applied sciences

MDPI

Article

Gated Recurrent Attention for Multi-Style Speech Synthesis

Sung Jun Cheon, Joun Yeop Lee, Byoung Jin Choi, Hyeonseung Lee and Nam Soo Kim *

Department of Electrical and Computer Engineering and the Institute of New Media and Communications, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea; sjcheon@hi.snu.ac.kr (S.J.C.); jylee@hi.snu.ac.kr (J.Y.L.); bjchoi@hi.snu.ac.kr (B.J.C.); hslee@hi.snu.ac.kr (H.L.)
* Correspondence: nkim@snu.ac.kr; Tel.:+82-02-880-8419

Received: 3 July 2020; Accepted: 30 July 2020; Published: 31 July 2020

Abstract: End-to-end neural network-based speech synthesis techniques have been developed to represent and synthesize speech in various prosodic style. Although the end-to-end techniques enable the transfer of a style with a single vector of style representation, it has been reported that the speaker similarity observed from the speech synthesized with unseen speaker-style is low. One of the reasons for this problem is that the attention mechanism in the end-to-end model is overfitted to the training data. To learn and synthesize voices of various styles, an attention mechanism that can preserve longer-term context and control the context is required. In this paper, we propose a novel attention model which employs gates to control the recurrences in the attention. To verify the proposed attention's style modeling capability, perceptual listening tests were conducted. The experiments show that the proposed attention outperforms the location-sensitive attention in both similarity and naturalness.

Keywords: attention; speech synthesis; style-modeling; global style token; attention; encoder-decoder; neural network

1. Introduction

It is crucial to produce natural prosody which includes tone, intonation, stress, and rhythm to generate a natural speech utterance. In the context of acoustics or speech signal processing, the prosody is interpreted as the composition of the duration of each phoneme in the utterance, the pitch contour over the whole sentence, and the spectral characteristics including amplitude. Each of these elements can be analyzed and synthesized individually, or jointly, depending on the modeling methods [1,2]. The prosody can be affected by various factors including identity of the speaker, emotion of the speaker, purpose of the speech, form of the speech, how lexical items relate to each other semantically or syntactically, and where the speaker places emphasis [3]. Styles are common prosodic characteristics in an utterance or several utterances, affected by shared factors in the speech. For example, we designate a style of a particular speaker identity with some common prosodic attributes observed from the utterances of the speaker. In the field of speech synthesis, techniques for modeling or generating styles have been studied [1,2,4].

In the statistical parametric speech synthesis (SPSS) studies [5–7], the speech is synthesized by first estimating the duration of each phoneme or phoneme state according to the context, and then regressing acoustic features such as the pitch and spectral information for the estimated duration. Although explicit statistical modeling of duration and acoustic features followed by speech parameter generation improves robustness against data sparseness, it degrades the quality and naturalness of the synthesized speech mostly due to the oversmoothed parameter trajectories [6]. In other words, the prosodical information may be discarded in the statistical modeling stage and cannot

be recovered in the synthesis stage by generating speech parameters merely from the means and variances of acoustic features. In order to compensate for the loss during statistical modeling, SPSS studies have used adaptation techniques, such as maximum likelihood linear regression (MLLR) [1,8], factored MLLR [9], learning hidden unit contribution (LHUC) [10,11], or retraining with control vector [12], to adapt to the style of various speakers and emotions. Though successful for some specific application [1,11], these techniques still suffer from oversmoothing, poor adaptation performance with less data, and difficulty in reflecting the style of an individual sentence. Furthermore, the duration and acoustic features are usually adapted separately, which can be considered undesirable for natural style expression. Another shortcoming of these approach is that they require labeled data for supervised training of the adaptation model.

Most of the aforementioned shortcomings in style modeling for SPSS can be somewhat mitigated in end-to-end (e2e) speech synthesis models. The e2e speech synthesis systems [13–15] based on sequence-to-sequence (seq2seq) models with attention [16,17] aim to synthesize the speech directly from a sequence of characters. The seq2seq model for speech synthesis consists of an encoder, an attention-based decoder, and a postprocessing network. The encoder embeds a sequence of characters into a sequence of embedding vectors. The attention-based decoder generates a stream of intermediate spectral features such as mel-spectrogram by attending on the character embedding vectors extracted from the input texts. The postprocessing network converts the intermediate spectral features into spectrograms with larger dimensionality. Unlike SPSS, which explicitly estimates the duration of each phoneme state and generates the acoustic parameters according to that duration, the e2e speech synthesis framework implicitly determines the duration by updating the decoder hidden state through an attention mechanism. This implicitly determined duration enables to apply a style that affects both the duration and spectral features with a single vector of style representation.

Several embedding methods have been proposed to represent styles for e2e speech synthesis: an embedding lookup to give speaker identity as a condition [18], an embedding vector from a reference encoder for modeling the style of an utterance [2,4], and a speaker embedding trained for speaker verification [19]. The style embedding vector is concatenated with the output of the text encoder in the e2e speech synthesis systems. The reference-encoder-based techniques [2,4,20] have shown that a speech with similar prosodic characteristics can be generated by conditioning the e2e speech synthesis model with an unsupervised style representation derived from a reference utterance. However, when synthesizing speech with styles rarely found in the training data, the prosodic similarity between the reference speech and the synthesized speech tends to be low. Among various factors of the prosody, it is more difficult to model the local style such as rhythm or duration, rather than the global style which is usually determined by tone or speaker identity. It has been reported that the speaker similarity observed from the speech synthesized with unseen speaker-style is low although the naturalness is good [19]. One of the reasons for these problems is that the attention-update at each decoder step is too simple to cover various styles in a multi-speaker database. Such simplicity causes the attention-based decoder to be overfitted to the styles that have been seen during the training session. For these reasons, in order to learn and synthesize voices of various styles, an attention mechanism that can preserve longer-term context and adjust the duration depending on the context is highly required.

In this paper, we propose a novel attention model based on gated recurrence, which we call gated recurrent attention (GRA). GRA controls the contextual information by employing two gates: the update gate and the scoring gate. The update gate determines how much the attention forgets the previous recurrent state which retains the attention alignments from the previous steps. The scoring gate determines how much the previous recurrent state is reflected to derive the attention alignment and decoder output. The two gates are computed according to the attention key, query, and the previous recurrent state.

The GRA's recurrent states are the convolved attention weights, which are also known as the location [21]. The recurrent attention (RecAtt) and recurrent neural network (RNN) attention

(RNNAtt) [22] apply a vanilla RNN to weighted input embeddings, which are also known as context [22] or glimpse [23]. Gated recurrent unit (GRU)-gated attention (GAtt) [24] uses a GRU to generate a recurrent representation of the input embeddings depending on the previous decoder state. GRA differs from the three other attention methods in that the GRA seeks attention alignments according to where it was, not to what was, focused in the previous time-step. For this reason, GRA is useful for tasks that generate different outputs depending on the location even if the inputs are the same, such as speech synthesis. GRA can be regarded as a gated recurrent version of the location-sensitive attention (LSA) [21] as the gates in GRA operate similarly to the gates in the GRU [25]. GRA is expected to be advantageous for retaining long short-term information and creating shortcuts between multiple temporal steps in a similar fashion to the GRU [26]. GRA has been found useful for seq2seq tasks to deal with long sequences where there exists huge variability in the alignments between the input and output sequences.

