SeDepTTS: Enhancing the Naturalness via Semantic Dependency and Local Convolution for Text-to-Speech Synthesis

Chenglong Jiang, Ying Gao*, Wing W. Y. Ng*,
Jiyong Zhou, Jinghui Zhong, Hongzhong Zhen

School of Computer and Engineering, South China University of Technology, Guangzhou, China
{csjiangcl_gx,202121044812,cshzzhen}@mail.scut.edu.cn, {gaoying,jinghuizhong}@scut.edu.cn, wingng@ieee.org

Abstract

Self-attention-based networks have obtained impressive performance in parallel training and global context modeling. However, it is weak in local dependency capturing, especially for data with strong local correlations such as utterances. Therefore, we will mine linguistic information of the original text based on a semantic dependency and the semantic relationship between nodes is regarded as prior knowledge to revise the distribution of self-attention. On the other hand, given the strong correlation between input characters, we introduce a one-dimensional (1-D) convolution neural network (CNN) producing query(**Q**) and value(**V**) in the self-attention mechanism for a better fusion of local contextual information. Then, we migrate this variant of the self-attention networks to speech synthesis tasks and propose a non-autoregressive (NAR) neural Text-to-Speech (TTS): SeDepTTS. Experimental results show that our model yields good performance in speech synthesis. Specifically, the proposed method yields significant improvement for the processing of pause, stress, and intonation in speech.

Introduction

Speech synthesis aims to generate natural and expressive speech from text. In recent years, the proposal of the seq2seq (Cho et al. 2014) paradigm has significantly improved the performance of neural TTS. Autoregressive (AR) neural TTS models are known to be able to generate speech with naturalness such as Tacotron (Wang et al. 2017). However, these models use recurrent neural network units, so parallel training is not applicable. Recently, Transformer-based NAR models (e.g., FastSpeech 1/2 (Ren et al. 2019, 2020)) have been proposed for the TTS task. Compared with AR models, the NAR neural TTS has unique advantages in synthesis speed and global context modeling. But many studies show that the self-attention mechanism has an inherent issue that dispersal the attention, which weakens the attention of local information (Li et al. 2019; Guo, Zhang, and Liu 2019; Pilault, Pal et al. 2020), especially for data with relatively strong local correlation, such as text sentences. Therefore, in this paper, we will focus on the local modeling performance of the self-attention mechanism, especially on the distribution of attention in self-attention. An example of our work as shown in Figure 1, we propose a method for fusing the semantic dependency prior knowledge into the attention distribution produced by self-attention.

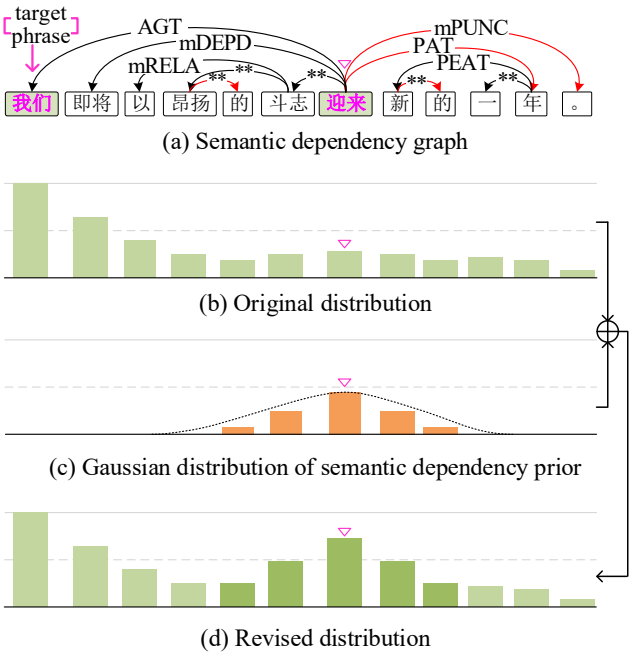


Figure 1: Semantic dependency prior. (a) illustrates the semantic dependency graph (SDG), where the phrase, connecting line and token indicates node, semantic dependency arc and semantic relation token, respectively. ▽ denote the central position corresponding to the target phrase;(b) The distribution of attention for each phrase under the original self-attention networks;(c) a Gaussian distribution of semantic dependence prior is introduced at the central position; (d) The attention corrected by the prior knowledge.

Prior knowledge can effectively enhance the local modeling performance of the self-attention mechanism (Yang et al. 2020; Guo, Zhang, and Liu 2019; Ying et al. 2021). Yang et al. (Yang et al. 2018) propose introducing a learnable Gaussian bias to reconstruct the attention distribution in the self-attention mechanism for machine transla-

*Corresponding Author

tion, then this method was extended to TTS tasks (Yang et al. 2020). Similarly, Gaussian prior was also used in Gaussian-Transformer (Guo, Zhang, and Liu 2019), which has achieved good results in the natural language classification task. However, those methods must establish a central position for Gaussian distribution to revise attention distribution. Usually, an algorithm is introduced to predict the corresponding central position for each query of self-attention. Then, the reliability of prediction results will directly affect the model performance. On the other hand, there may be multiple central positions in long sentences because richer linguistic information is contained. So, a single central position was not satisfied for multiple local dependents. Therefore, semantics prior knowledge is proposed to solve the above issues.

