M}_{\text{TTS}}$ with regard to $\mathcal{L}_{\text{emo}}$
  end for
end while

The loss function for voice $\mathcal{L}_{\text{voice}}$ follows Eq. (1). In the second step, the adapted $\mathcal{M}_{\text{TTS}}$ is updated by optimizing the emotion loss function,

$$\mathcal{L}_{\text{emo}} = \text{Dist}(\mathbf{e}, \hat{\mathbf{e}}), \tag{2}$$

where $\mathbf{e}$ and $\hat{\mathbf{e}}$ are emotion embedding vectors extracted from recorded and synthesized expressive speech, respectively. We use L_1 as a distance metric Dist. In our preliminary experiments, other distance metrics such as cosine distance also showed similar results. In this work, $\hat{\mathbf{e}}$ is extracted from a synthesized expressive speech of a target speaker's voice $\hat{\mathbf{S}}^{\text{tgt,emo}}$. $\mathbf{e}$, which is an input of $\mathcal{M}_{\text{TTS}}$ for synthesizing speech expressively, is extracted from $\mathbf{S}^{\text{src,emo}}_{\text{ref}}$, since the target speaker's speech database does not contain expressive speech. By comparing the emotion embedding vectors, the TTS model can be updated even when there is no expressive speech of the target speaker.

IV. EXPERIMENTS AND ANALYSIS

A. EXPERIMENTAL SETUP

We prepared two Korean speech databases: one expressive and one neutral style speech. The expressive speech database for the source speaker consists of four emotion classes, namely neutral (NEU), joyful (JOY), angry (ANG), and sad (SAD), recorded by a single professional voice actress. The total amount of speech is about 11 hours. The scripts for each emotion class were different from the others. The target speaker's neutral speech database was recorded by another professional actress and consisted of approximately 1 hour of speech. Both databases were recorded at a 16 kHz sampling rate. The amounts of data for the training, validation, and test sets were set to 90%, 5%, and 5%, respectively. To enhance trainability, we excluded speech data longer than 10 seconds and trimmed the silence regions at the beginning and end of each sentence.

The network architectures and hyperparameters of the E2E-TTS module and the emotion encoder module were set to follow the original papers [8], [13], except that the emotion embedding vector was concatenated with the text embedding vectors. Consequently, the dimension of the audio embedding vectors was also adjusted to match that of the concatenated embedding vector. 80-dimensional mel-spectrograms were extracted at 12.5 ms frame intervals with 50 ms frame lengths from speech segments. All networks were trained using the Adam optimization algorithm [27] with a learning rate of 0.001 when training the baseline expressive TTS model, and 0.0001 when performing adaptations to obtain the target speaker's expressive TTS model.

B. EXPERIMENTAL RESULTS

To quantitatively verify the effectiveness of the proposed approach, we measured the equal error rate (EER) for speaker verification and the emotion classification accuracy for the target speaker's synthesized expressive-speech samples: lower EER indicates higher speaker similarity, and high classification accuracy indicates that the synthesized speech faithfully expresses emotion corresponding to the expressiveness style of the input emotion embedding vector. For the speaker verification task, we utilized the Kaldi toolkit [28]. We built a universal background model with data from 100 Korean speakers and enrolled an additional 11 speakers including the target speaker. For the emotion classification task, we used the naïve Bayes classification method [29]. The model parameters, such as the mean and variance of each class, were obtained from emotion embedding vectors extracted from the source speaker's recorded expressive speech. The emotion classification accuracy for the expressive speech synthesized by the source speaker's expressive TTS model was 93.5% (baseline).

Figure 4 shows evaluation results for the conventional and proposed methods as the adaptation progresses, where **emoTgtConv** and **emoTgtProp** are the target speaker's expressive speech synthesized from the TTS model adapted by the conventional (w/o emotion loss) and proposed approaches (w/ emotion loss), respectively. In both approaches, the synthesized speech became more similar to the target speaker's voice until the 40-th epoch, and emotion classification accuracy decreased. However, **emoTgtProp** maintained expressiveness characteristics with much higher accuracy than **emoTgtConv**, even though it changed voice characteristics slowly.

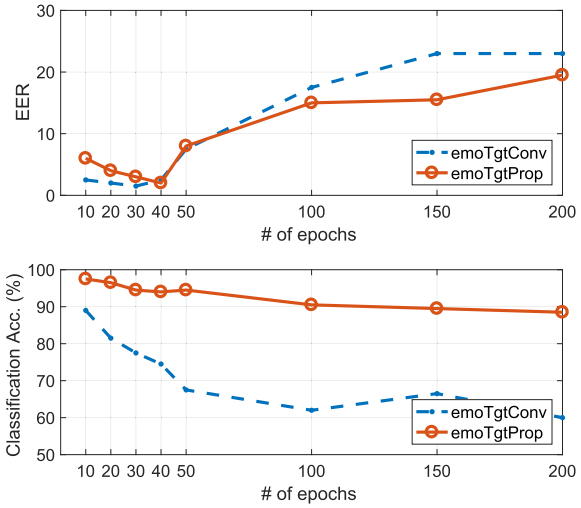

FIGURE 4. Speaker verification results (top) and emotion classification accuracy (bottom) with respect to number of epochs.