2. Backgrounds

2.1. Gated Recurrent Unit

GRU is a RNN that is motivated by the long short-term memory (LSTM) [27,28], but simpler to compute and implement [25]. GRU updates the recurrent hidden state h_t to be a linear interpolation between the previous recurrent hidden state h_{t-1} and the candidate recurrent hidden state $\tilde{h}_t$ as follows,

$$h_t = (1 - z_t)\, h_{t-1} + z_t \tilde{h}_t \tag{1}$$

where an update gate z_t controls how much the unit updates its hidden state. The update gate z_t is determined by

$$z_t = \sigma\left(W_z x_t + U_z h_{t-1}\right) \tag{2}$$

given an input x_t at time t in the input sequence $x = (x_1, x_2, ..., x_T)$ with σ representing the logistic sigmoid function. In Equations (2)–(4), W_* and U_* are parameter matrices. The candidate recurrent hidden state is obtained by

$$\tilde{h}_t = \phi\left(W x_t + U\left(r_t \odot h_{t-1}\right)\right) \tag{3}$$

where ϕ is a nonlinear activation function such as tanh, r_t is a reset gate, and $\odot$ denotes element-wise multiplication [26]. The reset gate r_t is determined according to

$$r_t = \sigma\left(W_r x_t + U_r h_{t-1}\right). \tag{4}$$

Unlike the vanilla RNN, which always replaces the recurrent hidden state with the current input and the previous recurrent hidden state, i.e.,

$$h_t = \phi\left(x_t, h_{t-1}\right), \tag{5}$$

GRU maintains the existing context h_{t-1} and add the new content $\tilde{h}_t$ as in Equation (1) [26]. This additive nature of GRU makes it easy for each unit to remember the existence of a specific feature in the input sequence for long-time-steps.

2.2. Tacotron

Tacotron [13] is a seq2seq model equipped with attention [16,17] designed for speech synthesis. This model consists of an encoder, an attention-based decoder, and a postprocessing network as shown in Figure 1. The encoder first encodes a one-hot vector, which represents each character in a text sequence into a continuous vector. Then, a set of nonlinear transformations called the pre-net, which includes two layers of fully connected network with ReLU activation [29] and dropout [30], is applied to each layer. A convolution bank with highway connection and bidirectional-GRU (CBHG) module

summarizes contextual information in the text. The content-based attention introduced in [16,17] is adopted for the attention-based decoder. The decoder RNNs consist of a stack of GRUs with residual connections [31]. As the training target of the decoder, a mel-scale spectrogram is used instead of a high-dimensional raw spectrogram. In the postprocessing network, another CBHG module is used to transform the mel-scale spectrogram into the linear-scale spectrogram with larger number of bands for synthesizing high quality audio. For the vocoder, the Griffin–Lim algorithm [32] and WaveNet [33] vocoder have been applied. The content-based attention mechanism has a problem that the attention scores are nearly unchanged regardless of the context for the equivalent or very similar inputs. For this reason, an attention mechanism that scores according to the attention alignment from the previous step, LSA [21], is used in an extended version of the Tacotron, also known as Tacotron 2 [15]. In Tacotron 2, a vanilla LSTM and convolutional layers are used in the encoder and decoder instead of CBHG-stacks and GRU recurrent layers. Training is performed to minimize the summed mean square error obtained before and after the post-net:

$$\mathcal{L}_{Taco2} = \mathcal{L}_{before-post-net} + \mathcal{L}_{after-post-net}. \quad (6)$$

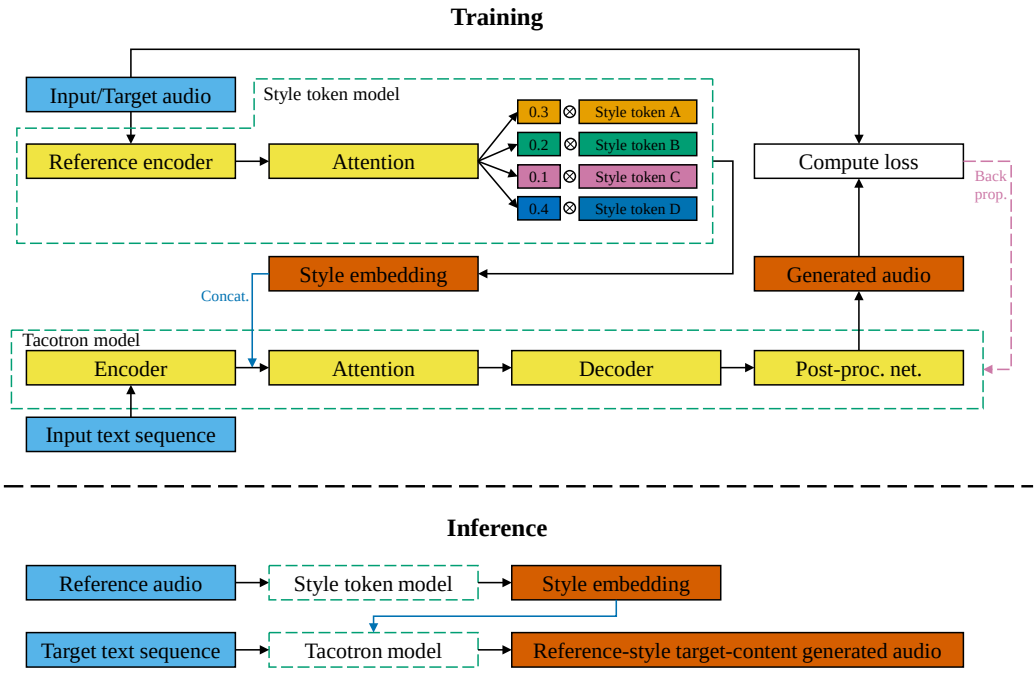


Figure 1. Overall architecture of Tacotron with global style tokens.

2.3. Global Style Tokens

The global style tokens (GSTs) [4] are a set of embeddings jointly trained with the Tacotron to represent speech style in an unsupervised manner. The style token model includes a reference encoder, an attention module, and randomly initialized tokens as shown in Figure 1. The reference encoder encodes a variable-length reference audio into a fixed-length vector with convolutional filters and a layer of RNN to represent the style of the reference audio. In the style token model, the attention module is not used to learn an alignment. The attention module learns the similarity between the reference embedding and each token. The attention module outputs a set of weights that represent the contribution of each token to the reference embedding. The weighted sum of the GSTs is concatenated with the Tacotron text embedding at every time-step. The style token model is jointly trained with the Tacotron by backpropagating the Tacotron decoder reconstruction loss. The trained GST model can encode the style extracted from any reference audio into a style embedding vector.

The Tacotron conditioned on the style embedding vector synthesizes speech with prosody similar to that of the reference audio.

2.4. Location-Sensitive Attention

LSA [21], which extends the additive content-based attention [16] to cumulate attention weights from the previous decoder time-step, was applied to the Tacotron 2. The LSA stochastically generates a decoder outputs $y = (y_1, y_2, ..., y_T)$ from an encoded input sequence $x = (x_1, x_2, ..., x_N)$. Note that the encoded input sequence and the decoder state sequence may have different time-scales, e.g., for speech synthesis, the decoder state represents a latent acoustic feature at each frame, whereas the encoded input represents a character or phoneme. The LSA updates a recurrent state $f_{n,t} \in \mathbb{R}^k$ at each decoder time-step t with the attention weights $\alpha_{n,t}$ by convolving it with a matrix $F \in \mathbb{R}^{k \times r}$:

$$f_{n,t} = f_{n,t-1} + F * \alpha_{n,t} \tag{7}$$

where $*$ denotes the convolution operation and the attention weights $\alpha_{n,t}$, which are also called as alignment between the encoder and decoder are computed by the score $e_{n,t}$ as follows,

$$\alpha_{n,t} = \frac{\exp\left(e_{n,t}\right)}{\sum_{v=1}^{N} \exp\left(e_{v,t}\right)}. \tag{8}$$

