An Effective Style Token Weight Control Technique for End-to-End Emotional Speech Synthesis

Ohsung Kwon, Inseon Jang, ChungHyun Ahn, and Hong-Goo Kang

***Abstract*—In this letter, we propose a high-quality emotional speech synthesis system, using emotional vector space, i.e., the weighted sum of global style tokens (GSTs). Our previous research verified the feasibility of GST-based emotional speech synthesis in an end-to-end text-to-speech synthesis framework. However, selecting appropriate reference audio (RA) signals to extract emotion embedding vectors to the specific types of target emotions remains problematic. To ameliorate the selection problem, we propose an effective way of generating emotion embedding vectors by utilizing the trained GSTs. By assuming that the trained GSTs represent an emotional vector space, we first investigate the distribution of all the training samples depending on the type of each emotion. We then regard the centroid of the distribution as an emotion-specific weighting value, which effectively controls the expressiveness of synthesized speech, even without using the RA for guidance, as it did before. Finally, we confirm that the proposed controlled weight-based method is superior to the conventional emotion label-based methods in terms of perceptual quality and emotion classification accuracy.**

***Index Terms*—Emotional speech synthesis, end-to-end, text-to-speech, global style token, emotional vector space, emotion weight values.**

I. Introduction

A TEXT-TO-SPEECH (TTS) technique that generates an artificial speech waveform from a given text is essential for natural human-computer interaction systems, such as automatic announcements, conversational agents, and voice assistance for the impaired. For the last few decades, unit selection synthesis and statistical parametric speech synthesis methods have largely dominated the speech synthesis community [1]–[7]. However, an end-to-end (E2E)-TTS synthesis system that generates high-quality speech in a single unified deep neural network has recently gained much attention [8]–[13].

With the great success of the E2E-TTS paradigm, e.g., the Tacotron framework, research interests have been expanding to synthesize *expressive* or *emotional* speech [14]–[19]. In particular, the global style token (GST) on the Tacotron framework, i.e., the GST-Tacotron system, is a representative unsupervised style-modeling example for expressive speech synthesis in the E2E-TTS framework [16]. The GST is a high-dimensional embedding vector that implicitly contains acoustic information, such as prosody, duration, and pitch, from the training speech dataset. The weighted sum of the GSTs, i.e., style embedding vector, represents the target speaker's speaking style, and it is embedded in the encoded text sequence of the Tacotron to generate expressive speech.

In the inference phase, there are two types of style embedding generation methods: (1) the reference audio (RA)-based style transfer method and (2) the weighting-based style control method [16]. The RA-based transfer method first selects an RA sample with a specific speaking style, and then the RA is fed into the trained GST network to extract a style embedding vector. And, the weighting-based control method generates style embedding through a linear combination of weight values for the trained GSTs to control the style of synthesized speech. In the RA-based transfer method, however, determining an appropriate RA to obtain high-quality and faithful expressiveness in the synthesis process remains a challenge. In addition, no one has yet determined how to effectively control the weights of the GSTs to represent a specific speaking style.

In this letter, we propose a Tacotron2-based speech synthesis system that not only generates various types of high-quality emotional speech but also flexibly controls the type of emotion. In particular, we focus on analyzing the relationship between GSTs and three types of emotional representation. Our contributions are as follows: (1) To alleviate the aforementioned problems, we propose an algorithm to obtain weight values to generate various types of high-quality emotional speech; the weight values are defined by investigating the distribution of each emotion in the emotional vector space, e.g., GSTs. (2) By analyzing the characteristics of each GST in terms of a specific emotion, we verify the feasibility of using specific weight values to flexibly control each type of emotion. (3) Comparing our proposed method with the conventional emotion label-based method, we demonstrate the effectiveness of the former.

The frameworks of Tacotron and global style token are briefly explained in Section II. Section III describes the RA-based and the proposed controlled weight (CW)-based methods in detail. Then, experiments and conclusion are followed in Sections IV and V, respectively.

