Journal of Computer Science and Cybernetics, V.38, N.4 (2022), 369–379
DOI no 10.15625/1813-9663/18236

EMOTION TRANSPLANTATION APPROACH FOR VLSP 2022

THANG NGUYEN VAN[1], LONG LUONG THANH[1], HUAN VU[2*]

[1]*Innovation Center, VNPT-IT, 57 Huynh Thuc Khang Street, Dong Da District, Ha Noi, Viet Nam*
[2]*University of Transport and Communications, 3 Cau Giay Street, Lang Thuong Ward, Dong Da District, Ha Noi, Viet Nam*

Abstract. Emotional speech synthesis is a challenging task in speech processing. To build an emotional Text-to-speech (TTS) system, one would need to have a quality emotional dataset of the target speaker. However, collecting such data is difficult, sometimes even impossible. This paper presents our approach that addresses the problem of transplanting a source speaker's emotional expression to a target speaker, one of the Vietnamese Language and Speech Processsing (VLSP) 2022 TTS tasks. Our approach includes a complete data preprocessing pipeline and two training algorithms. We first train a source speaker's expressive TTS model, then adapt the voice characteristics for the target speaker. Empirical results have shown the efficacy of our method in generating the expressive speech of a speaker under a limited training data regime.

Keywords. Emotional speech synthesis, emotion transplantation, text-to-speech.

1. INTRODUCTION

Traditional TTS systems aim to synthesize human-like speech from texts. It is an important feature that is utilized widely in many applications such as virtual assistance, virtual call centers,... Thanks to recent advances in deep learning, models such as Tacotron 2 [1], Fastspeech 2 [2], and VITS [3] have successfully shown to be able to generate high-quality speech.

To expand further, researchers have tried to develop TTS models that are able to include emotional expression to generat speech [4–8]. These approaches often rely on an emotional speech dataset from the target speaker, along with emotion embedding techniques that help the model learn different characteristics of each emotion. However, such a dataset is not always available for every speaker, and building such a dataset for a chosen speaker is an extremely challenging task. A speaker, even when demanded, might be unable to express certain emotions naturally during the recording process.

To tackle this problem, another approach is widely studied, namely the emotion transplantation approach. This approach aims to transfer the ability of a model to express emo-

*Corresponding author.
E-mail addresses: thangnv97@vnpt.vn (T. Nguyen Van); longlt97@vnpt.vn (L. Luong Thanh); huan.vu@utc.edu.vn (H.Vu).

tions from one speaker to another. In this way, one would only need a quality emotional dataset from the source speaker, along with a traditional (neutral) speech dataset from the target speaker. The key obstacle of this approach is how to adjust the model so that it replicates the target speaker's voice while maintaining the capacity to express the desired emotions. Some adaptation approaches have been proposed recently, for example [9–12].

This paper presents our approach to the emotion transplantation challenge in VLSP 2022. Our approach contains a complete data pre-processing pipeline, the details of our model architecture, the training process of a baseline model, and the adaptation process to the target speaker. Several experiments were also conducted to show the quality of our approach.

2. DATA PRE-PROCESSING

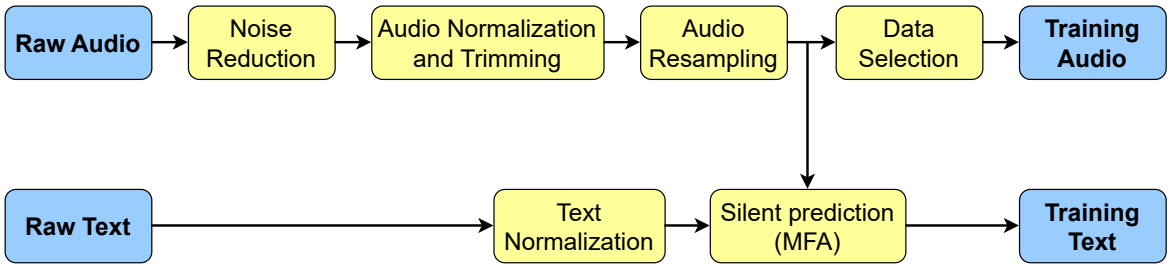

Figure 1: Our data pre-processing pipeline

Two datasets are provided for this task. The first dataset is an emotional dataset crawled from a television (TV) series and interviews. In this set, audio files from speaker A are accompanied by the corresponding transcript and an emotion label. The second dataset is a neutral speech dataset from another speaker B (only audio samples and transcripts are provided). The aim is to build a system that is able to generate speech with voices from the speaker B, and that has an additional user-defined emotion label.