The score $e_{n,t}$ rates how much the encoder output at the location n is to be attended for generating the decoder state around the location t. The LSA uses the previous recurrent state $f_{n,t-1}$ to compute the score $e_{n,t}$:

$$e_{n,t} = v^T \tanh\left(W s_{t-1} + V x_n + U f_{n,t-1} + b\right) \tag{9}$$

where s_{t-1} is the $(t-1)$-th decoder recurrent state; v and b are vectors; and U, V, and W are matrices. The attention-based decoder generates an output y_t by focusing on the relevant elements of x as follows,

$$g_t = \sum_{n=1}^{N} \alpha_{n,t} x_n, \tag{10}$$

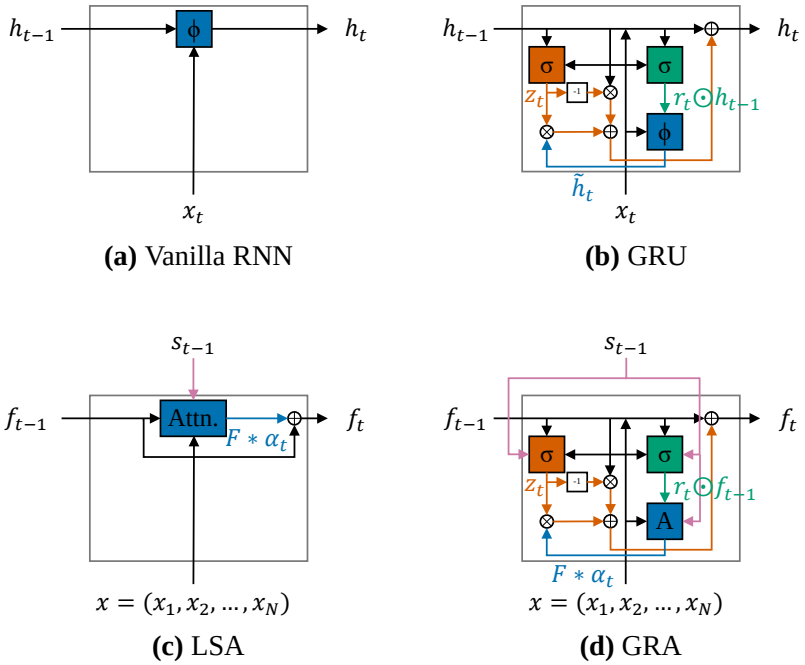
$$s_t \sim recurrency\left(s_{t-1}, g_t, y_t\right), \tag{11}$$

$$y_t \sim generator\left(s_{t-1}, g_t\right) \tag{12}$$

where g_t is the glimpse [23]. The recurrent activation and output function of the LSTM [27] are typically used as the *recurrency* and *generator*, respectively.The score $e_{n,t}$ is location-sensitive in the sense that the previous recurrent state $f_{n,t-1}$ is determined by the convolved previous attention weights $F * \alpha_{n,t-1}$, which contain the location of the focus from the previous step. The LSA is also known as hybrid attention because the score $e_{n,t}$ is a function of both the content x_n and the previous recurrent state $f_{n,t-1}$.

3. Gated Recurrent Attention

The LSA aligns a sequence of the decoder states with the corresponding sequence of the encoder outputs, conditioned on the previous recurrent state at each decoder time-step. The mechanism of updating the recurrent state in the LSA is similar to that of updating the hidden state in the vanilla RNN as shown in Figure 2a,c. Similar to Equation (5), where the hidden state is updated through a function of the content x_n and the previous hidden state h_{t-1}, the recurrent state $f_{n,t}$ in LSA is updated by filtering the normalized score of content x_n and the previous recurrent $f_{n,t-1}$ in Equations (7)–(9). As described in [34], although the vanilla RNN has a powerful ability to represent context, due to the problem of vanishing gradient, it is inadequate for the tasks involving long-term dependencies. Likewise, we can assume that the LSA is not suitable for reflecting long-term contexts.



(a) Vanilla RNN **(b)** GRU

(c) LSA **(d)** GRA

Figure 2. Illustration of recurrences in (**a**) Vanilla recurrent neural network (RNN), (**b**) Gated recurrent unit (GRU), (**c**) Location-sensitive attention (LSA), and (**d**) Gated recurrent attention (GRA).

For e2e speech synthesis, it is important to deal with the **long-term contexts** in an attention model. Other tasks using attention-based seq2seq models such as machine translation [25] and speech recognition [21] generate a sequence of characters or words. However, for speech synthesis, the attention-based decoder generates a speech sequence which is usually much longer than the input sequence of characters or words. Analogous to the vanishing gradient problem of the vanilla RNN, as the length of the decoder state sequence increases, the simple recursive structure in LSA may deteriorate the performance of the attention.

Moreover, it is necessary for an attention model to be able to control the degree of **updating the recurrent state** $f_{n,t}$. The duration and acoustic characteristics of each character in spoken utterances differ by the character, utterance, and context. A vowel is a short or long monophthong or diphthong according to the context, e.g., a vowel *i* sounds [ɪ] in *bit*, [i] in *police*, or [aɪ] in *bite*. To model the intra-character and inter-character phoneme transitions in various prosody, the recurrent state which contains contextual information should be updated faster near the transition-boundary than far from the transition-boundary.

Furthermore, the **location-sensitivity** is important for speech synthesis. The score's sensitivity to the recurrent state is crucial for generating not only the alignment but also the spectrogram. The decoder state at the t-th step y_t is generated by attending encoded inputs x corresponding to the score $e_{n,t}$ as in Equations (8)–(11). The decoder state sequence is used to synthesize the output mel-spectrogram in Tacotron. From a single character, several frames of spectrograms which have different spectral characteristics should be generated in speech synthesis. In speech recognition, the encoder encodes the spectral characteristics which have positional dependencies. However, in speech synthesis, the encoder output merely specifies the character to be synthesized. In other words, to generate the decoder state y_t, it is necessary to know where the frame is located by conditioning the previous recurrent state $f_{n,t-1}$. Thus, for e2e speech synthesis, an attention mechanism that can control the **location-sensitivity** and **updating the recurrent state** while considering a **long-term context** is required.

We propose an attention mechanism which can model and generate alignment between two sequences with style-variability and long-term dependency. It includes an update gate z_t and a scoring gate r_t, which are presented respectively by

$$z_t = \sigma\left(W_z s_{t-1} + V_z x_n + U_z f_{n,t-1} + b_z\right), \tag{13}$$

$$r_t = \sigma\left(W_r s_{t-1} + V_r x_n + U_r f_{n,t-1} + b_r\right) \tag{14}$$

where σ is a logistic sigmoid function. In Equations (13)–(15), U_*, V_*, and W_* denote parameter matrices and b_* denotes a bias vector. The score $e_{n,t}$ weighs how much the n-th encoded input should be relatively attended by the t-th decoder state given as follows,

$$e_{n,t} = v^T \tanh\left(W_e s_{t-1} + V_e x_n + U_e\left(r_t \odot f_{n,t-1}\right) + b_e\right) \tag{15}$$

where v is a vector and $\odot$ denotes element-wise multiplication. The attention alignment $\alpha_{n,t}$ is also given by Equation (8) as in the LSA. The recurrent state $f_{n,t}$ is updated by the previous recurrent state $f_{n,t-1}$, alignment $\alpha_{n,t}$, and update gate z_t:

$$f_{n,t} = \left(1 - z_t\right) \odot f_{n,t-1} + z_t \odot F * \alpha_{n,t} \tag{16}$$

where F is a convolutional filter and $*$ denotes the convolution operation.