II. Background

A. Tacotron Speech Synthesis Framework

Tacotron [12], [13] is a deep learning-based E2E-TTS synthesis framework that generates a time-domain speech waveform

Manuscript received May 26, 2019; revised July 9, 2019; accepted July 13, 2019. Date of publication July 29, 2019; date of current version August 8, 2019. This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (2019-0-00447, Development of emotional expression service to support hearing/visually impaired). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Odette Scharenborg. *(Corresponding author: Hong-Goo Kang.)*

O. Kwon and H.-G. Kang are with the Department of Electrical and Electronics, Yonsei University, Seoul 03722, South Korea (e-mail: osungv@dsp.yonsei.ac.kr; hgkang@yonsei.ac.kr).

I. Jang and C. Ahn are with the Media Research Division, ETRI, Daejeon 34129, South Korea (e-mail: jinsn@etri.re.kr; hyun@etri.re.kr).

Digital Object Identifier 10.1109/LSP.2019.2931673

from an input text sequence, which is based on an encoder-decoder model with an attention mechanism [20]–[24]. We exploit the Tacotron2 synthesis system [13] as a baseline E2E-TTS framework for high-quality speech synthesis. The system comprises a spectrogram prediction network and a waveform generation network that are trained separately.

The spectrogram prediction network is trained to find a mapping rule between the input text sequence and target output mel-spectrogram, using a minimization of mean squared error (MMSE) between the target reference and generated mel-spectrogram. Specifically, input text sequence is first converted into character embedding sequence by a character embedding module, and they are encoded into transcript embedding, using a transcript encoder. They are then passed through the location-sensitive attention network [20] for alignment with target acoustic sequence, i.e., the frame-level mel-spectrogram. After that, a decoder generates multiple frames of mel-spectrogram in every decoding step by utilizing a spectrogram obtained at the last decoding stage and a context vector from an output of the attention network. In the decoding process, one projection layer predicts a stop token to determine the completion time of decoding, and the other projection layer, followed by a post-net, generates the mel-spectrogram where the post-net predicts the residual component of the generated spectrogram to further improve prediction accuracy.

The waveform generation network, i.e., WaveNet [8], is trained to autoregressively generate the speech waveform with the predicted mel-spectrogram as a conditional input, using a criterion of minimizing a negative log-likelihood. Specifically, the WaveNet vocoder predicts the gain, mean, and log-scale parameters of mixtures by assuming that speech samples can be represented by a discretized mixture of logistic distributions.

In the speech synthesis phase, an arbitrary text sequence is fed into the spectrogram prediction network to predict the mel-spectrogram, then the spectrogram is used as a conditional input of the WaveNet vocoder to generate a speech waveform.

B. GST-Based Style Modeling

GSTs are a set of embedding vectors that contain the target speaker's prosody and speaking style information. The idea of the GST was first proposed in the original Tacotron framework, and the GST network was trained without using an explicit style label [12], [16].

In this approach, the GST network located in the encoding framework constructs GST embeddings such that the weighted sum of the embeddings presents style information. In the training phase, the prosody encoder first encodes the mel-spectrogram of input RA signal to extract the hidden prosodic representation of the RA, i.e., the prosody embedding vector. Then, the multi-headed attention network [25] measures the similarity between the prosody embedding and each GST, and predicts the style embedding vector which is the weighted sum of the GSTs. Note that the multi-headed attention network learns how to control the contribution of each GST via weight values. Finally, the estimated style embedding is broadcasted and combined with the transcript embeddings generated by the transcript encoder. The transcript embeddings fully represent the linguistic information of the speech signal, and the style embedding represents the acoustic information of the reference signal. In the inference phase, the style embedding can be generated by feeding the RA signal into the trained GST network or controlling the weight values related to each trained GST for weighted sum.

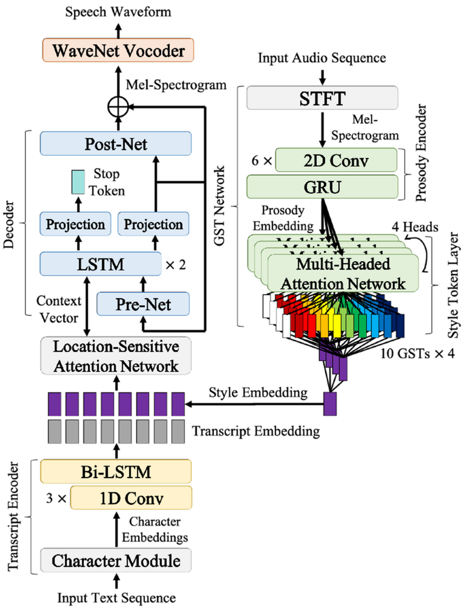

Fig. 1. A block diagram of the GST-Tacotron2.