In Figure 5, emotion embedding vectors, extracted from expressive speech synthesized from **emoTgtConv** and **emoTgtProp**, were plotted using the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm [30]. Emotion embedding vectors of the same class grouped were more condensed for **emoTgtProp**, which means that much more distinct expressive speech was synthesized. On the other hand, the emotion embedding vectors for **emoTgtConv** showed an ambiguous boundary between emotions, especially NEU and ANG. This means that the synthesized expressive-speech did not faithfully express the emotion corresponding to the given input emotion embedded vector.

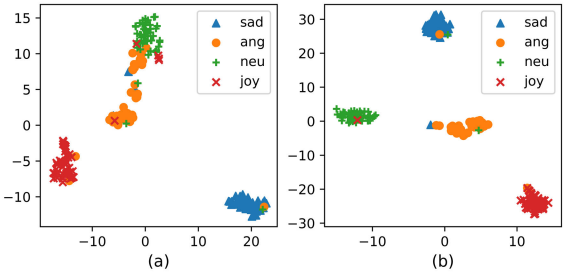

FIGURE 5. T-SNE plots of emotion embedded vectors. These were extracted from two types of synthesized expressive speech. (a) emoTgtConv (b) emoTgtProp.

C. SUBJECTIVE LISTENING TESTS

We conducted mean opinion score (MOS) tests on expressiveness, speaker similarity, sound quality, and naturalness for the synthesized speech[1] to evaluate the performance of the proposed approach. 20 native Korean speakers participated in the tests. A total of 20 sentences were randomly selected from the test set and speech samples were generated using each method identified above. Considering the variation of L_{TTS} and L_{emo}, we chose the TTS model trained up to the 50-th epoch. For the MOS tests, speech samples were synthesized using a neural vocoder, WaveGlow [31], [32], trained using both source and target speakers' speech databases.

To evaluate expressiveness, we asked participants to rate the expressiveness degree of a speech signal given the emotion label, using the following five responses: 1 = Absolutely different from the annotated emotion label; 2 = Ambiguous to the annotated emotion label; 3 = Slightly expressive; 4 = Very expressive; 5 = Extremely expressive. The source speaker's recorded expressive speech (**emoSrcRec**) was also compared to provide an upper bound. The top of Figure 6 shows that **emoTgtProp** scored slightly worse than **emoSrcRec** but was much more expressive than **emoTgtConv**, confirming the superiority of the proposed approach over the conventional one.

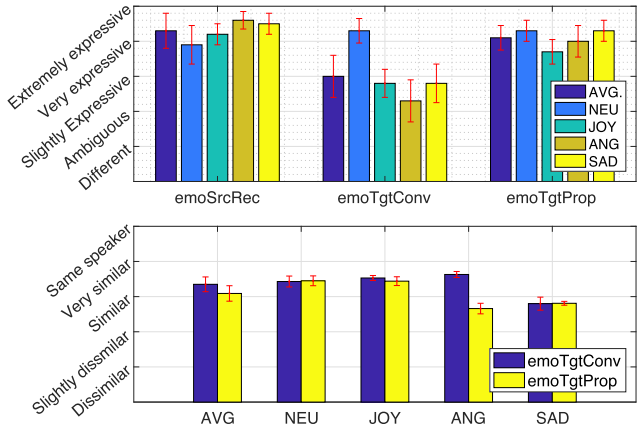

FIGURE 6. MOS test results for expressiveness in speech (top) and for speaker similarity (bottom).

To evaluate speaker similarity, participants were asked to rate voice similarity for the synthesized speech with the target speaker's voice, using a scale from 1 to 5: 1 = Dissimilar; 2 = Slightly dissimilar; 3 = Similar; 4 = Very similar; 5 = Absolutely the same speaker. Each synthesized speech sample was compared to recorded speech samples randomly selected from the target speaker's neutral speech database. We asked listeners to focus only on the degree of expressiveness and voice similarity, excluding speech content or

[1] https://nc-ai.github.io/speech/publications/emotion-tts/

emotional state differences. The bottom of Figure 6 shows that there were no significant differences between **emoTgtProp** and **emoTgtConv** except for ANG. In the case of ANG, because angry speech generated by **emoTgtConv** sounded like neutral speech, listeners felt that it was similar to the target speaker's voice. Both approaches received low scores for SAD. Sad speech has a low pitch and trembling characteristics, which results in missing speaker characteristics. Although we asked listeners to evaluate only voice similarity, excluding emotional state and content differences between speech samples, they were still somewhat influenced by the features of emotion.