The attention mechanism we propose in this work, GRA, is a gated recurrent variant of the LSA [21]. The recurrences in the vanilla RNN, GRU, LSA, and GRA are illustrated in Figure 2 for comparison. The GRU is designed to update the recurrent hidden state h_t to be a weighted average of the previous recurrent hidden state h_{t-1} and the candidate recurrent hidden state $\tilde{h}_t$ with the update gate z_t, which is a function of the input at time t and the previous recurrent hidden state h_{t-1} as in Equation (1). Similarly, in Equation (16), the GRA updates the recurrent state $f_{n,t}$ to be a weighted average of the previous recurrent state $f_{n,t-1}$ and the convolved alignment $F * \alpha_t$ with the update gate z_t, which is a function of the query, key, and the previous recurrent state $f_{n,t-1}$. The candidate recurrent hidden state $\tilde{h}_t$ in Equation (3) is a function of the input at time t and the multiplication of the previous recurrent hidden state h_{t-1} and the reset gate r_t, which is computed by Equation (4). The score $e_{n,t}$ in Equation (15) is a function of the query, key, and the product of the previous recurrent state $f_{n,t-1}$, which comes from the previous alignment $\alpha_{n,t-1}$, with the scoring gate r_t.

The proposed GRA is expected to generate better alignment between the encoded input and the decoder state than the LSA by reflecting the **long-term context** and creating shortcuts between multiple temporal steps, as the GRU models a sequential data better than the vanilla RNN [26]. The update gate z_t, scoring gate r_t, and score $e_{n,t}$ in GRA are **sensitive to the location** as they are computed conditioned on the previous recurrent state as well as the content. Each of the two gates controls **how much the recurrent state is updated** or **how much the score is sensitive to the recurrent state**.

4. Experiments and Results

To demonstrate the GRA's alignment and style modeling performance, we conducted a set of experiments, comparing multi-style e2e speech synthesis systems with different attention methods. Tacotron2 [15] with GSTs [4] for style modeling was trained on M-AILABS US dataset which includes 102 hours of speech clips spoken by three US English speakers [35]. To guide the attentions to have straight line alignments in early training-stage, we used a variant of guided attention described in [14].

4.1. Tacotron2 with Global Style Tokens

The dimensionality of the character embedding generated by the Tacotron text encoder was set to be 512. The proposed GRA and the conventional LSA were compared for style modeling. The decoder and the postprocessing networks in [4] were used to generate the mel-spectrograms from the input text sequence. The reference encoder in the GST model was composed of 6 convolutional layers followed by a single-layer of 128-unit unidirectional GRU. Instead of encoding 10 randomly

initialized 256-dimensional tokens [3], we encoded 32,512-dimensional tokens into a style embedding vector to have the same dimensionality with the character embedding, by applying a content-based multi-head tanh attention [36]. The Tacotron2-GST model was trained with a batch size of 16 for 200 epochs on one NVIDIA M40 GPU. The learning rate was 10^{-3} initially and exponentially decayed from 40k to 100k iterations by 10^{-5}. To facilitate the reproduction of the experiment, implementation details not mentioned in this paper are the same with an open source [37] Tacotron2 with GSTs. To synthesize the audio from mel-spectrogram, we used Parallel WaveGAN vocoder [38]. The Parallel WaveGAN was trained on the three speakers in the M-AILABS US dataset and a subset of VCTK corpus [39], which consists of all speech clips of 12 selected speakers.

4.2. Decaying Guided Attention

Guided attention has been proposed to make attention alignment $\alpha_{n,t}$ closer to a straight line by utilizing prior knowledge that attention is mostly aligned close to a straight line in speech synthesis [14]. For the input sequence with length N and the output sequence with length T, the guided attention loss $\mathcal{L}_{ga}$ is computed by

$$\mathcal{L}_{ga} = \mathop{\mathbb{E}}_{n,t}[\alpha_{n,t} W_{n,t}] \tag{17}$$

where $W_{n,t}$ is given as

$$W_{n,t} = 1 - \exp\left(-\frac{(n/N - t/T)^2}{2g^2}\right). \tag{18}$$

In [14], the constant g is set to 0.2. Adding guided attention loss to the training loss helps attentions to be quickly optimized in the early stage. However, in the latter stages of training, adding guided attention loss can disrupt the reconstruction of spectrogram, and the creation of natural prosody by forcing attention alignment close to a straight line. To avoid these problems, we added guided attention loss to the reconstruction loss only up to the first 5k training iterations then removed it from the loss. Before removing the guided attention, attention was aligned very close to a straight line, then after removal, attention alignment was excessively blurred or split between 10k to 50k iterations. Based on the assumption that this problem is due to the sudden disappearance of the strong constraints that have forced the attention alignment, we added decaying guided attention loss $\mathcal{L}_{dga}$ to the Tacotron2 loss:

$$\mathcal{L}_{total} = \mathcal{L}_{Taco2} + \mathcal{L}_{dga}, \tag{19}$$

$$\mathcal{L}_{dga} = \begin{cases} \mathbb{E}\left[\dfrac{A\alpha_{n,t} W_{n,t}}{\sqrt{iteration + 1}}\right] & \text{if } iteration <= 5000 \\ 0 & \text{if } iteration > 5000 \end{cases} \tag{20}$$

where $A = 100$ and $g = 0.4$ in this work. The attention alignment plots during the early stage of training the Tacotron2-GST with guided attention and decaying guided attention can be found in the Appendix A. It seems that decaying guided attention loss helps the attentions to be trained stably than the original guided attention.

4.3. Datasets and Feature Processing

The M-AILABS US corpus [35] consists of three speakers: Elliot, Judy, and Mary. Moreover, the total audio duration for each speaker is 38 h 23 min, 36 h 43 min, and 27 h, respectively. All speech clips are sampled at 16 kHz. In our experiments, M-AILABS US dataset was used during both training and inference stages. VCTK corpus includes speech data spoken by 109 native English speakers with various accents [39]. There are approximately 400 clips per speaker sampled at 48 kHz. Among the 109 speakers, 6 female and 6 male speakers with distinctive accents and regional characteristics are selected for the experiments. The speaker information used in Tacotron2 with GSTs experiments is given in Table 1. VCTK dataset was used only as a reference style of Tacotron2-GSTs when synthesizing speech.

We downsampled the audio in the VCTK corpus to 16 kHz. The texts in the M-AILABS US corpus were normalized and symbolized for character embedding. The silent part at the front and back of the audio clips in both corpus were trimmed. The log mel-spectrogram feature was computed with 50 ms window, 12.5 ms hop-length short-time Fourier transform, and 80-channel mel-scale filterbank.

Table 1. M-AILABS US and VCTK speaker information.

Speaker ID	Gender	Accent
MAILABS_Elliot	M	American
MAILABS_Judy	F	American
MAILABS_Mary	F	American
VCTK_p226	M	English
VCTK_p248	F	Indian
VCTK_p255	M	Scottish
VCTK_p270	M	English
VCTK_p282	F	English
VCTK_p293	F	Northern Irish
VCTK_p295	F	Irish
VCTK_p299	F	American
VCTK_p306	F	American
VCTK_p334	M	American
VCTK_p360	M	American
VCTK_p376	M	Indian

4.4. Evaluation Methods

We evaluated the performance of LSA and GRA through two kinds of subjective listening tests. Synthesized speech samples from Tacotron2-GSTs trained with the two attentions were rated by 16 speech-expert listeners with a headphone. The listeners are from 25 to 33 years old and fluent in English. The speech samples were synthesized on style-references in M-AILABS US and VCTK. Sixteen and 4 speech clips were picked for each reference speaker in the M-AILABS US and VCTK, respectively, i.e., 48 clips per corpus were picked as the style-references. Two sentences which were not included in the training dataset nor style-references were randomly picked for the listening test per the style-references. The synthesized audio samples were randomly shuffled on each attention mechanism as well as each style-reference. To compare which attention transfers style better, the listeners rated how much the prosodic style of synthesized samples is similar to that of the reference audio by the comparison category rating (CCR) method [40]. The listeners used the scale in Table 2 a to judge the prosodic similarity of the Tacotron2-GST-GRA samples relative to that of Tacotron2-GST-LSA samples, to the reference style. The measured similarity difference is presented by the comparison mean opinion score (CMOS). To measure the prosodic naturalness of the synthesized speech, we used the absolute category rating (ACR) method [40]. Listeners were asked to rate how much natural the synthesized speech is with the scale in Table 2b. The measured naturalness is given in terms of the mean opinion score (MOS).