The second dataset is of quite high quality since it is crawled from ebooks. However, many problems were found in the first dataset, including:

- Multiple files have background noise such as background music, traffic noise, laughing noise, crying noise, and voices from other speakers.
- Files originating from the interviews have different speaking styles compared to audio from the TV series.
- The speaking rate and prosody are inconsistent.
- Mislabeled text.
- Emotion labels are often ambiguous, especially between "happy" and "neutral".

To overcome these issues, we have applied several data pre-processing techniques as follows.

2.1. Noise reduction

Due to the high amount of noise in the audio utterances, a noise reduction technique named Music Source Separation [13] is applied first. This technique separates the audio into multiple sources. To retrieve vocals, we use their pre-trained MobileNet Subbandtime model with 2 input channels and 2 output channels.

As Music Source Separation may fail at times when dealing with audio samples that has a high noise ratio, the FullSubNet model [14] is used to enhance speech in these files.

2.2. Audio normalization

Another important factor that may harm the performance of a speech synthesis model is speech volume, which differs from one file to another. The difference is higher in the provided dataset due to its origin in a TV series. This issue was dealt with by simply normalizing the audio files to a top level of 20dB.

Attention-based speech synthesis models often rely on the linearity between texts and audio. Audios that have silences at the beginning and the end make it harder for the model to correctly predict the first phonemes. Hence, it is also important to trim these silences in the audio files.

2.3. Punctuation prediction and text normalization

After noise reduction and audio normalization, we also need to deal with text transcripts. First, the transcripts provided are mostly punctuation-free. However, acoustic models build upon punctuation to predict silences in the output audio. To add punctuation in transcripts, we use the Montreal Forced Aligner (MFA) tool [15], an open-source tool that learns alignment between text and audio. MFA is trained on the clean dataset to detect silences and add punctuation. We decided to add a "," for a silence between $0.2 - 0.4$ seconds and a "." for a silence of more than 0.4 seconds.

Moreover, some minor text normalization techniques are also used to facilitate training. Some English words presented in the text are replaced by Vietnamese equivalents (e.g., casting $\rightarrow$ cát sx tinh). Numbers are also replaced by their transcript (e.g., 2 $\rightarrow$ hai).

2.4. Data selection

Finally, we found out that audio also have different speaking rates, depending on the emotion. An inconsistent speaking rate in training data may lead to an inferior model. After carefully examining the training set, we decided to use the words per second (WPS) value to measure speaking rate and filtered out utterances in which the either spoke very slowly (WPS < 3) or very fast (WPS > 5.5).

We also discover that lots of audios have screaming or laughing voices. These audios are commonly very short ones. Therefore, audios that are less than 0.8 seconds in length are deleted from the training set. Finally, our team members reviewed the cleaned audio files and voted for removing or re-labeling the files that have high emotional ambiguity. Since the number of training samples is limited, during this selection process, we intend to keep as much audio as possible. Only samples which are agreed by all members are removed.

The final data pre-processing pipeline is presented in Figure 1. After pre-processing the original dataset of 8,800 files, we achieved a cleaned dataset of 8,333 files. The total length is reduced from 3.85 hours to about 3.2 hours. The final statistics are presented in Table 1.

Table 1: Data pre-processing statistics

Emotion	**Original Data**	**Cleaned Data**
Neutral	6,620	6,199
Angry	1,148	1,050
Sad	475	532
Happy	557	552
Total	8,800 (3.85h)	8,333 (3.2h)

For the second dataset, since it is of good quality, we only apply Audio Normalization and Data Selection. We resampled both datasets to 22,050Hz instead of 44,000Hz due to the limited timeline of the challenge and our computational capabilities.

3. PROPOSED EMOTION TRANSPLANTATION APPROACH

An adaptation method which can transfer expressive speech synthesis from one speaker to another [12] is used for this task. This can be highly effective when we have an emotional dataset from a source speaker and a neutral dataset from a target speaker. This approach ensures that the system provides speech with voices from the target speaker, and with the desired emotion. It is divided into two phases. In the first phase, we aim to build a baseline expressive TTS model for the source speaker using the first dataset. Following that, we propose an adaptation technique to finetune this model to the target speaker's acoustic characteristics using the second dataset. Details of the model architecture, the training process of the baseline model and the adaptation technique are described in the following subsections.