Participants also rated the naturalness and sound quality of the synthesized speech samples. Table 1 shows that there were no significant differences between the two approaches.

TABLE 1. MOS test result for naturalness and sound quality with 95% confidence intervals.

	emoTgtConv	emoTgtProp
Naturalness	4.54 ±0.20	4.41 ±0.22
Sound quality	3.88 ±0.10	3.76 ±0.15

V. CONCLUSION

This paper proposed an effective emotion transplantation approach within an E2E-TTS framework for generating expressive speech for a target speaker whose speech database includes only neutral style speech. By alternately updating a pre-trained expressive TTS model in two directions, we not only generated the target speaker's voice characteristics but also successfully maintained the expressiveness characteristics of the pre-trained model. Thus, the proposed method successfully transplanted the ability to express appropriate emotions in the target speaker's voice. We verified the superior performance of the proposed approach through various quantitative and qualitative evaluations.

REFERENCES

[1] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 7962–7966.
[2] Y. Fan, Y. Qian, F. Xie, and F. K. Soong, "TTS synthesis with bidirectional LSTM based recurrent neural networks," in *Proc. Interspeech*, 2014, pp. 1964–1968.
[3] E. Song, F. K. Soong, and H.-G. Kang, "Effective spectral and excitation modeling techniques for LSTM-RNN-based speech synthesis systems," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 25, no. 11, pp. 2152–2161, Nov. 2017.
[4] Z.-C. Liu, Z.-H. Ling, and L.-R. Dai, "Statistical parametric speech synthesis using generalized distillation framework," *IEEE Signal Process. Lett.*, vol. 25, no. 5, pp. 695–699, May 2018.
[5] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," 2016, *arXiv:1609.03499*. [Online]. Available: http://arxiv.org/abs/1609.03499
[6] Y. Wang, R. J. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, Aug. 2017, pp. 4006–4010.
[7] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, S. O. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," in *Proc. ICLR*, 2018, pp. 1–16.
[8] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4784–4788.
[9] L. Chen, N. Braunschweiler, and M. J. F. Gales, "Speaker and expression factorization for audiobook data: Expressiveness and transplantation," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 23, no. 4, pp. 605–618, Apr. 2015.
[10] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," in *Proc. NIPS*, 2017, pp. 1–6.
[11] G. E. Henter, J. Lorenzo-Trueba, X. Wang, and J. Yamagishi, "Principles for learning controllable TTS from annotated and latent variation," in *Proc. Interspeech*, Aug. 2017, pp. 3956–3960.
[12] R. Skerry-Ryan, E. Battenberg, Y. Xiao, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80. PMLR, 2018, pp. 4693–4702.
[13] Y. Wang, D. Stanton, Y. Zhang, R. J. Skerry-Ryan, F. 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. Int. Conf. Mach. Learn. (ICML)*, vol. 80. PMLR, 2018, pp. 5180–5189.
[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, vol. 1. Cambridge, MA, USA: MIT Press, 1986, pp. 318–362.
[15] Y. Wang, R. J. Skerry-Ryan, Y. Xiao, D. Stanton, J. Shor, E. Battenberg, R. Clark, and R. A. Saurous, "Uncovering latent style factors for expressive speech synthesis," 2017, *arXiv:1711:00520*. [Online]. Available: https://arxiv.org/abs/1711.00520
[16] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Trans. Audio, Speech, Language Process.*, vol. 17, no. 1, pp. 66–83, Jan. 2009.
[17] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Multi-speaker modeling and speaker adaptation for DNN-based TTS synthesis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2015, pp. 4475–4479.
[18] Z. Wu, P. Swietojanski, C. Veaux, S. Renals, and S. King, "A study of speaker adaptation for DNN-based speech synthesis," in *Proc. Interspeech*, 2015, pp. 879–883.
[19] Y.-S. Joo, W.-S. Jun, and H.-G. Kang, "Efficient deep neural networks for speech synthesis using bottleneck features," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA)*, Dec. 2016, pp. 1–4.
[20] R. Doddipatla, N. Braunschweiler, and R. Maia, "Speaker adaptation in DNN-based speech synthesis using d-Vectors," in *Proc. Interspeech*, Aug. 2017, pp. 3404–3408.
[21] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo, J. Ferreiros, J. Yamagishi, and J. M. Montero, "Emotion transplantation through adaptation in HMM-based speech synthesis," *Comput. Speech Lang.*, vol. 34, no. 1, pp. 292–307, Nov. 2015.
[22] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *Proc. Interspeech*, 2015, pp. 1–5.
[23] K. Inoue, S. Hara, M. Abe, N. Hojo, and Y. Ijima, "An investigation to transplant emotional expressions in DNN-based TTS synthesis," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1253–1258.
[24] J. Parker, Y. Stylianou, and R. Cipolla, "Adaptation of an expressive single speaker deep neural network speech synthesis system," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 5309–5313.
[25] D. Griffin and J. Lim, "Signal estimation from modified short-time Fourier transform," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 32, no. 2, pp. 236–243, Apr. 1984.
[26] A. van den Oord *et al.*, "Parallel WaveNet: Fast high-fidelity speech synthesis," in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 80. PMLR, 2018, pp. 3918–3926. [Online]. Available: http://proceedings.mlr.press/v80/oord18a.html
[27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[28] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlícek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Veselý, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011, pp. 1–4.
[29] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification and Scene Analysis*, vol. 3. New York, NY, USA: Wiley, 1973.
[30] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[31] R. Prenger, R. Valle, and B. Catanzaro, "Waveglow: A flow-based generative network for speech synthesis," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 3617–3621.
[32] R. Prenger and R. Valle. (2018). *Waveglow*. [Online]. Available: https://github.com/NVIDIA/waveglow