4.5. Evaluation Results

The result of subjective listening tests is presented with 95% confidence intervals in Table 3. For all the reference speakers in the training or test dataset, GRA outperformed LSA in terms of the similarity of the synthesized style to the reference style. The difference in CMOS between the two attention methods on the VCTK was more significant than on the M-AILABS. The prosody of the speech synthesized with GRA was found more natural than or at the same level with the LSA for all the reference speakers. The p-values from independent-samples unequal-variance one-sided t-tests for the two attentions' scores were 3.80×10^{-8} and 1.16×10^{-12}, respectively, with MAILABS and VCTK style-references, which were much less than 0.05. Thus, the naturalness difference caused

by the two attentions is statistically significant. The difference in MOS between the two attention methods on the VCTK was also more significant than on the M-AILABS. Note that the Tacotron2-GSTs models were trained on M-AILABS. The two differences imply that GRA is more robust to overfitting and better at generalization than LSA in style speech synthesis. A subset of speech samples used in the evaluation is accessible on a webpage (https://gra-example.github.io/).

Table 2. (**a**) Comparison category rating CCR method for similarity comparison mean opinion score (CMOS) and (**b**) absolute category rating (ACR) method for naturalness mean opinion score (MOS).

(**a**) CCR	
Similarity	**Score**
GRA is much more similar	3
GRA is more similar	2
GRA is slightly more similar	1
About the same	0
LSA is slightly more similar	−1
LSA is more similar	−2
LSA is much more similar	−3
(**b**) ACR	
Quality	**Score**
Excellent	5
Good	4
Fair	3
Poor	2
Bad	1

Table 3. Similarity CMOS and naturalness MOS.

Speaker of Reference Audios	**CMOS**	**MOS**	
		LSA	**GRA**
MAILABS_Elliot	0.447 ± 0.110	3.133 ± 0.085	**3.303 ± 0.080**
MAILABS_Judy	0.096 ± 0.114	**4.113 ± 0.065**	4.100 ± 0.078
MAILABS_Mary	0.438 ± 0.120	3.793 ± 0.089	**4.213 ± 0.069**
MAILABS_everage	0.327 ± 0.067	3.680 ± 0.051	**3.872 ± 0.048**
VCTK_p226	0.828 ± 0.191	2.953 ± 0.158	**3.313 ± 0.147**
VCTK_p248	1.172 ± 0.197	2.594 ± 0.156	**3.578 ± 0.192**
VCTK_p255	0.453 ± 0.175	2.594 ± 0.172	**2.672 ± 0.157**
VCTK_p270	0.227 ± 0.214	**3.086 ± 0.151**	2.961 ± 0.163
VCTK_p282	0.906 ± 0.196	2.484 ± 0.160	**3.242 ± 0.194**
VCTK_p293	0.719 ± 0.264	3.938 ± 0.169	**4.188 ± 0.132**
VCTK_p295	0.102 ± 0.223	3.969 ± 0.175	**4.266 ± 0.126**
VCTK_p299	0.219 ± 0.206	4.047 ± 0.171	**4.125 ± 0.154**
VCTK_p306	0.781 ± 0.254	3.789 ± 0.171	**4.313 ± 0.130**
VCTK_p334	0.086 ± 0.161	2.641 ± 0.152	**2.648 ± 0.147**
VCTK_p360	0.055 ± 0.214	**3.070 ± 0.157**	3.055 ± 0.154
VCTK_p376	0.930 ± 0.182	2.930 ± 0.146	**3.039 ± 0.153**
VCTK_average	0.540 ± 0.063	3.174 ± 0.055	**3.450 ± 0.054**

5. Conclusions

We proposed and evaluated the Gated Recurrent Attention, a novel attention mechanism with gated recurrence for enriching style modeling capability in multi-style speech synthesis. Although the GRA is not a technique for directly modeling style such as global style tokens, the experimental results show that the GRA can improve the similarity and naturalness of synthesized speech. GRA was found to be more effective in transferring unseen styles, which implies that the generalization performance

of GRA is better than that of conventional techniques. Other gating architectures such as LSTM can be easily applied to the GRA. If a prominent gate-based RNN is introduced, the effect of the GRA may be boosted by it. As the GRA is a general attention method that can be applied to any sequential process, GRA can be evaluated on time series data or video processing in further studies.

Author Contributions: Conceptualization, S.J.C. and N.S.K.; methodology, S.J.C. and J.Y.L.; software, S.J.C. and H.L.; validation, S.J.C. and N.S.K.; formal analysis, S.J.C.; investigation, S.J.C. and B.J.C.; resources, S.J.C. and N.S.K.; data curation, S.J.C.; writing—original draft preparation, S.J.C.; writing—review and editing, J.Y.L., B.J.C., H.L., and N.S.K.; visualization, S.J.C.; supervision, N.S.K.; project administration, N.S.K.; funding acquisition, N.S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the research fund of Signal Intelligence Research Center supervised by the Defense Acquisition Program Administration and Agency for Defense Development of Korea.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript.

SPSS	statistical parametric speech synthesis
MLLR	maximum likelihood linear regression
LHUC	learning hidden unit contribution
e2e	end-to-end
seq2seq	sequence-to-sequence
GRA	gated recurrent attention
GRU	gated recurrent unit
RecAtt	recurrent attention
RNN	recurrent neural network
RNNAtt	RNN attention
GAtt	GRU-gated attention
LSTM	long short-term memory
CBHG	convolution bank with highway network and bidirectional-GRU
LSA	location-sensitive attention
GST	global style token
CCR	comparison category rating
CMOS	comparison mean opinion score
ACR	ablsolute category rating
MOS	mean opinion score

Appendix A. Guided Attention and Decaying Guided Attention

We added guided attention $\mathcal{L}_{ga}$ to the reconstruction loss as follows,

$$\mathcal{L}_{total} = \mathcal{L}_{Taco2} + \mathcal{L}_{ga} \tag{A1}$$

where

$$\mathcal{L}_{ga} = \begin{cases} \mathbb{E}\left[A\alpha_{n,t}W_{n,t}\right] & \text{if } iteration <= 5000 \\ 0 & \text{if } iteration > 5000 \end{cases}, \tag{A2}$$

$A = 100$, and $g = 0.4$. We propose to add decaying guided attention $\mathcal{L}_{dga}$ in Equations (19) and (20). To compare the two guiding methods, we attach the attention alignment plot during training Tacotron-GST-LSA with guided attention on Figure A1, Tacotron-GST-LSA with decaying guided attention on Figure A2, Tacotron-GST-GRA with guided attention on Figure A3, and Tacotron-GST-GRA with decaying guided attention on Figure A4. The attention alignment plots at $step = 2000$ and 4000 in Figures A1 and A3, respectively, show that the guided attention too much forces the alignments to be close to a straight line. The forced alignments make the decoder to focus only on one input embedding without considering the context and may produce a weird prosody. The plots at $step = 20{,}000$ and

50,000 in Figure A1 and $step$ = 10,000 and 20,000 in Figure A3 show the attention failures caused by the sudden disappearance of the guided attention. The plots in Figures A2 and A4 show that the attention is stably trained with decaying guided attention.