The text contains abundant linguistic knowledge such as syntax, semantics and pragmatics. For TTS tasks, mining linguistic information from the original text is also a key to achieving expressive TTS modeling (Tan et al. 2021). Traditional neural TTS models usually use phoneme sequences as input which haven't fully utilized the contextual semantic information of the target sentence. Therefore, many works improve the expressiveness of TTS by introducing syntactic information (Sun et al. 2020; Liu, Sisman, and Li 2021), those methods by explicitly associating input phoneme embedding with syntactic relations. A word-level semantic representation method is proposed in (Zhou et al. 2021) which is based on dependent structure and pre-trained BERT. Although the above works explore the validity of linguistic knowledge, the local modeling performance has not improved significantly on the TTS task.

In addition, considering a solid relationship between adjacent elements for text, convolutional neural networks (CNNs) are widely used to extract local semantic features at different levels to fully consider the context information of the target position (Soni, Chouhan, and Rathore 2022; Johnson and Zhang 2017; Kim 2014). In the tasks of time series forecasting, Li et al. (Li et al. 2019) propose a ConvTrans algorithm that uses a causal convolutional to solve the problems of local context. These works show that CNNs play an essential role in local modeling.

Inspired by previous research, this work proposes two approaches to address the local modeling of the self-attention mechanism. Here, the semantic relation is considered prior knowledge of the self-attention mechanism. This specific relationship avoids predicting the central position and solves the issue of multiple local dependence. Besides, 1-D CNNs are introduced to focus on local context information. Specifically, our proposed techniques are built on Transformer's self-attention mechanism. Firstly, SDG (Wang et al. 2018) was transformed into a semantic dependency matrix (SDM) with Gaussian distribution and each cell of the SDM represents a specific semantic relationship score. Then, the SDM as prior knowledge was added to the attention distribution produced by self-attention, where the direct semantic relationship yielded can better revise the distribution of attention. Secondly, we introduce 1-D CNNs producing **Q** and **V** in the self-attention mechanism to enhance the sensitivity of models for local contextual information, where **Q** and **V** are feature maps focusing on the local context. For each key(**K**), **V**-**Q** matching aware of local context by calculating the dot product of **K** and **Q**, which achieves the purpose of enhancing the local context attention. Lastly, we propose SeDepTTS: A speech synthesis modeling method with semantic dependency prior and local convolution.

Overall, the contributions of this paper are summarized as follows:

- Based on the self-attention mechanism, a novel local modeling enhancement method is proposed, which consists of semantic dependency prior and local convolution to provide linguistic information and local information, respectively.
- We migrate this variant of self-attention mechanism to speech synthesis tasks and propose a NAR neural TTS: SeDepTTS.
- Experimental results show that our model yields good performance in speech synthesis. Specifically, the naturalness of pause, stress and intonation has been significantly improved[1].

Related Works

Attention with Prior for Transformer

The approach of attention with prior is either to add new knowledge to the original attention map or to replace the standard attention with prior knowledge. It is observed that the attention distribution of adjacent layers is similar for Transformer, so one approach is to add the low-level attention map to current weight directly (He et al. 2021), which is identical to the residual network. Lazyformer (Ying et al. 2021) is proposed further to share the attention map among multiple adjacent layers. The advantage of this approach is that the attention map is computed only once and is reused many times in subsequent layers, thus reducing the computation cost. However, these methods mainly solve the problem of computational complexity, which is difficult to adjust the attention distribution at a specific position.

A learnable Gaussian bias is proposed to model localness for self-attention in (Yang et al. 2018). The bias as prior is then incorporated into the original attention distribution to form a revised distribution. This prior knowledge was applied to Transformer-TTS and the effect is significantly improved (Yang et al. 2020). Similarly, Gaussian distribution prior also achieved good results on multi-classification natural language reasoning tasks (Guo, Zhang, and Liu 2019). Furthermore, a more direct approach is to discard the generated distribution by self-attention and only use prior knowledge. You et al. (You, Sun, and Iyyer 2020) used Gaussian distribution as hard coding to calculate attention distribution, this is similar to the previous work (Yang et al. 2020; Guo, Zhang, and Liu 2019) that the attention distribution is focused on a local, it achieved better performance when applied to machine translation tasks. Although those self-attention variants have a specific improvement in the seq2seq studies, the Gaussian prior still needs an algorithm

[1]Synthesized speech samples can be found at: https://jclgx.github.io/

to predict the central position which may cause errors. Moreover, a single central position was not satisfied for multiple local dependents.

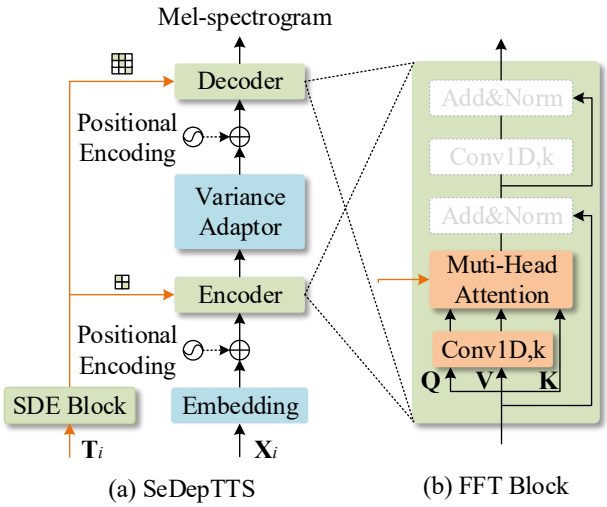
Convolution Neural Network for Text Mining