3.1. Architecture of our model

Our end-to-end expressive TTS model has two main components: (1) an emotion encoder using the Global style token (GST) approach [6] and (2) an end-to-end TTS model that synthesize emotional speech from input texts and expressive condition vectors from the GST module. The overall architecture is shown in Figure 2. The details of each module are described as follows:

A. Emotion encoder

The source speaker dataset contains four emotional labels: "Neutral", "Angry", "Sad" and "Happy". However, the target speaker dataset only consists of one emotion: "Neutral". Therefore, including a lookup table emotion embedding is ineffective for this task since the embedding stays the same for the second dataset. In our approach, we propose to use a global style token approach as an emotion encoder for both datasets. GST-based TTS system allows the model to synthesize speech with a style transferred from a reference audio. This reference audio provides additional information for our model while training and generating speech.

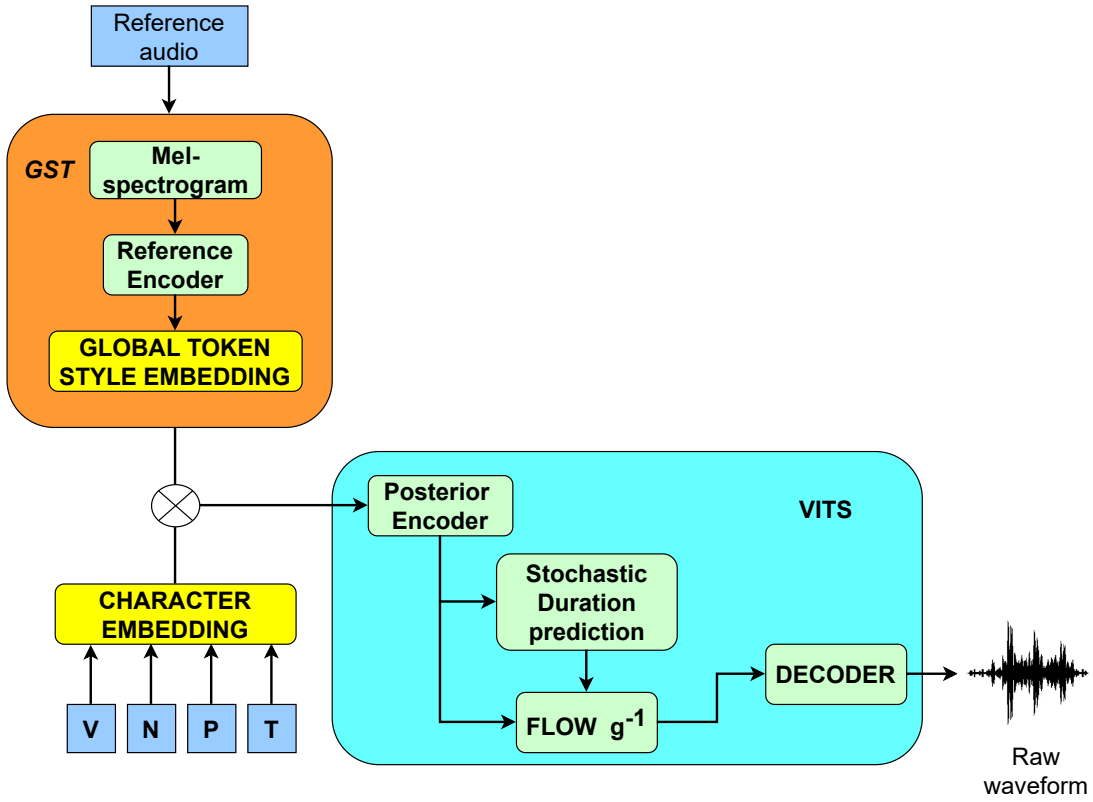

Figure 2: Model architecture

The GST module takes as input a reference audio, followed by a reference encoder with several convolutional layers and a Gated Recurrent Units (GRU) network [16] that produces a fixed-length vector that carries the acoustic features of the reference audio. This vector is then passed through a multihead attention layer to provide the final embedding vector for the reference audio.

B. End-to-end TTS model

In our proposed approach, we use VITS [3] as our backbone model. VITS is an end-to-end TTS model that takes advantage of parallel training and computing. Generally, end-to-end models are shown to underperform compared to two-staged models that consist of an acoustic model and a vocoder. However, since its introduction, VITS, which uses several techniques such as variational autoencoder, stochastic duration prediction, and adversarial training, has been shown to produce comparable audio quality to baseline two-staged models, while still keeping the advantages of end-to-end models.