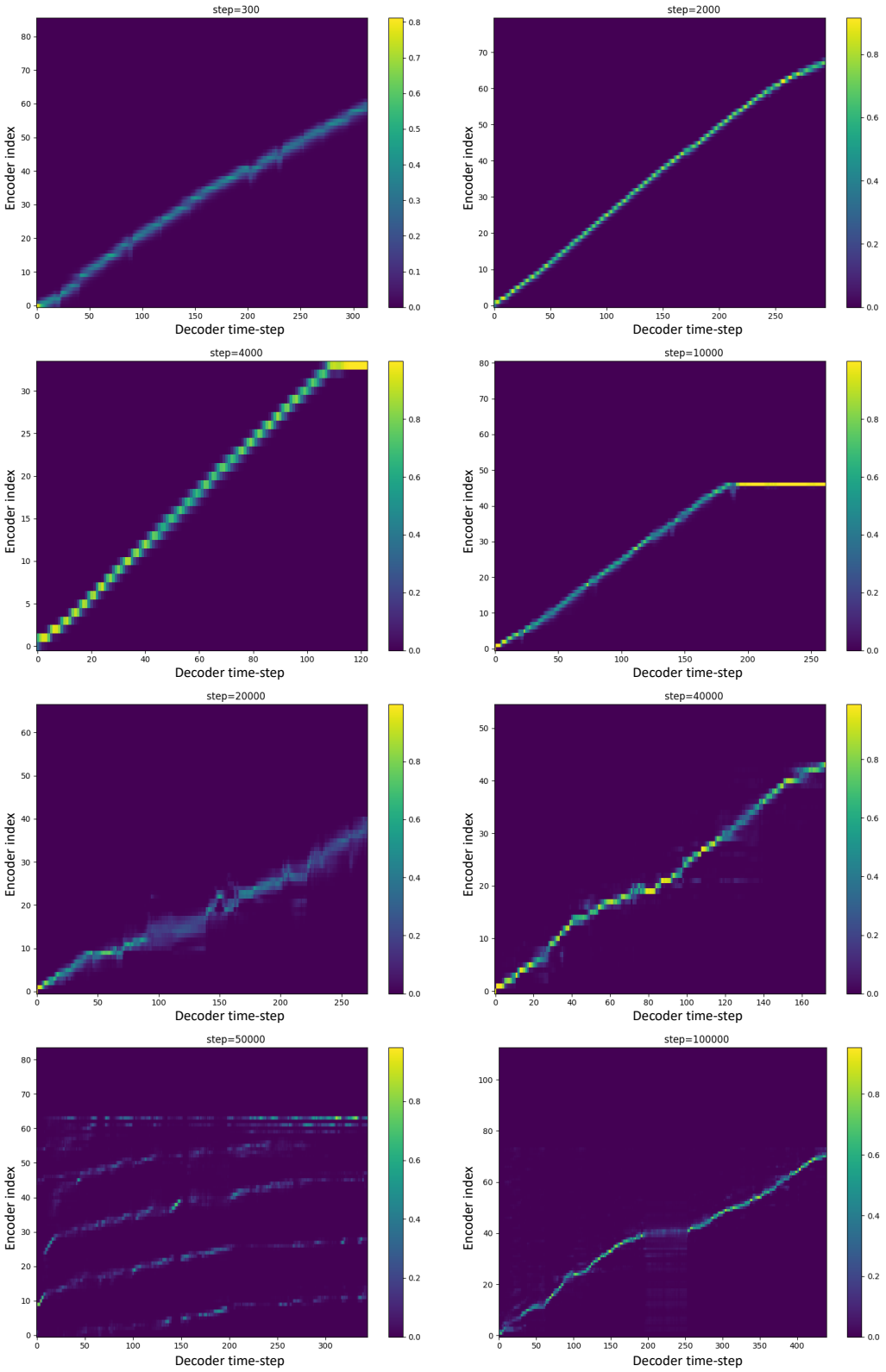


Figure A1. Attention plot during training Tacotron2-GST-LSA with guided attention.

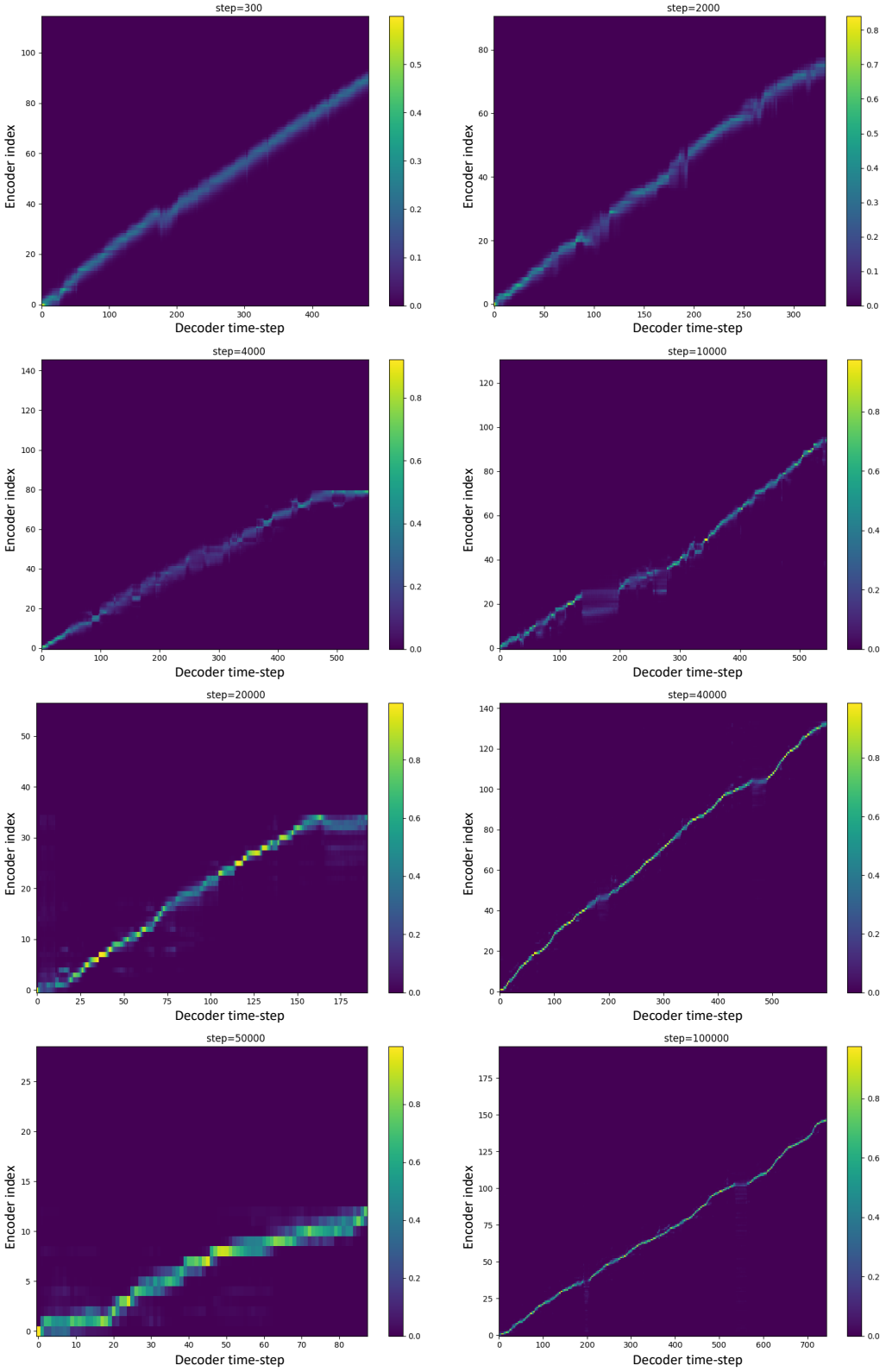


Figure A2. Attention plot during training Tacotron2-GST-LSA with decaying guided attention.