CNNs were usually designed to extract the context semantics of the target location, which is an effective method to enhance local modeling. The model in (Kim 2014) utilizes simple 1-D CNNs and performs remarkably well in sentiment analysis and question classification for text. This model has served as the benchmark architecture for many recent models. Given the limited text features of short texts, it is necessary to explore the characteristics of short texts from various angles. Therefore, Feng et al. (Feng and Cheng 2021) propose a novel sentiment analysis model based on multi-channel CNN with multi-head attention mechanisms. To extract remote relationship features from the text, the authors of (Johnson and Zhang 2017) propose a low-complexity word-level deep CNN architecture to generate an efficient representation of long-range associations in text. Most of the above works adopt 1-D convolving filters, where each filter specializes in features of a particular input word embedding. TextConvo-Net (Soni, Chouhan, and Rathore 2022) is proposed to adopt 2-D convolution filtering to extract the intra-sentence n-gram features and capture the inter-sentence n-gram features. Lots of work shows that CNNs play an essential role in text-mining tasks. Inspired by this, we introduce CNNs to alleviate the issue of the context semantics that were not considered in the original TTS task.

Non-autoregressive Neural TTS

The NAR neural TTS model can generate sequences of all tokens in parallel, giving faster speech generation speed and more flexible controllability. Ren et al. (Ren et al. 2020) propose FastSpeech1/2 which is a NAR neural TTS based on Transformer. Then, many NAR neural TTS models were developed to improve speech quality. Using dynamic programming in AlignTTS (Zeng et al. 2020) to train the duration model for text-to-Mel aligns. Shah et al. (Shah et al. 2021) propose a NAR model by replacing the attention module of the conventional attention-based TTS model with an external duration model for low-resource and highly expressive speech. Peng et al. (Peng et al. 2020) propose a ParaNet model for NAR TTS, extracting attention from the AR TTS model and then realizing the parallelization.

In addition, the generation model benefits NAR neural TTS, constructing a NAR TTS utilizing a deep variational autoencoder with a residual attention mechanism that subtly refines the text-to-sound alignment (Liu et al. 2021). Excessive smoothing is a severe problem for the NAR TTS model, which harms the performance of the NAR TTS model (Ren et al. 2022). Therefore, relevant work (Guo et al. 2022; Kim et al. 2020) is based on the richer conditional input information and enhanced modeling methods to strengthen the ability of the model fitting complex data distribution. Ye et al. (Ye et al. 2022) propose the SyntaSpeech, a lightweight NAR neural TTS sensitive to syntax. Besides, PortaSpeech is proposed in (Ren, Liu, and Zhao 2021), which is a way



(a) SeDepTTS (b) FFT Block

Figure 2: The over architecture for SeDepTTS. Blue and green blocks in subfigure (a) denote the basic structure of FastSpeech2 and new functional blocks introduced by SeDepTTS respectively. The Encoder (or Decoder) details are shown in subfigure(b). The SDE block of SeDepTTS is shown in Figure 3.

of integrating syntactic information into the prosodic module. Those methods improve the expressiveness of speech synthesis. In this work, we propose a novel approach to enhance the local modeling in self-attention-based and migrate it to NAR neural TTS.

Methodology

Framework

The overall architecture of SeDepTTS is shown in Figure 2. The backbone framework of our proposed model is FastSpeech2-based which includes the semantic dependency extractor (SDE) block, Encoder, Variance Adaptor and Decoder, where the Encoder (or Decoder) introduces a local convolution architecture that yields a better fusion of local contextual information. The input sequence is divided into phonemes $\mathbf{X}_i$ and the Chinese text $\mathbf{T}_i$. $\mathbf{X}_i$ combines positional encoding as Encoder input while $\mathbf{T}_i$ directly inputs the SDE block to extract semantic dependency prior knowledge. The details of the SDE block and local convolution are presented follow sections.

Semantic Dependency Extractor Block

The SDG is an extension of the syntax or semantic representation of the tree structure in which some words are allowed to have multiple parent nodes. The specific implementation of SDG is referred to as the (Wang et al. 2018). As shown in Figure 3, the SDE block includes four parts: Chinese word segmentation (CWS), semantic dependency parsing (SDP), SDM and Upsample. CWS and SDP are performed using jieba[2] and ltp[3], respectively. SDM and Upsample are de-

[2]https://github.com/fxsjy/jieba
[3]https://github.com/HIT-SCIR/ltp

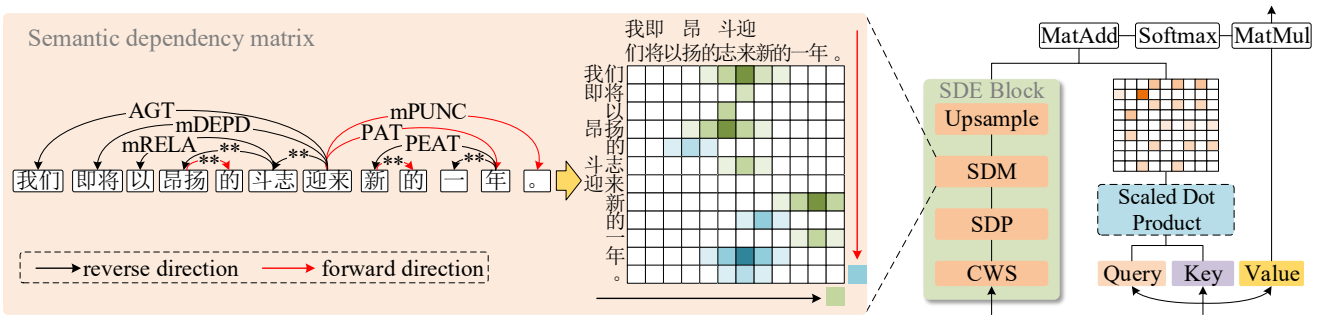


Figure 3: Semantic dependency attention module. The transcript is "We are about to usher in the New Year with high morale.".

scribed as follows.