By using a variational autoencoder, VITS connects the two components of two-staged models into one single component. Furthermore, to address the one-to-many problem of a speech synthesis system, VITS introduces the stochastic duration prediction mechanism. This mechanism allows the model to learn and predict the speaking rate from input text more easily, leading to more natural speech. Finally, adversarial training is introduced to control the training process of the generator model.

3.2. Training the source speaker's expressive TTS model

Algorithm 1: The expressive TTS training algorithm

Data: $(\mathbf{X}, \mathbf{S}^{src,emo}) \in D_{src,emo}$
while $\mathcal{M}_{TTS}$ *not converged* **do**
  $\mathbf{e} \leftarrow \mathbf{GST}_{emo}(\mathbf{S}^{src,emo})$
  $\hat{\mathbf{S}}^{src,emo} \leftarrow \mathcal{M}_{TTS}(\mathbf{X},\mathbf{e})$
  $\mathcal{L}_{TTS} \leftarrow \mathcal{L}_{TTS}(\mathbf{S}^{src,emo}, \hat{\mathbf{S}}^{src,emo})$
  Update $\mathcal{M}_{TTS}$
end

Algorithm 1 describes the training process of the baseline expressive TTS model using the source speaker emotional speech dataset. The training input requires the input text $\mathbf{X}$ and the referenced expressive speech $\mathbf{S}^{src,emo}$ drawn from the training set $D_{src,emo}$. First, the output of the GST module $\mathbf{e}$ from $\mathbf{S}^{src,emo}$ is computed. $\mathbf{e}$ is then passed as input, along with $\mathbf{X}$ to train a VITS model normally. The aim is to minimize a loss between the referenced expressive speech $\mathbf{S}^{src,emo}$ and the synthesized expressive speech $\hat{\mathbf{S}}^{src,emo}$. This loss is computed using the original VITS loss as follows

$$\mathcal{L}_{TTS}(\mathbf{S}, \hat{\mathbf{S}}) = \mathcal{L}_{recon} + \mathcal{L}_{kl} + \mathcal{L}_{dur} + \mathcal{L}_{adv}(G) + \mathcal{L}_{adv}(G).$$

3.3. Adapting voice characteristics for the target speaker

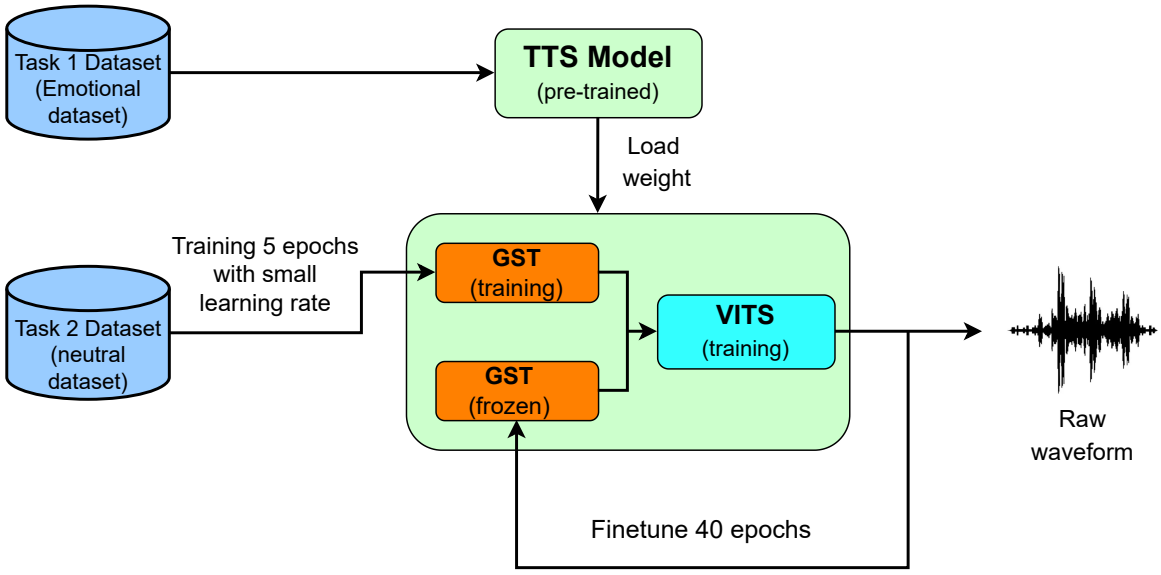

Figure 3: Voice adaptation process.

