IMPROVING SPEECH PROSODY OF AUDIOBOOK TEXT-TO-SPEECH SYNTHESIS WITH ACOUSTIC AND TEXTUAL CONTEXTS

Detai Xin[1,2*], *Sharath Adavanne*[1], *Federico Ang*[1], *Ashish Kulkarni*[1],
Shinnosuke Takamichi[2], *Hiroshi Saruwatari*[2]
Rakuten Group, Inc., Japan[1],
Graduate School of Information Science and Technology, The University of Tokyo, Japan[2]

ABSTRACT

We present a multi-speaker Japanese audiobook text-to-speech (TTS) system that leverages multimodal context information of preceding acoustic context and bilateral textual context to improve the prosody of synthetic speech. Previous work either uses unilateral or single-modality context, which does not fully represent the context information. The proposed method uses an acoustic context encoder and a textual context encoder to aggregate context information and feeds it to the TTS model, which enables the model to predict context-dependent prosody. We conducted comprehensive objective and subjective evaluations on a multi-speaker Japanese audiobook dataset. Experimental results demonstrate that the proposed method significantly outperforms two previous works. Additionally, we present insights about the different choices of context - modalities, lateral information and length - for audiobook TTS that have never been discussed in the literature before.

Index Terms— text-to-speech synthesis, TTS, audiobook, speech prosody, context modeling

1. INTRODUCTION

Recent text-to-speech (TTS) systems based on deep neural networks (DNNs) have been able to synthesize natural read-out speech [1, 2, 3]. However, how to synthesize speech with a lot of prosody variations like audiobooks remains unsolved. Synthesizing such speech requires the system to not only transform linguistic information but also para-/non-linguistic information such as emotions, and intentions from text to speech [4, 5, 6]. Speech prosody in audiobooks produced by professional speakers depends on several factors including characteristics, context, and styles (narration or dialogue) [7]. Among these factors, context, either acoustic or textual, is popularly utilized in the literature to improve the prosody of audiobook TTS [8, 9, 10, 11]. This is because (1) consecutive utterances always have sequential relations in audiobooks; (2) unlike characteristics and other factors, context requires no additional cost to get.

However, existing work either (1) uses only single-modality context or (2) uses unilateral context which cannot fully leverage the power of context information. Gallegos et al. first proposed to use acoustic context to improve the prosody of audiobook TTS [8]. In their following work, they further used acoustic and textual contexts in audiobook TTS [9], but only preceding contexts were used. Xu et al. first proposed to use pretrained bidirectional encoder representations from transformers (BERT) [12] to encode preceding and succeeding sentences to incorporate textual context information in audiobook TTS [10]. Nakata et al. also used BERT embeddings, but only in an implicit way by encoding the target sentences with the bilateral context [11]. All of these works only used textual context with one or two sentences without trying longer context.

In this paper, we present a multi-speaker Japanese audiobook TTS system that fully and explicitly utilizes preceding acoustic context and bilateral textual context to improve the prosody of synthetic speech. The proposed method first uses an acoustic context encoder (ACE) [8] to encode preceding mel-spectrograms as acoustic context representations. Moreover, we propose a textual context encoder (TCE) based on attention mechanisms to extract textual context representations from BERT embeddings of bilateral textual context. These context representations are then fed to a multi-speaker TTS model to guide it to synthesize the target utterance with appropriate prosody. We conducted comprehensive experiments with both objective and subjective evaluations to verify the effectiveness of the proposed method. We further compared how different modalities, laterals, and lengths of context influence the prosody of audiobook TTS, which was never studied in previous work. Our contributions are summarized as follows:

- We propose a multi-speaker Japanese audiobook TTS system with acoustic and textual context encoding mechanisms.
- We conduct comprehensive objective and subjective experiments to show the effectiveness of the proposed method.
- We further conduct experiments to find the best combination of context modalities, laterals, and lengths for audiobook TTS.

Audio samples are publicly available[1].

2. RELATED WORK

Methods of audiobook TTS can be roughly grouped into two categories: single-sentence and multi-sentence methods. Single-sentence methods synthesize one utterance at a time, and improves speech prosody by feeding auxiliary features like context [8, 9, 10], emotions [13], and features extracted by DNNs like variational autoencoder [14, 15] and global style tokens (GST) [16]. Sufficient and correct prosody information should be maintained to guarantee satisfactory speech quality in such methods.

Multi-sentence methods, on the other hand, synthesize multiple sentences at a time. The length of the target sequence ranges from three sentences [17] to a whole paragraph [18]. While such methods are potentially better than single-sentence methods, it requires more memory and sophisticated training process.