In the Chinese text $\mathbf{T}_i$, assume that the embedding of semantic relation token between nodes is $\mathbf{e}$. Then, referring to the attention mechanism and the semantic dependency attention score between nodes can be expressed as follows:

$$\mathbf{s}_j = \frac{\mathbf{e}_j \cdot \mathbf{e}_j^{\mathrm{T}}}{\sqrt{d_e}}, \tag{1}$$

where, d_e and j are the dimensions of embedding and the j-th semantic relation tokens, respectively. The semantic relation phrase corresponding to the target phrase is taken as the central position to introduce a probability density function of Gaussian:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma}}, \tag{2}$$

μ is the center location, σ is a constant. Then, the semantic dependency matrix is defined as follows:

$$\begin{aligned}\mathbf{S}_j =& [\cdots, f(x_i) \times \mathbf{s}_j, \cdots, f(x_1) \times \mathbf{s}_j, f(x_0) \times \mathbf{s}_j, f(x_1) \\ & \times \mathbf{s}_j, \cdots, f(x_i) \times \mathbf{s}_j, \cdots], x_i = \mu + i (i \in \mathbb{N}_+).\end{aligned} \tag{3}$$

$$\mathbb{S}_{j\times j} = [\mathbf{S}_1, \mathbf{S}_2, \cdots, \mathbf{S}_j]^{\mathrm{T}}, \tag{4}$$

where x_i is the position around the central position μ. $\mathbf{S}_j$ is the j-th tokens, we calculate both the semantic dependency score of the central position and its neighbors, then all the tokens are combined into matrix $\mathbb{S}_{j\times j}$. In Vaswani et al. (Vaswani et al. 2017), Transformer adopts a point-multiple attention mechanism and is defined as:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}})\mathbf{V}, \tag{5}$$

Where $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ indicate query, key and value, $\sqrt{d}$ is the scaling factor related to the dimension of $\mathbf{K}$. In Figure 3, the SDM of the SDE block is transformed according to the description in Formula (1-4), and each of the cells in the SDM represents scores between nodes in SDG. In this way, we create absolute prior knowledge based on SDG.

Because the SDG adopts phrases as nodes while the self-attention takes phonemes as input in the TTS task, alignment is needed before adding the SDM to the attention map of self-attention. Here, we introduce Upsample method which is expanding by the number of phonemes of the nodes to obtain: $\mathbb{S}_{j\times j} \Rightarrow \widehat{\mathbb{S}}_{m\times m}, m \geq j$. Note that the Encoder and Decoder use different Upsample dimensions. Then, the attention with prior is redefined as follows:

$$Attention(\mathbf{Q}, \mathbf{K}, \mathbf{V})_{\mathbf{T}_i} = Softmax[\widehat{\mathbb{S}}_{m\times m} + \frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{d}}]. \tag{6}$$

Local Convolution

Local convolution consists of a two-layer feedforward bottleneck layer with ReLU activation, which employs 1-D CNNs with upsampling and downsampling. Before the point-multiple attention calculation of self-attention, the kernel sizes consisting of $k \times n$ and $1 \times n$ are used to local convolution for $\mathbf{Q}$ and $\mathbf{V}$, respectively. Where k is the hyperparameter (i.e., windows size of target location context), n denotes latent feature vector dimension. $\mathbf{K}$ value is original without transformation to ensure that the superposition of semantic dependency prior knowledge is meaningful. $\mathbf{Q}$ and $\mathbf{V}$ are processed with the same structure.

Experiments

We conduct experiments on the single-speaker dataset (BIAOBEI), a public dataset provided by Biaobei Technology Company[4]. The BIAOBEI dataset consists of 10,000 sentences with each consisting of a Mandarin speech audio and text annotation of the speech. The effective voice duration of the data is about 12 hours and the voice sampling rate is 22.05kz. Hopping sizes of 12.5ms and window sizes of 50ms are used to extract the sound spectrogram. In the experiment, the corpus data are randomly divided into 9388 sentences for the training set and 512 sentences for the validation set. The remaining 100 sentences are used as the testing set.

The structure of the SeDepTTS is FastSpeech2-based, where the hyperparameters of the variance adapter, Encoder and Decoder in the model are the same as the original. In local convolution, the filter size of upsampling and downsampling are 9×1 and 1×1, respectively, and the corresponding number of channels is 1024 and 256, respectively. In the SDE block, $x_i \in [u, u+2] \in \mathbb{N}_+$ and σ is set to $1/\sqrt{2\pi}$

[4]https://www.data-aker.com/open_source.html

for the probability density function of Gaussian (Formula (2)). Adam optimizer (Kong, Kim, and Bae 2020) was used with β_1 = 0.9, β_2 = 0.98 and $\epsilon = 10^{-9}$, and the learning rate schedule referring to (Vaswani et al. 2017). The total number of training steps is 100,000 and gradient clipping is performed every 30,000. To reconstruct audio from the predicted Mel-spectrogram, we trained a HiFi-GAN (Kong, Kim, and Bae 2020) vocoder which serves on the baselines and our SeDepTTS with the same standard.

Baseline Methods

Three baseline models are used for comparison. The baseline model we chose is based on the principle that only phonemes are used as input in the training and prediction stages. The comparison models are all Biaobei datasets. Each baseline model is described in detail below.