Figure 3 and Algorithm 2 describe the adaptation process. This update is done using the second dataset with a much lower learning rate. It is observed that in the second phase, finetuning directly from the pre-trained TTS model $\mathcal{M}_{TTS}$ leads to a trade-off between the

Algorithm 2: The emotion transplantation algorithm.

Data: $(\mathbf{X}, \mathbf{S}^{tgt,neu}) \in D_{tgt,neu}$
Pretrained model: $\mathcal{M}_{TTS}$
Step 1: First 5 epochs
Step 2: Following 40 epochs
while $\mathcal{M}_{TTS}$ *not converged* **do**
  if *Step == 2* **then**
    Freeze emotion encoder (GST)
  end
  for $i \leftarrow 1, \mathcal{N}_{tgt}$ **do**
    $\mathbf{e} \leftarrow \mathbf{GST}_{emo}(\mathbf{S}^{tgt,neu})$
    $\hat{\mathbf{S}}^{tgt,emo} \leftarrow \mathcal{M}_{TTS}(\mathbf{X},\mathbf{e})$
    $\mathcal{L}_{TTS} \leftarrow \mathcal{L}_{TTS}(\mathbf{S}^{tgt,neu}, \hat{\mathbf{S}}^{tgt,emo})$
    Update $\mathcal{M}_{TTS}$
  end
end

emotional quality (i.e., the quality to express the desired emotion), and the target speaker's voice characteristics. In the first 5 epochs, the ability to express emotion is quite good, but the generated voice lacks the characteristic of the target speaker. To tackle this issue, we divide the finetuning process into two small steps. In the first 5 epochs, we finetuned the entire pre-trained TTS model with the target speaker data $D_{tgt,neu}$. After 5 epochs, we freeze the weights of the GST module $\mathbf{GST}_{emo}$, and finetune the rest of the model for an additional 40 epochs. The finetuning loss is the TTS loss $\mathcal{L}_{TTS}(\mathbf{S}^{tgt,neu}, \hat{\mathbf{S}}^{tgt,emo})$ between the referenced neutral speech of the target speaker and the synthethized emotional speech.

4. EXPERIMENTAL SETUP

The cleaned dataset is divided into a training set, a validation set and a test set with ratios of 90%, 5%, and 5%, respectively.

The same network architecture and hyperparameters are used for the End-to-end TTS model and the GST module as in its original paper [3,6]. 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 12.5ms frame intervals with 50ms frame length from speech segments. All networks are trained using an Adam optimizer [17] with a learning rate of 2×10^{-4} while training the baseline expressive TTS model (source speaker's model), and 4×10^{-5} while finetuning (adapting) for the target speaker. All training is done using in NVIDIA V100 GPU with a batch size of 32. Training details are shown in Table 2.

Table 2: Training details

Steps	**LR**	**Epochs**	**Time**
Baseline	2e-4	300	4 days
Finetuning			
Step 1	4e-5	5	2h
Step 2	4e-5 (GST frozen)	40	16h

5. EXPERIMENTAL RESULTS

The synthesized speech in this challenge needs to have the voice characted of the target speaker while expressing different emotions. Therefore, two different experiments are conducted. The first experiment is a similarity test to measure the difference between the synthesized voice and the target speaker's voice. The second test is an emotion quality test. Finally, the VLSP 2022 official score is presented.

5.1. Similarity test

To measure the similarity between the voices of the synthesized speech and the target speaker, 20 files were chosen from the target speaker's test set and speeches with all four emotions were generated. The average cosine similarity of each synthesized speech and each original speech is shown in Table 3.

Table 3: Cosine similarity of the generated and original speech

Emotion	**Cosine similarity**
Neutral	0.76
Angry	0.54
Sad	0.77
Happy	0.71

We notice that the speeches are quite similar. The cosine similarity score is about 0.75 in most cases, except for the "Angry" emotion. "Angry" audio files generated by the proposed models usually have a much higher pitch compared to "neutral" files, which leads to a lower similarity score.

5.2. Emotional test

The second test which was conducted was an emotional test. 10 participants will listen to 60 audios selected randomly in the source speaker's test set (15 audios for each emotion). Participants are asked to rank the emotional expression of audio files with 5 different levels, which 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 final results are shown in Table 4.