However, the criterion or quality impact of selecting the RA signal in the dataset and controlling the GST weight values to a specific speaking style has not been studied thoroughly. In the next section, we introduce a GST-Tacotron2 framework for emotional speech synthesis and describe the limitation of the RA-based method in detail. To address this issue, we investigate the relationship between GST and various types of emotions in the emotional vector space; we then propose a guideline to effectively control the weight values.

III. Emotion Analysis Based on GST-Tacotron2

A. GST-Tacotron2

Fig. 1 presents the detailed structure of the GST-Tacotron2. In the training phase, the GST network and the spectrogram prediction network are jointly trained via MMSE between the target and the predicted spectrogram. Transcript and style embeddings are generated by the transcript encoder and the GST network, respectively. They are then concatenated to provide a joint style-transcript embedding for the decoding network. The joint embedding vector is then passed through the decoder to predict the mel-spectrogram used to synthesize a speech waveform via the WaveNet vocoder.

B. Emotional Speech Synthesis With RA-Based Method

Our previous research showed the feasibility of using GST-based emotional speech synthesis with the RA-based transfer method [26]. However, there was no criterion to determine the RA. For example, the quality and characteristics of synthesized speech varied in terms of the selected RA, but it was not easy to set a relevant criterion for the selection process.

To verify the performance variation in terms of perceptual quality and emotional expressiveness caused by choosing differentRA samples, we generated three types of emotional speech, using four different RA samples per emotion randomly selected in the training dataset.[1]

[1] Experimental setups for this preliminary test are the same as those summarized in Section IV.

TABLE I
MOS WITH A 95% CONFIDENCE INTERVAL FOR FOUR DIFFERENT RA SAMPLES

RA	Happy	Sad	Neutral
A	4.30±0.13	4.31±0.22	3.98±0.14
B	4.41±0.17	4.29±0.30	3.93±0.11
C	4.25±0.15	4.28±0.24	3.95±0.13
D	4.42±0.13	4.25±0.28	4.03±0.16

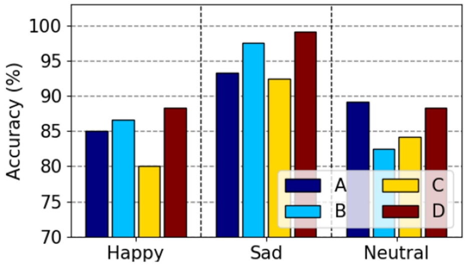

Fig. 2. Emotion classification results (%) with four different RA samples.

TABLE II
P-VALUES FOR FOUR DIFFERENT RA SAMPLES. THE NUMBERS REPRESENTS HAPPY, SAD, AND NEUTRAL, RESPECTIVELY

RA	B	C	D
A	**0.041**, 0.838, 0.504	0.463, 0.738, 0.638	0.089, 0.551, 0.539
B	-	**0.037**, 0.910, 0.830	0.908, 0.490, 0.317
C	-	-	**0.010**, 0.713, 0.312

Table I and Fig. 2 depict the mean opinion score (MOS) and the accuracy of emotion classification tests, respectively; in the two results, A, B, C, and D denote four different RA samples per emotion. Also, Table II summarizes p-values for every combination of four different RA samples; the values less than 0.05 are highlighted with bold font. The results clearly show the variations in respect to emotional expressiveness caused by the selection of RA samples, but it is difficult to set a criterion to appropriately select an RA signal. To ameliorate the problem, we investigate a method to effectively control the weights of GSTs. We call it the *controlled weight (CW)-based method*.

C. Emotion Embedding With the GST Framework

Emotional information contained in speech signals can be represented by the distribution of style embedding if the GST network is trained with an emotional speech corpus. We thus redefine the output of the GST network as *emotion embedding*. Especially, the GSTs capture distinct emotional information of input RA signals with the help of the prosody encoder and the multi-headed attention network in the training phase.