- FastSpeech2: Our model structure followed the non-autoregression architecture and the hyperparameters used by FastSpeech2.
- SAG-Tacotron: Yang et al. (Yang et al. 2018, 2019) propose a local modeling method of Gaussian distribution in self-attention, which is migrated to the speech synthesis tasks. Similar to our proposed method, attention with prior is used to solve local modeling issues. In addition, part of synthetic speech samples[5] are disclosed in SAG-Tacotron, which makes the comparison more credible.
- BERT-Dep: This model is based on dependent structure and pre-trained BERT to generate word-level semantic representation information, which is fused into the latent representations of Tacotron2 as additional features (Zhou et al. 2021). It is the latest based on semantic information to improve the naturalness and expressiveness of speech synthesis[6].

Evaluation Methods

Objective evaluation indicators:(1) Mel cepstral distortion (MCD) (Kubichek 1993) is used to evaluate the synthetic speech quality. MCD score is to quantify the distortion degree between two audio Mel frequency cepstral coefficients, a smaller MCD indicates a better-synthesized speech quality. In practice, Dynamic time warps (Kruskal 1983) are used to align the predicted spectra with the target. (2) R^2 is the goodness of fit to evaluate the gap of F0, the larger the R^2 value, the better the performance is.

Subjective evaluation indicators: Subjective evaluation is scored by listening to the audio. (1) Mean Opinion Score (MOS) is used for evaluation in the case of raw audio reference. MOS test rules are that participants score the given audio on a range from 1 to 5, where 1 means very poor and 5 means excellent. (2) In the absence of raw audio, the performance of speech expression was evaluated by the ABX preference test. Participants were asked to choose one over the other based on their feelings about listening to two audio clips. Alternatively, participants could remain neutral when they thought the two sounds were close. All subjective rating experiments were conducted in a quiet studio by native Mandarin-speaking wearing headphones. Each audio was listened to at least three times to get an average rating. In particular, we ask the participants to focus on the authenticity of the sound quality, the placement of the pauses, and the expressiveness of the sound. In addition to the objective and subjective rating, we draw a spectrum and F0 curves to more intuitively show the differences between audio.

[5]https://fyyang1996.github.io/as191245/

[6]https://thuhcsi.github.io/interspeech2022-dependency-semantic-tts/

Experimental Results

Subjective Evaluation

Test audio is model-synthesized or disclosed by the baseline. Please refer to the section Evaluation Methods for specific test procedures. The MOS test results are shown in Table 1, MOS of SeDepTTS is 4.10 $\pm$ 0.13, which is higher than the baseline's, where SAG-Tacotron and BERT-Dep both are auto-regressive (AR) TTS models. MOS score results show that speeches generated by SeDepTTS have pleasing naturalness in the sense of hearing, even beyond the AR TTS.

Model	MOS↑
Ground truth	4.43±0.12
FastSpeech2 (NAR)	3.72±0.12
SAG-Tacotron (AR)	3.97±0.13
BERT-Dep (AR)	3.93±0.10
SeDepTTS (NAR)	$\mathbf{4.10 \pm 0.13}$

Table 1: The MOS of different models with 95% confidence intervals.

FastSpeech2	SAG-Tacotron	BERT-Dep	N/P	SeDep-TTS ↑	$p \downarrow$
11.8	/	/	23.4	**64.8**	<0.001
/	18.2	/	**54.5**	27.3	0.061
/	/	16.8	25	**58.2**	<0.001

Table 2: Results of the ABX preference tests (%) among the three baselines and the proposed model. N/P denotes "no preference" and p means the p-value of a t-test between two models.

The ABX preference test is a direct auditory perception of the tester for audio, and a better choice is made between the baseline and the proposed, results as shown in Table 2. It shows that the average preference scores for SeDepTTS outperform the baseline model by a significant margin. Except for the SAG-Tacotron model, the proposed method significantly outperforms the others (p-value smaller than 0.05).

Noted that, to observe the difference among the synthesized speech visually, we analyzed a case of speech presented in the Mel spectrogram, including pause, stress, intonation, etc. The Mel spectrogram of a natural recording and other synthesized speech as shown in Figure 4, compared with different baselines, the synthesized speech produced by our proposed model is more similar and natural to the ground truth. In Figure 4, a-i to a-iv columns is zoom-in of

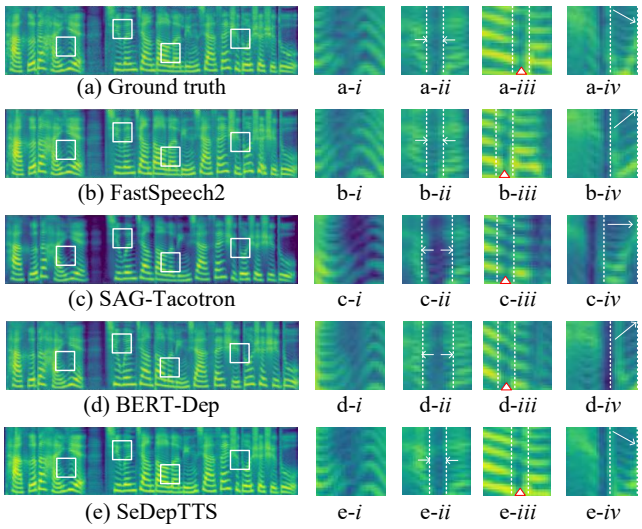


Figure 4: The Mel spectrograms of one case. The transcript is "The cold night in *Tahe*, the temperature dropped to more than minus thirty degrees Celsius.".