Table 4: Our conducted emotional test score

Emotion	Original	Synthesised
Neutral	4.58	3.83
Angry	4.31	3.33
Sad	4.27	3.48
Happy	3.88	2.69

Based on our conducted experiments, our model has successfully expressed the chosen emotion. However, due to the unbalanced dataset and the ambiguity between "Happy" and "Neutral" in the training set, which has not been dealt with completely during the data pre-processing phase, the "Happy" emotion has not achieved satisfactory results.

5.3. Final VLSP 2022 score

Table 5: Our VLSP 2022 results

Test	Score
Naturalness test (Mean Opinion Score - MOS)	3.762
Intelligibility (Syllable Error Rate - SUS)	25.80%
Speaker Similarity	2.286

Table 5 presents the final VLSP 2022 score of our team. Due to the unreliability of the emotion evaluation results, the organizers decided not to include this score in the final results. The final score includes the MOS score (ranging from 1-5 with 5 representing a voice that is identical to a human). The second score is the intelligibility test, measured using syllable error rate. Finally, to compare the synthesized voice and the target speaker's voice, a speaker similarity score is measured (ranging from 1-4). A speaker similarity of 4 represents a voice that is identical to the original voice. Our naturalness and speaker similarity scores rank second, while our intelligibility score ranks third among the teams. Our approach achieves the second overall ranking, confirming the effectiveness of the chosen method.

6. CONCLUSIONS

This paper details our approach to the VLSP 2022 TTS task 2, namely the emotion transplantation task. Our approach includes a complete data pre-processing pipeline and our model architecture. The training process and the hyperparameters used in the expriments are also presented in detail. Results show that our model successfully generates quality speech with voice characteristics from the target speaker while maintaining the ability to express emotions. In future work, more transplantation strategies will be explored, especially when dealing with unbalanced datasets and the ambiguity between emotions.

ACKNOWLEDGMENT

This work was supported by the Innovation Center, VNPT-IT, Vietnam Posts and Telecommunications Group.

REFERENCES

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018.

[2] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *ArXiv*, vol. abs/2006.04558, 2021.

[3] J. Kim, J. Kong, and J. Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[4] Y. Lee, A. Rabiee, and S.-Y. Lee, "Emotional end-to-end neural speech synthesizer," *ArXiv*, vol. abs/1711.05447, 2017.

[5] Y. Wang, D. Stanton, Y. Zhang, R. J. 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 *ICML*, 2018.

[6] O. Kwon, I. Jang, C. H. Ahn, and H.-G. Kang, "An effective style token weight control technique for end-to-end emotional speech synthesis," *IEEE Signal Processing Letters*, vol. 26, pp. 1383–1387, 2019.

[7] S. Um, S. Oh, K. Byun, I. Jang, C. H. Ahn, and H.-G. Kang, "Emotional speech synthesis with rich and granularized control," *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7254–7258, 2020.

[8] N. Tits, K. E. Haddad, and T. Dutoit, "Exploring transfer learning for low resource emotional tts," in *IntelliSys - Intelligent Systems Conference*, 2019.

[9] J. Lorenzo-Trueba, R. Barra-Chicote, R. San-Segundo-Hernández, J. Ferreiros, J. Yamagishi, and J. M. Montero-Martínez, "Emotion transplantation through adaptation in hmm-based speech synthesis," *Computer Speech & Language*, vol. 34, pp. 292–307, 2015.

[10] Y. Ohtani, Y. Nasu, M. Morita, and M. Akamine, "Emotional transplant in statistical speech synthesis based on emotion additive model," in *INTERSPEECH Conference*, 2015.

[11] J. Parker, Y. Stylianou, and R. Cipolla, "Adaptation of an expressive single speaker deep neural network speech synthesis system," in *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing*, 2018, p. 5309–5313.

[12] Y.-S. Joo, H. Bae, Y.-I. Kim, H.-Y. Cho, and H.-G. Kang, "Effective emotion transplantation in an end-to-end text-to-speech system," *IEEE Access*, vol. 8, pp. 161 713–161 719, 2020.

[13] Q. Kong, Y. Cao, H. Liu, K. Choi, and Y. Wang, "Decoupling magnitude and phase estimation with deep resunet for music source separation." in *ISMIR - International Society for Music Information Retrieval Conference*, 2021.

[14] X. Hao, X. Su, R. Horaud, and X. Li, "Fullsubnet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," *ICASSP - IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6633–6637, 2021.

[15] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *INTERSPEECH Conference*, 2017.

[16] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder–decoder approaches," in *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 2014.

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2015.

Received on April 09, 2023
Accepted on October 06, 2023