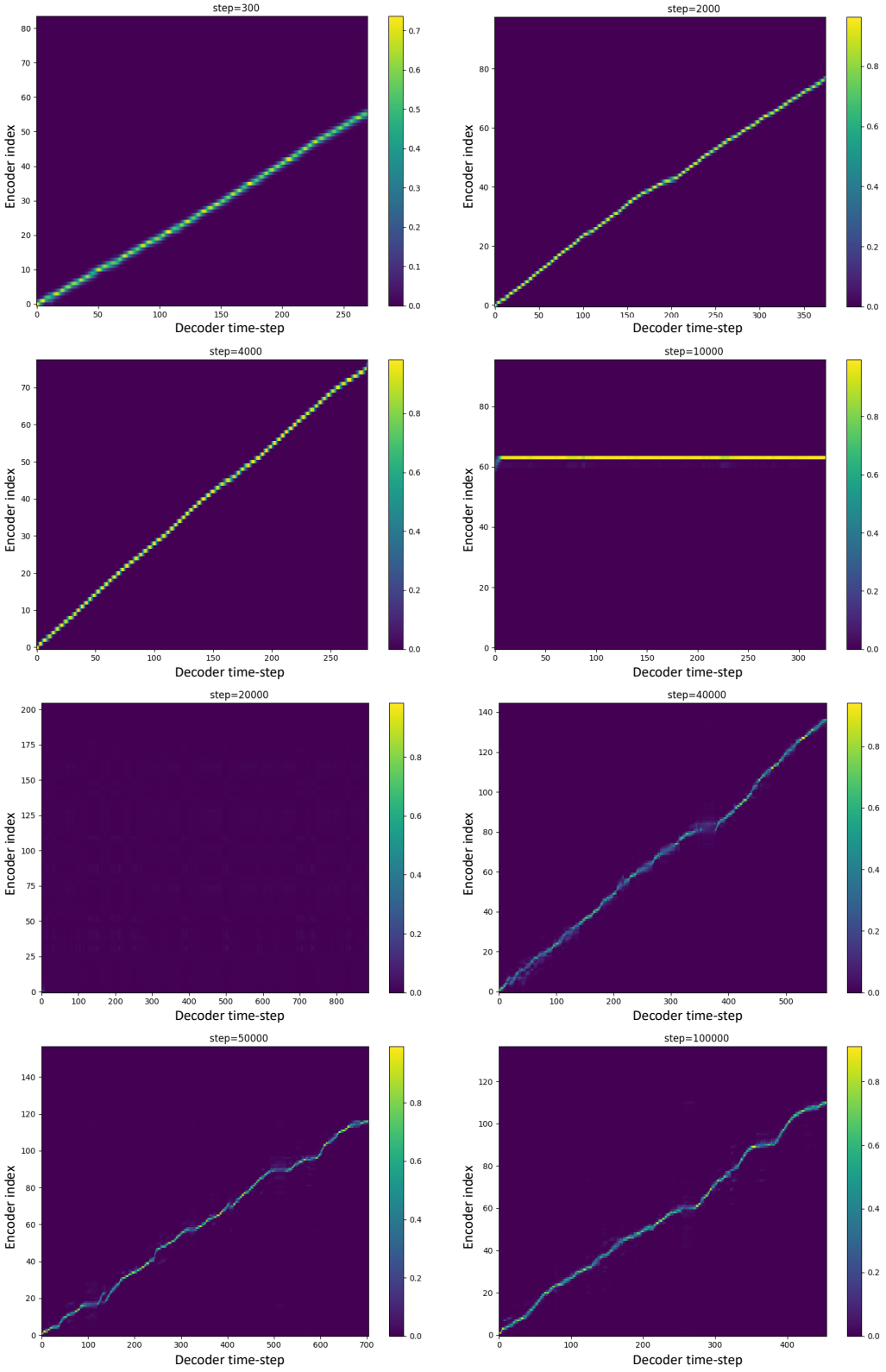


Figure A3. Attention plot during training Tacotron2-GST-GRA with guided attention.

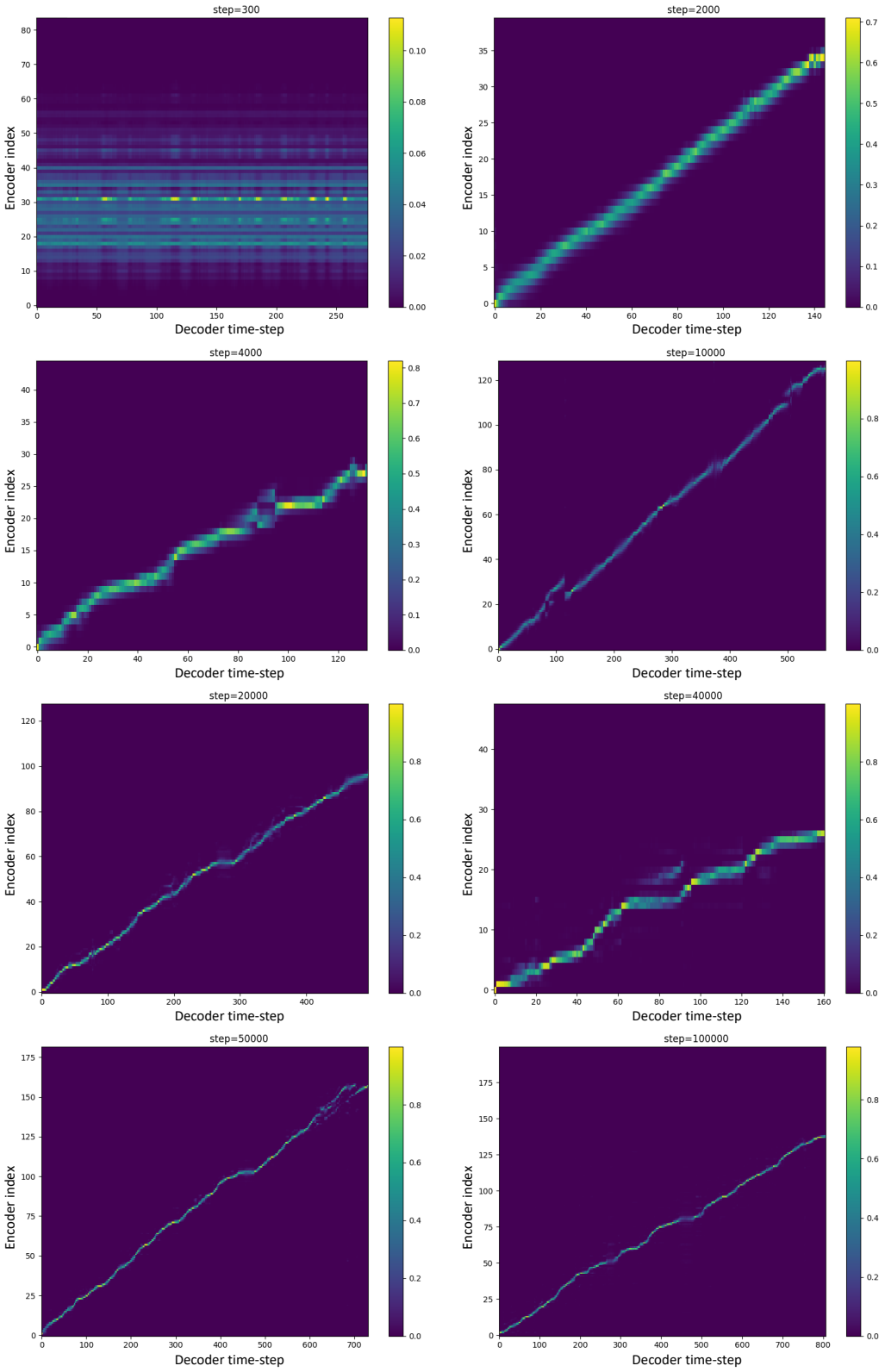


Figure A4. Attention plot during training Tacotron2-GST-GRA with decaying guided attention.