the local spectrogram marked by rectangles in the left-most column which mainly sampling pauses between phrases and details in the different moments. Comparing the local Mel spectrogram from a-i to e-i, the Mel spectrogram generated by SAG-Tacotron is darker overall meaning that the voice sounded muffled. Then, we observe a-ii column, the c-ii and d-ii have been a significant expansion of silent time than a-ii. The column of a-iii indicates that SeDepTTS yields better alignments with the natural spectrogram and the pause position is closer to the ground truth. At the same time, the other three baseline models have poor alignments. The right-most column illustrates that SeDepTTS has better intonation performance. Here, ground truth intonation is a stressed falling tone, while FastSpeech2 (b-iv) and BERT-Dep (d-iv) both stressed rising tones, SAG-Tacotron (c-iv) is a soft even voice. In contrast, SeDepTTS (e-iv) is a complete falling tone and consistent with a natural tone. These results indicate that SeDepTTS yields more accurate inferences of audios' durations and retains better details.

Objective Evaluation

We first compared the distortion of synthetic audio with real audio by MCD in different models, allowing us to compare the gap between those different models objectively. On the other hand, we further analyze the audio's fundamental frequency(F0) using the YIN algorithm (De Cheveigné and Kawahara 2002). We use the goodness of fit to evaluate the gap of F0 where the goodness of fit is computed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - Y_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y_i})^2}, \quad (7)$$

where Y_i is the real value, y_i is the reference value, and Y is the average of the real value $\overline{Y_i}$. The range of R^2 value is $(-\infty, 1]$ and the larger the better. The objective evaluation results are presented in Table 3, showing that

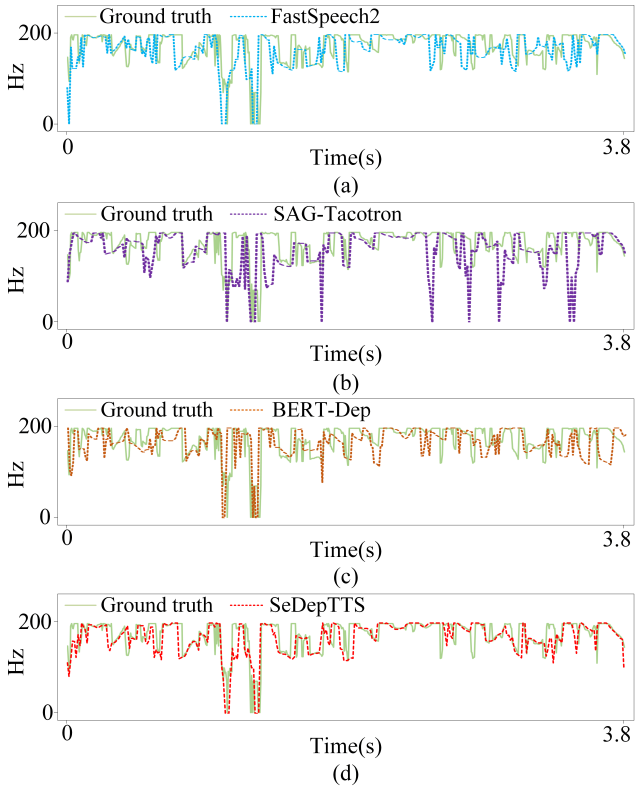


Figure 5: F0 contours of audios. Note that F0 contours for audios generated by other models are aligned to the ground truth.

the MCD of SeDepTTS is reduced by about 12.7%, 5.19% and 7.12% compared with FastSpeech2, SAG-Tacotron and BERT-Dep, respectively. Our model significantly outperforms other models. The R^2 means that SeDepTTS is closest to the real pitch while SAG-Tacotron is the opposite.

Model	MCD↓	R^2 ↑
FastSpeech2	6.28	0.784
SAG-Tacotron	5.78	0.607
BERT-Dep	5.90	0.678
SeDepTTS	**5.48**	**0.838**

Table 3: The performance of different models is evaluated with MCD and R^2.

Figure 5 shows F0 contours for audios of ground truth, FastSpeech2, SAG-Tacotron, BERT-Dep and SeDepTTS, respectively. In Figure 5, the F0 contour of FastSpeech2 (Figure 5(a)) has significant shifts, especially in the latter half. This point is consistent with the Mel spectrogram of Figure 4. The fluctuation range of the F0 contour is significantly more extensive than that of other models in Figure 5(b), which is also the reason for the voice sounding muffled and the lowest of R^2. Although the frequency fluctuation range in Figure 5(c) is generally close to the ground truth, the degree of fit is not good. So, the R^2 of BERT-Dep

is also low. In contrast, our model (Figure 5(d)) is superior to the three baselines in both ranges of fluctuation and degree of fit. Corresponding R^2 of SeDepTTS is the highest than others.

Ablation Studies

In this section, the validity of several techniques in the proposed method is verified through specific studies, mainly including SDE and local convolution (LC). On the one hand, subjective evaluation is used to evaluate our proposed techniques. MOS test results are shown in Table 4 and ABX preference test results are shown in Table 5. We could observe that adding SDE and LC to the self-attention improves about 10 percent on MOS, and using SDE has excellent effects on the score of models, the model yields the worst MOS when SDE is not used, indicating the significance of the semantic dependency prior. LC is also indispensable, when LC was added, the MOS score was significantly improved. The average preference scores of SeDepTTS in Table 5 also illustrate that SeDepTTS outperform other models without SDE or LC. Significantly, the SDE block is essential for the SeDepTTS when without the SDE($-$SDE) and most testers will directly choose the SeDepTTS in the ABX preference test.