Fig. 3 depicts the sample distribution of each emotion embedding projected in the two-dimensional (2D) space, using the t-distributed stochastic neighbor embedding (t-SNE) algorithm [27], which visualizes the rough characteristic of the 256-dimensional emotion embedding vectors to the 2D representation. Three different colors in the figure denote three distinct *emotion clusters*: sad, happy, and neutral emotions. Each emotion embedding sample of the figure was generated by the RA-based method, using the emotional speech signals in the training dataset. It is true that the GST network has a capability of distinguishing different types of emotional classes if we can utilize the distribution of emotion embeddings in the emotional vector space.

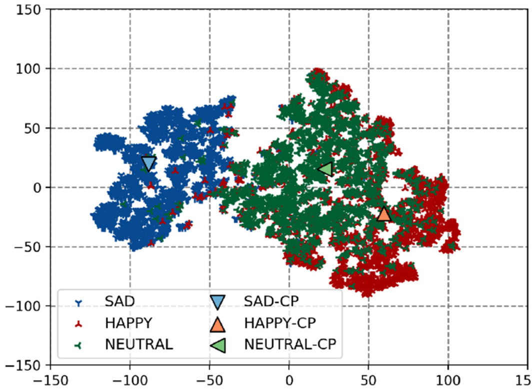

Fig. 3. Visualization of emotion embedding, using the t-SNE. The suffix "CP" denotes the center point of emotion cluster.

TABLE III
NUMBER OF UTTERANCES IN DIFFERENT SETS AND EMOTIONS FOR EMOTIONAL SPEECH CORPUS

Emotion	Training	Test
Happy	4135 (3.6 h)	433 (0.3 h)
Sad	3986 (3.6 h)	388 (0.3 h)
Neutral	2015 (3.6 h)	203 (0.3 h)
Total	10136 (10.8 h)	1024 (0.9 h)

Therefore, we conclude that it is possible to flexibly control the emotions of the GST-Tacotron2 synthesis system if we establish the relationship between the GSTs and each emotion. A simple method is to set each emotion embedding by a weighted sum of the trained GSTs. We call it an *emotion weight values*, and the weight values are set to the centroid of each emotion cluster. Note that the centroid of each emotion category is computed by the element-wise average of emotion embeddings included in each emotion cluster. In Fig. 3, "CP" denotes the center point of each emotion cluster. Consequently, weight values of each emotion are used to control the emotional expressiveness of synthesized signals.

IV. EXPERIMENTS

A. Database, Feature, and Character Module

An emotional speech corpus recorded by a professional Korean female speaker was used for the experiments, and Table III shows the number of utterances in the speech corpus. Note that the speaker was guided to record the emotions clearly and the recorded corpus was examined by a sound engineer. The speech signals were sampled at 24 kHz, and each sample was quantized by 16 bits. The frame and shift sizes were set to 50-ms and 12.5-ms, respectively.

A modified version of the front-end character module in the transcript encoder was used to process the Korean language. Among several approaches, sub-character architecture was used to solve the data sparsity problem caused by the unique compositional principle of Korean letters [28].

B. E2E Emotional Speech Synthesis Systems

1) Tacotron2 Framework: In the Tacotron2 framework, the transcript encoder converted the input text sequence into 512-dimensional transcript embedding, where the encoder comprised three convolution layers, followed by a single bi-directional long short-term memory (Bi-LSTM) layer with 256 memory blocks for each direction. Each convolution layer contained 10 × 1 kernel with 512 channel sizes. The decoder then predicted the mel-spectrogram of the current frame with a previously generated spectrogram and context vector extracted from the transcript embedding through the location-sensitive attention network. Specifically, the previously generated spectrogram passed through the two fully connected layers with 256 units for each. The passed representation and the context vector were then fed into two LSTM layers with 1024 units per layer, followed by two projection layers, which generate stop token and mel-spectrogram components, respectively. The generated mel-spectrogram was improved by the post-net, which comprised five convolution layers containing 5 × 1 kernel and 512 channel sizes per layer. The WaveNet vocoder, which comprised three dilated convolution stacks with ten layers for each stack, converted the generated spectrogram into a time-domain speech waveform.