References

1. Tamura, M.; Masuko, T.; Tokuda, K.; Kobayashi, T. Adaptation of pitch and spectrum for HMM-based speech synthesis using MLLR. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Salt Lake City, UT, USA, 7–11 May 2001; Volume 2, pp. 805–808.
2. Skerry-Ryan, R.; Battenberg, E.; Xiao, Y.; Wang, Y.; Stanton, D.; Shor, J.; Weiss, R.; Clark, R.; Saurous, R.A. Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 4693–4702.
3. Wagner, M.; Watson, D.G. Experimental and theoretical advances in prosody: A review. *Lang. Cognit. Process.* **2010**, *25*, 905–945. [CrossRef] [PubMed]
4. Wang, Y.; Stanton, D.; Zhang, Y.; Ryan, R.S.; Battenberg, E.; Shor, J.; Xiao, Y.; Jia, Y.; Ren, F.; Saurous, R.A. Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; Volume 80, pp. 5180–5189.
5. Zen, H.; Tokuda, K.; Black, A.W. Statistical parametric speech synthesis. *Speech Commun.* **2009**, *51*, 1039–1064. [CrossRef]
6. Tokuda, K.; Nankaku, Y.; Toda, T.; Zen, H.; Yamagishi, J.; Oura, K. Speech synthesis based on hidden Markov models. *Proc. IEEE Inst. Electr. Electron Eng.* **2013**, *101*, 1234–1252. [CrossRef]
7. Zen, H.; Senior, A.; Schuster, M. Statistical parametric speech synthesis using deep neural networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 26–30 May 2013; pp. 7962–7966.
8. Yamagishi, J.; Masuko, T.; Kobayashi, T. MLLR adaptation for hidden semi-Markov model based speech synthesis. In Proceedings of the International Conference on Spoken Language Processing, Jeju, Korea, 4–8 Ocotober 2004.
9. Kim, N.S.; Sung, J.S.; Hong, D.H. Factored MLLR adaptation. *IEEE Signal Process. Lett.* **2010**, *18*, 99–102. [CrossRef]
10. Swietojanski, P.; Renals, S. Learning hidden unit contributions for unsupervised speaker adaptation of neural network acoustic models. In Proceedings of the IEEE Spoken Language Technology Workshop, South Lake Tahoe, NV, USA, 7–10 December2014; pp. 171–176.
11. Wu, Z.; Swietojanski, P.; Veaux, C.; Renals, S.; King, S. A study of speaker adaptation for DNN-based speech synthesis. In Proceedings of the Conference of the International Speech Communication Association, Dresden, Germany 6–10 September 2015.
12. Zhu, X.; Xue, L. Building a controllable expressive speech synthesis system with multiple emotion strengths. *Cognit. Syst. Res.* **2020**, *59*, 151–159. [CrossRef]
13. Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R.J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. Tacotron: Towards end-to-end speech synthesis. In Proceedings of the Interspeech, ISCA, Toronto, ON, Canada, 24–28 June 2017; pp. 4006–4010.
14. Tachibana, H.; Uenoyama, K.; Aihara, S. Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4784–4788.
15. Shen, J.; Pang, R.; Weiss, R.J.; Schuster, M.; Jaitly, N.; Yang, Z.; Chen, Z.; Zhang, Y.; Wang, Y.; Skerrv-Ryan, R.; et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Calgary, AB, Canada, 15–20 April 2018; pp. 4779–4783.
16. Bahdanau, D.; Cho, K.; Bengio, Y. Neural machine translation by jointly learning to align and translate. In Proceedings of the International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
17. Vinyals, O.; Kaiser, U.; Koo, T.; Petrov, S.; Sutskever, I.; Hinton, G. Grammar as a foreign language. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QB, Canada, 7–12 December 2015; pp. 2773–2781.

18. Gibiansky, A.; Arik, S.; Diamos, G.; Miller, J.; Peng, K.; Ping, W.; Raiman, J.; Zhou, Y. Deep voice 2: Multi-speaker neural text-to-speech. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–10 December 2017; pp. 2962–2970.
19. Jia, Y.; Zhang, Y.; Weiss, R.; Wang, Q.; Shen, J.; Ren, F.; Nguyen, P.; Pang, R.; Moreno, I.L.; Wu, Y. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. In Proceedings of the Advances in neural information processing systems, Montreal, QB, Canada, 3–8 December 2018; pp. 4480–4490.
20. Kwon, O.; Jang, I.; Ahn, C.; Kang, H.G. Emotional speech synthesis based on style embedded Tacotron2 framework. In Proceedings of the IEEE 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC), Jeju, Korea, 23–26 June 2019; pp. 1–4.
21. Chorowski, J.K.; Bahdanau, D.; Serdyuk, D.; Cho, K.; Bengio, Y. Attention-based models for speech recognition. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; pp. 577–585.
22. Feng, S.; Liu, S.; Li, M.; Zhou, M. Implicit distortion and fertility models for attention-based encoder-decoder NMT model. *arXiv* **2016**, arXiv:1601.03317.
23. Mnih, V.; Heess, N.; Graves, A. Recurrent models of visual attention. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014, pp. 2204–2212.
24. Zhang, B.; Xiong, D.; Xie, J.; Su, J. Neural machine translation with GRU-gated attention model. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**. [CrossRef] [PubMed]
25. Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; Bengio, Y. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, Doha, Qatar, 25–29 October 2014; pp. 1724–1734.
26. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. In Proceedings of the NIPS 2014 Workshop on Deep Learning, Montreal, QC, Canada, 8–14 December 2014.
27. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
28. Graves, A. Generating sequences with recurrent neural networks. *arXiv* **2013**, arXiv:1308.0850.
29. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning, Atlanta, GA, USA, 21–14 June 2010; pp. 807–814.
30. Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **2014**, *15*, 1929–1958.
31. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144.
32. Griffin, D.; Lim, J. Signal estimation from modified short-time Fourier transform. *IEEE Trans. Acoust. Speech Signal Process.* **1984**, *32*, 236–243. [CrossRef]
33. van den Oord, A.; Dieleman, S.; Zen, H.; Simonyan, K.; Vinyals, O.; Graves, A.; Kalchbrenner, N.; Senior, A.; Kavukcuoglu, K. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499.
34. Bengio, Y.; Simard, P.; Frasconi, P. Learning long-term dependencies with gradient descent is difficult. *IEEE Trans. Neural Netw.* **1994**, *5*, 157–166. [CrossRef] [PubMed]
35. The M-AILABS Speech Dataset. 2019. Available online: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/ (accessed on 30 March 2020).
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, u.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
37. Kuchaiev, O.; Ginsburg, B.; Gitman, I.; Lavrukhin, V.; Li, J.; Nguyen, H.; Case, C.; Micikevicius, P. Mixed-precision training for NLP and speech recognition with OpenSeq2Seq. *arXiv* **2018**, arXiv:1805.10387.
38. Yamamoto, R.; Song, E.; Kim, J.M. Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 6199–6203.

39. Christophe, V.; Junichi, Y.; Kirsten, M. CSTR VCTK Corpus: English Multi-Speaker Corpus for CSTR Voice Cloning Toolkit. 2016. Available online: https://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html (accessed on 12 September 2019).
40. P. 800: Methods for Subjective Determination of Transmission Quality. 1996. Available online: https://www.itu.int/ITU-T/recommendations/rec.aspx?rec=3638 (accessed on 30 March 2020).