Methods	MOS↑
SeDepTTS	$\mathbf{4.10 \pm 0.13}$
$-$SDE	3.79±0.10
$-$LC	3.93±0.11
$-$SDE$-$LC	3.72±0.12

Table 4: Ablation studies of proposed techniques.

SeDepTTS	-SDE-LC	-SDE	-LC	N/P	$p\downarrow$
64.8	11.8	/	/	23.2	<0.001
56.7	/	16.7	/	26.6	<0.001
40.6	/	/	31.3	28.1	0.084

Table 5: Results of the ABX preference tests (%) among the proposed techniques. N/P denotes "no preference" and p means the p-value of a t-test between two models.

On the other hand, we show the changing process of Mel spectrograms in one case, which is affected by the proposed techniques, as shown in Figure 6. In the figure, the red and pink boxes were used to mark the positions of mood pauses and intonation changes, respectively. Figure 6(a) shows that an audio Mel spectrogram is generated by ground truth. When SDE and LC are not considered in our model, the audio Mel spectrogram caused by this structure as shown in Figure 6(b), the intonation of the Mel spectrogram enclosed by the pink box is soft even in Mandarin and it is not consistent with the ground truth. Moreover, the tone pause is also not significant enough in the red box compared to the ground truth. Figure 6(c) shows that after the LC is introduced into our model and the intonation of Mel spectrogram enclosed by the pink box was changed, the intonation of the phrase "*tend to*" had two consecutive falling tones. At the same time, the raw recordings should be first rising and then falling in this position. After introducing the SDE block, Mel spectrogram shown in Figure 6(d), we can observe a clear pause between the "*The reason is that*" and "*many donors*", moreover, the Mel spectrogram enclosed by the pink was also first rising and then falling, those changes indicate that the utterance's intonation more expressive. These ablation studies verify the effectiveness of our proposed techniques in improving speech quality.

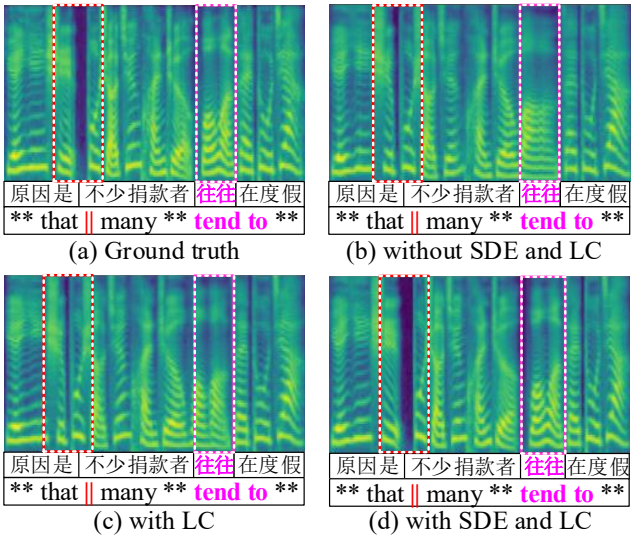


Figure 6: The evolution of the Mel spectrograms. The red box ("||") and pink box ("*tend to*") indicate the position of pauses and intonation, respectively. The ground-truth intonation enclosed by the pink box rises and then falls. The transcript is "The reason is that many donors tend to be on vacation.".

Conclusion

This paper proposed a variant of the self-attention mechanism and migrated it to the TTS tasks. Based on the semantic dependency prior knowledge, a Gaussian distribution is introduced into the semantic relation nodes to generate an SDM to revise the attention distribution generated by self-attention. In addition, we present 1-D CNNs producing **Q** and **V** in the self-attention mechanism to enhance the context information of the tag location. Experimental results show that speech synthesized by our proposed has a better naturalness than the three baselines because the proposed method performs better in alignment, pauses and intonation. Ablation studies confirmed the effectiveness of the proposed techniques, especially the SDE block is of great significance.

Furthermore, Chinese sentences have rich semantics and different semantics greatly influence the prosodic expression of speech, especially in the professional field of audio reading, such as performing a poetry recitation with emotion. Therefore, one of our future works is to mine semantic information to improve the prosodic performance of speech synthesis in the professional field.

Acknowledgments

This work was partly supported by the International Cooperation Project of Guangdong Province under Grant 2021A0505030017, the National Natural Science Foundation of China under Grant 61876066 and the Program for Guangdong Introducing Innovative and Enterpreneurial Teams (Grant No.: 2017ZT07X183).

References

Cho, K.; van Merrienboer, B.; Gülçehre, Ç.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

De Cheveigné, A.; and Kawahara, H. 2002. YIN, a fundamental frequency estimator for speech and music. *The Journal of the Acoustical Society of America*, 111(4): 1917–1930.

Feng, Y.; and Cheng, Y. 2021. Short text sentiment analysis based on multi-channel CNN with multi-head attention mechanism. *IEEE Access*, 9: 19854–19863.

Guo, H.; Lu, H.; Wu, X.; and Meng, H. 2022. A Multi-Scale Time-Frequency Spectrogram Discriminator for GAN-based Non-Autoregressive TTS. *arXiv preprint arXiv:2203.01080*.

Guo, M.; Zhang, Y.; and Liu, T. 2019. Gaussian transformer: a lightweight approach for natural language inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 6489–6496.

He, R.; Ravula, A.; Kanagal, B.; and Ainslie, J. 2021. RealFormer: Transformer Likes Residual Attention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 929–943.

Johnson, R.; and Zhang, T. 2017. Deep pyramid convolutional neural networks for text categorization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 562–570.