2) GST Network: In the GST network, the prosody encoder generated a prosody embedding from a mel-spectrogram of the RA, which was used to generate 256-dimensional emotion embedding with the style token layer. The prosody encoder comprised six 2D convolutional layers, followed by a single gated recurrent unit (GRU) layer with 128 units. Each convolutional layer had 3 × 3 kernel and 2 × 2 stride, and they contained 32, 32, 64, 64, 128, and 128 channels per layer. In the style token layer, a four-headed attention network and a bank of ten GST embeddings were used to measure similarity and to capture emotional information, respectively.[2] The generated emotion embedding was broadcasted and concatenated to the transcript embedding to synthesize emotional speech. In the inference phase, the emotion embedding for each emotion was generated by a matrix multiplication to each emotion weight values and the ten trained GSTs.

3) Tacotron2 With Emotion Label-Based Method: To demonstrate the effectiveness of the proposed CW-based emotion modeling method, we implemented a baseline system: Tacotron2 trained with an additional condition vector of the emotion label [19]. Note that the emotion label was converted into a three-dimensional (3D) one-hot vector and it was broadcasted and concatenated to the transcript embedding for emotional speech synthesis.

C. Subjective Results

We conducted two subjective listening tests, i.e., an MOS test and an emotion classification test, to evaluate the perceptual quality and emotional expressiveness of synthesized speech, respectively. Twelve native Korean listeners were asked to evaluate 30 randomly selected utterances generated by different network models, i.e., three types of emotions, ten utterances per emotion. Note that the baseline system, i.e., the emotion label-based Tacotron2 system, and the original recorded audio, i.e., "RAW," were also evaluated for comparison with the proposed CW-based method. In the two results, the performances of the RA-based method were obtained by averaging the output scores and accuracies of four different RA samples depicted in Table I and Fig. 2, respectively.

TABLE IV
MOS WITH A 95% CONFIDENCE INTERVAL FOR SEVERAL SYNTHESIS METHODS (P-VALUES WERE CALCULATED BETWEEN THE BASELINE AND PROPOSED CW-BASED METHOD)

System	Happy	Sad	Neutral
RAW	4.83±0.09	4.83±0.12	4.72±0.15
Baseline	4.15±0.20	4.15±0.35	3.81±0.17
RA	4.34±0.13	4.28±0.25	3.97±0.10
CW	**4.40±0.20**	**4.48±0.23**	**4.07±0.18**
p-value	**0.009**	**0.014**	**0.010**

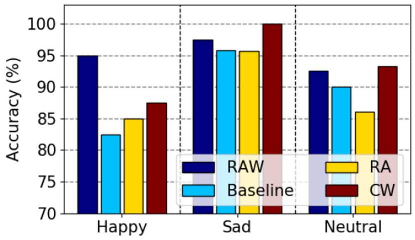

Fig. 4. Emotion classification results (%) with several synthesis methods.

Table IV depicts the average MOS results of the "RAW," and of three synthesis methods. The table shows that the proposed CW-based model significantly outperforms the baseline system ($p < 0.05$) in all three emotion cases.

In the emotion classification test, four categories were used to classify the output, i.e., happy, sad, neutral, and other; "other" category was chosen if the listeners had difficulty determining a specific emotion of the synthesized speech. Fig. 4 depicts the results, which show that the proposed CW-based method generates more emotionally expressive speech than the baseline system and the RA-based method in all three emotion cases.

V. DISCUSSION AND CONCLUSIONS

In this letter, we proposed a high-quality emotional E2E-TTS system by controlling the weights of each GST obtained via training with an emotion database. The distribution of emotion embedding formed distinct clusters in the emotional vector space, which indicated that the GST network could be used to model different types of emotions even without explicit supervision. By further investigating the relationship between the GSTs and three types of emotions as weight values, we confirmed that our proposed method could flexibly control various types of emotions through controlled weight values. The subjective evaluation results in terms of perceptual quality and emotional expressiveness demonstrated the superiority of the proposed system over conventional emotion label-based approach.

In the future, we would like to further extend the proposed algorithm to be applicable for generating fine-grained or multiple types of emotional speech in a single sentence. Combining the proposed method with the stochastic-based network structure such as the one used in GMVAE-Tacotron [29] could be a good strategy.

ACKNOWLEDGMENT

The authors would like to thank Clova Voice team at NAVER Corp., Seongnam, Korea, for their helpful comments and discussions.