The proposed method in this work is a single-sentence method utilizing both acoustic and textual contexts.

3. PROPOSED METHOD

The general architecture of the proposed method is illustrated in Figure 1. The model contains three components: an improved multi-

*This work was supported by JST SPRING, Grant Number JPMJSP2108.

[1] https://aria-k-alethia.github.io/2022rat-demo/

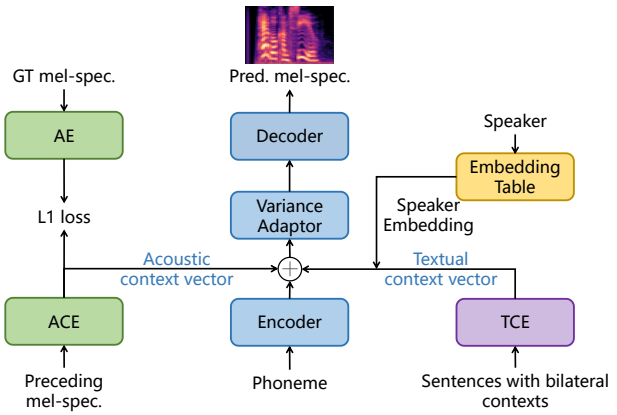

Fig. 1. Architecture of the proposed method.

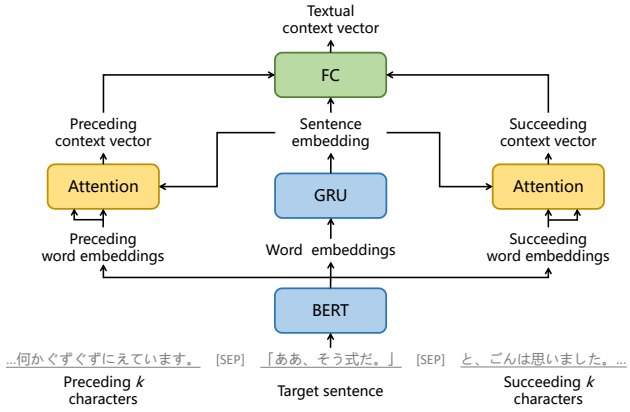

Fig. 2. Architecture of TCE. FC denotes fully connected layer.

speaker FastSpeech2 [3] for mel-spectrograms synthesis, an acoustic context encoder and a textual context encoder for context encoding. We introduce these components separately in the following sections.

3.1. Improved multi-speaker FastSpeech2

We follow the original FastSpeech2 [3] to construct the TTS model that contains a phoneme encoder, a variance adaptor, and a mel-spectrograms decoder. To adapt the model to a multi-speaker setting, we first create a look-up embedding table for speakers. The speaker embedding is summed to the output of the phoneme encoder and fed to the following modules to generate mel-spectrograms for the corresponding speaker. Second, we also use speaker-dependent pitch normalization [19] to disentangle speaker information from pitch values. Specifically, for a pitch value p_s of speaker s of a voiced frame, we normalize it to $\bar{p}_s$ by: $\bar{p}_s = \frac{p_s - \mu_s}{\sigma_s}$, where μ_s, σ_s are the mean and standard deviation values of the pitch of the speaker s, respectively.

We also notice that the utterances of audiobooks are longer than utterances in read-out corpus, but the absolute positional encoding [20] used in the original FastSpeech2 cannot well handle long sequences. To solve this problem, we replace the absolute positional encoding with relative positional encoding [21] so that the model can handle sequences with any length. We follow the previous work [21] and add relative positional embeddings to the attention layers of both the encoder and the decoder. The clipping distance is set to 4 so that the model can capture relative position differences within this value. For more details please refer to the original paper.

Finally, the original FastSpeech2 uses a length regulator in the variance adaptor before the pitch and energy predictors so that the model outputs frame-level pitch and energy, which makes the synthesized speech unstable. Therefore we move the length regulator after the two predictors to learn phoneme-level pitch and energy. In our preliminary experiments, we found all of the aforementioned modifications could improve the overall performance.

The deterministic nature of the above model makes it difficult to synthesize speech with various prosodies. Therefore, the proposed method further uses two context encoders to incorporate context information into the model.