Kim, J.; Kim, S.; Kong, J.; and Yoon, S. 2020. Glow-tts: A generative flow for text-to-speech via monotonic alignment search. *Advances in Neural Information Processing Systems*, 33: 8067–8077.

Kim, Y. 2014. Convolutional Neural Networks for Sentence Classification. In *EMNLP*.

Kong, J.; Kim, J.; and Bae, J. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in Neural Information Processing Systems*, 33: 17022–17033.

Kruskal, J. B. 1983. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM review*, 25(2): 201–237.

Kubichek, R. 1993. Mel-cepstral distance measure for objective speech quality assessment. In *Proceedings of IEEE pacific rim conference on communications computers and signal processing*, volume 1, 125–128. IEEE.

Li, S.; Jin, X.; Xuan, Y.; Zhou, X.; Chen, W.; Wang, Y.-X.; and Yan, X. 2019. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting. *Advances in neural information processing systems*, 32.

Liu, P.; Cao, Y.; Liu, S.; Hu, N.; Li, G.; Weng, C.; and Su, D. 2021. Vara-tts: Non-autoregressive text-to-speech synthesis based on very deep vae with residual attention. *arXiv preprint arXiv:2102.06431*.

Liu, R.; Sisman, B.; and Li, H. 2021. Graphspeech: Syntax-aware graph attention network for neural speech synthesis. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6059–6063. IEEE.

Peng, K.; Ping, W.; Song, Z.; and Zhao, K. 2020. Non-autoregressive neural text-to-speech. In *International conference on machine learning*, 7586–7598. PMLR.

Pilault, J.; Pal, C.; et al. 2020. Conditionally Adaptive Multi-Task Learning: Improving Transfer Learning in NLP Using Fewer Parameters & Less Data. In *International Conference on Learning Representations*.

Ren, Y.; Hu, C.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2020. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. In *International Conference on Learning Representations*.

Ren, Y.; Liu, J.; and Zhao, Z. 2021. Portaspeech: Portable and high-quality generative text-to-speech. *Advances in Neural Information Processing Systems*, 34: 13963–13974.

Ren, Y.; Ruan, Y.; Tan, X.; Qin, T.; Zhao, S.; Zhao, Z.; and Liu, T.-Y. 2019. Fastspeech: Fast, robust and controllable text to speech. *Advances in Neural Information Processing Systems*, 32.

Ren, Y.; Tan, X.; Qin, T.; Zhao, Z.; and Liu, T.-Y. 2022. Revisiting Over-Smoothness in Text to Speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 8197–8213.

Shah, R.; Pokora, K.; Ezzerg, A.; Klimkov, V.; Huybrechts, G.; Putrycz, B.; Korzekwa, D.; and Merritt, T. 2021. Non-autoregressive tts with explicit duration modelling for low-resource highly expressive speech. *arXiv preprint arXiv:2106.12896*.

Soni, S.; Chouhan, S. S.; and Rathore, S. S. 2022. TextConvoNet: A Convolutional Neural Network based Architecture for Text Classification. *arXiv preprint arXiv:2203.05173*.

Sun, A.; Wang, J.; Cheng, N.; Peng, H.; Zeng, Z.; and Xiao, J. 2020. Graphtts: graph-to-sequence modelling in neural text-to-speech. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6719–6723. IEEE.

Tan, X.; Qin, T.; Soong, F.; and Liu, T.-Y. 2021. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Che, W.; Guo, J.; and Liu, T. 2018. A neural transition-based approach for semantic dependency graph parsing. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Wang, Y.; Skerry-Ryan, R.; Stanton, D.; Wu, Y.; Weiss, R. J.; Jaitly, N.; Yang, Z.; Xiao, Y.; Chen, Z.; Bengio, S.; et al. 2017. Tacotron: Towards End-to-End Speech Synthesis. *Proc. Interspeech 2017*, 4006–4010.

Yang, B.; Tu, Z.; Wong, D. F.; Meng, F.; Chao, L. S.; and Zhang, T. 2018. Modeling Localness for Self-Attention Networks. In *EMNLP*.

Yang, F.; Yang, S.; Zhu, P.; Yan, P.; and Xie, L. 2019. Improving mandarin end-to-end speech synthesis by self-attention and learnable Gaussian bias. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, 208–213. IEEE.

Yang, S.; Lu, H.; Kang, S.; Xue, L.; Xiao, J.; Su, D.; Xie, L.; and Yu, D. 2020. On the localness modeling for the self-attention based end-to-end speech synthesis. *Neural networks*, 125: 121–130.

Ye, Z.; Zhao, Z.; Ren, Y.; and Wu, F. 2022. SyntaSpeech: Syntax-Aware Generative Adversarial Text-to-Speech. *arXiv preprint arXiv:2204.11792*.

Ying, C.; Ke, G.; He, D.; and Liu, T.-Y. 2021. LazyFormer: Self Attention with Lazy Update. *arXiv preprint arXiv:2102.12702*.

You, W.; Sun, S.; and Iyyer, M. 2020. Hard-Coded Gaussian Attention for Neural Machine Translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7689–7700.

Zeng, Z.; Wang, J.; Cheng, N.; Xia, T.; and Xiao, J. 2020. Aligntts: Efficient feed-forward text-to-speech system without explicit alignment. In *ICASSP 2020-2020 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 6714–6718. IEEE.

Zhou, Y.; Song, C.; Li, J.; Wu, Z.; Bian, Y.; Su, D.; and Meng, H. 2021. Enhancing Word-Level Semantic Representation via Dependency Structure for Expressive Text-to-Speech Synthesis. *arXiv e-prints*, arXiv–2104.