YOUNG-SUN JOO (Member, IEEE) received the B.S. degree in electronics engineering from Kwangwoon University, Seoul, South Korea, in 2011, and the M.S. and Ph.D. degrees in electrical and electronics engineering from Yonsei University, Seoul, in 2013 and 2019, respectively. In 2017, she joined NCSOFT Corporation, Gyeonggi-do, South Korea. She served her internships at Microsoft Research Asia, Beijing, China, from 2014 to 2015. Her research interests include speech signal processing, speech analysis/synthesis, speech synthesis, voice conversion, and machine learning.

HANBIN BAE received the B.S. degree in computer engineering and the M.S. degree in electrical and computer engineering from Sungkyunkwan University, Suwon, South Korea, in 2016 and 2018, respectively. He worked on a project for the Electronics and Telecommunications Research Institute (ETRI) during the B.S. degree. He is currently working as a member of the Voice Conversion Team, Speech AI Lab., NCSOFT Corporation, Gyeonggi-do, South Korea. His research interests include speech synthesis, voice conversion, speaker verification, pattern classification, machine learning, and deep learning.

YOUNG-IK KIM received the B.S. and M.S. degrees in control and instrumentation engineering from Korea Maritime University, Busan, South Korea, in 1997 and 1999, respectively, and the M.S. and Ph.D. degrees in applied mathematics from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2003 and 2007, respectively. From 2008 to 2015, he was a Senior Member of Engineering Staff at the Department of Speech and Language Information, Electronics, and Telecommunications Research Institute (ETRI), Daejeon. He is currently a Lead Speech AI Researcher with the Speech Synthesis Team, Speech AI Lab., NCSOFT Corporation, Gyeonggi-do, South Korea. His research interests include speech processing and recognition, end-to-end speech synthesis, and deep learning algorithm optimization.

HOON-YOUNG CHO received the B.S., M.S., and Ph.D. degrees in computer science from the Korea Advanced Institute for Science and Technology (KAIST), Daejeon, South Korea, in 1995, 1998, and 2003, respectively. He was a Visiting Researcher at the Machine Learning and Signal Processing (MSLP) Laboratory, University of California at San Diego, San Diego, from 2003 to 2004. From 2004 to 2006, he worked as a Senior Researcher at the LG Advanced Research Institute, South Korea. He worked as a Senior Researcher at the Electronics and Telecommunication Research Institute (ETRI), Daejon, South Korea, from 2006 to 2012. In 2008, he co-founded a venture company, Enswers Inc., South Korea. He joined the company as the head of a research center in 2012 when the company was acquired by Korea Telecom Corporation, South Korea. In 2015, he joined a venture company, Beagle, as a CTO. Since 2016, he has been the Head of Speech AI Lab., NCSOFT Corporation, Gyeonggi-do, South Korea.

HONG-GOO KANG (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronics engineering from Yonsei University, Seoul, South Korea, in 1989, 1991, and 1995, respectively. He was a Senior Member of Technical Staff at AT&T Labs Research from 1996 to 2002. In 2002, he joined the Department of Electrical and Electronics Engineering, Yonsei University, where he is currently a Professor. His research interests include speech signal processing, audio–visual signal processing, and machine learning.