[2] Since the attention network had four-headed attention, the weight values and ten GSTs were defined as 4 × 10 and 10 × 64 matrices, respectively. The matrix multiplication of the weight values and ten GSTs then made a 4 × 64 matrix flatten to generate a 256-dimensional emotion embedding vector.

References

[1] A. J. Hunt and A. W. Black, “Unit selection in a concatenative speech synthesis system using a large speech database,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 1996, vol. 1, pp. 373–376.

[2] A. W. Black and P. A. Taylor, “Automatically clustering similar units for unit selection in speech synthesis,” in *Proc. Eurospeech*, 1997, pp. 601–604.

[3] K.-S. Lee and S.-R. Kim, “Context-adaptive smoothing for concatenative speech synthesis,” *IEEE Signal Process. Lett.*, vol. 9, no. 12, pp. 422–425, Dec. 2003.

[4] H. Zen, K. Tokuda, and A. W. Black, “Statistical parametric speech synthesis,” *Speech Commun.*, vol. 51, no. 11, pp. 1039–1064, 2009.

[5] H. Zen, A. Senior, and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2013, pp. 7962–7966.

[6] H. Zen and H. Sak, “Unidirectional long short-term memory recurrent neural network with recurrent output layer for low-latency speech synthesis,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2015, pp. 4470–4474.

[7] 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.

[8] A. v. d. Oord *et al.*, “WaveNet: A generative model for raw audio,” 2016, *arXiv:1609.03499*.

[9] J. Sotelo *et al.*, “Char2Wav: End-to-end speech synthesis,” in *Proc. Int. Conf. Learn. Representations*, Apr. 2017.

[10] S. Ö. Arık *et al.*, “Deep voice: Real-time neural text-to-speech,” in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 195–204.

[11] A. Gibiansky *et al.*, “Deep voice 2: Multi-speaker neural text-to-speech,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2962–2970.

[12] Y. Wang *et al.*, “Tacotron: Towards end-to-end speech synthesis,” in *Proc. Conf. Int. Speech Commun. Assoc.*, 2017, pp. 4006–4010.

[13] J. Shen *et al.*, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2018, pp. 4779–4783.

[14] Y. Wang *et al.*, “Uncovering latent style factors for expressive speech synthesis,” in *Mach. Learn. Audio Signal Proc. Workshop at NIPS*, 2017.

[15] R. Skerry-Ryan *et al.*, “Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 4693–4702.

[16] Y. Wang *et al.*, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5180–5189.

[17] Y. Bian, C. Chen, Y. Kang, and Z. Pan, “Multi-reference Tacotron by intercross training for style disentangling, transfer and control in speech synthesis,” 2019, *arXiv:1904.02373*.

[18] Y.-J. Zhang, S. Pan, L. He, and Z.-H. Ling, “Learning latent representations for style control and transfer in end-to-end speech synthesis,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2019, pp. 6945–6949.

[19] Y. Lee, A. Rabiee, and S.-Y. Lee, “Emotional end-to-end neural speech synthesizer,” in *Mach. Learn. Audio Signal Proc. Workshop at NIPS*, 2017.

[20] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-based models for speech recognition,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 577–585.

[21] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.

[22] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate, in *Proc. Int. Conf. Learn. Repres.*, May 2015.

[23] M.-T. Luong, H. Pham, and C. D. Manning, “Effective approaches to attention-based neural machine translation,” in *Proc. Int. Conf. Empirical Methods NLP*, Sep. 2015, pp. 1412–1421.

[24] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, attend and spell: A neural network for large vocabulary conversational speech recognition,” in *Proc. Int. Conf. Acoust., Speech, Signal Process.*, 2016, pp. 4960–4964.

[25] A. Vaswani *et al.*, “Attention is all you need,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[26] O. Kwon, I. Jang, C. Ahn, and H.-G. Kang, “Emotional speech synthesis based on style embedded Tacotron2 framework,” in *Proc. Int. Tech. Conf. Circuits/Syst., Comput., Commun.*, 2019, pp. 344–347.

[27] L. v. d. Maaten and G. Hinton, “Visualizing data using t-SNE,” *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

[28] K. Stratos, “A sub-character architecture for Korean language processing,” in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 721–726.

[29] W.-N. Hsu *et al.*, “Hierarchical generative modeling for controllable speech synthesis,” 2018, *arXiv:1810.07217*.