3.2. Acoustic context encoding

In audiobooks it can be assumed that the prosodies of consecutive utterances have minor differences, hence using acoustic context can intuitively make consecutive utterances more coherent in single-sentence audiobook TTS methods. To this end, the proposed method uses ACE to encode acoustic context. Since during inference, only preceding utterances are available, ACE only encodes preceding acoustic context. Supposing the index of the target utterance for synthesis is N, we use GST as the implementation of ACE to extract a fixed-length acoustic context vector from the $(N-1)$-th mel-spectrogram. The context vector is summed to the output of the phoneme encoder to assist the synthesis of the N-th mel-spectrogram. Note that, although during training we use the ground truth (GT) $(N-1)$-th mel-spectrogram as the input of ACE, during inference we use the synthesized one as the input.

Following previous work [8], we set an extra next-prediction task for ACE. As Figure 1 illustrates, we use an acoustic encoder (AE) to extract a fixed-length vector from the N-th mel-spectrogram. Here AE is implemented as another GST module that has different parameters from the one of ACE. We then minimize the L1 distance between the two vectors extracted by ACE and AE. While previous work doesn't explain why the extra task is effective [8], we suppose this is because the extra task forces ACE to learn a corresponding relation between the $(N-1)$-th and the N-th mel-spectrograms. During training, we add the L1 loss term to the loss function of the TTS model and jointly train the whole model.

3.3. Textual context encoding

Textual context is also an informative source for producing appropriate prosody. Phrases like "a man/woman says", and "happily/sadly" in the contexts can be useful cues to predict the prosody of the target sentence. Therefore we propose a textual context encoder (TCE) to incorporate textual context information in the model. The architecture of TCE is illustrated in Figure 2. First, the bilateral context and the target text are fed to a pretrained BERT model to extract word embeddings. Note that, different from most previous works [9, 10, 11], we set the context length of each lateral to k characters instead of setting it to a certain number of sentences, by which we can easily evaluate model performance with different context lengths. The word embeddings of the target sentence are then fed to a gated recurrent unit (GRU) [22], and the last hidden state of GRU is used as the sentence embedding of the target sentence. Next, the sentence embedding is used as a query vector in two attention modules with the word embeddings of preceding and succeeding contexts as keys and values to extract bilateral context vectors. Finally, the two context vectors and the sentence embedding are concatenated together and fed to a fully connected (FC) layer to get the textual context vector. This vector is then summed to the output of

the phoneme encoder to incorporate textual context information in the TTS model.

4. EXPERIMENTS

4.1. Setup

We used J-MAC, a multi-speaker Japanese audiobook corpus produced by professional speakers, as the dataset [23]. We selected 9 speakers from J-MAC who at least have 3 books to ensure each speaker has sufficient data for training. We then randomly picked up gender-balanced 6 speakers from the 9 speakers as the test speakers and excluded one audiobook for each test speaker from the training set as the test set. The final training set contained about 7 hours of audio data. For the BERT model in TCE, we used "bert-base-japanese-v2"[2] pretrained on 4GB Japanese Wikipedia data. We used the output of the last layer of the BERT model as the word embeddings used in TCE. OpenJTalk[3] was used to convert Japanese characters to phonemes. For the forced alignment we used Julius [24] to get the duration of each phoneme. The pitch values were extracted by the WORLD vocoder [25]. We used the pretrained "`UNIVERSAL_V1`" HiFi-GAN model[4] to convert mel-spectrograms into time-domain waveforms.

For the TTS model we used the same parameter setting as the one of the previous work [3] except for the modifications mentioned in Section 3.1. The dimension of the speaker embedding was set to 256. In ACE, the GST token number was set to 10. Following the original work [26], we used multi-head attention with 8 heads to improve the robustness. In TCE, the number of hidden units in GRU was set to 256. The dimension of both the acoustic and the textual context vectors was set to 256 so that they could be summed to the output of the phoneme encoder.

During training, the batch size was set to 32. We used Adam [27] as the optimizer, with a scheduled learning rate proposed in [20]. However, for the fine-tuning of the pre-trained BERT model in TCE, we set the learning rate to 10^{-7}. The combined model converged in around 200k steps.

We trained several variations of the proposed model to study how different modalities, laterals, and lengths of context influence the prosody of the synthesized speech. We denote the proposed method as ATCE-{pre., suc., bi}, where the suffix represents textual context laterals (pre. for preceding, suc. for succeeding, and bi for bilateral context). We also trained the proposed models without ACE, which are denoted by TCE-{pre., suc., bi}. We used a method without context modeling and two previous methods as the baselines. The first baseline, denoted by Baseline, is the proposed method without ACE and TCE that doesn't use any context information. The second baseline, denoted by ACE, uses ACE to utilize acoustic context [8]. The third baseline uses one-sentence bilateral textual context implicitly by feeding the target sentence with one-sentence bilateral context to the BERT model but only inputting the word embeddings of the target sentence to the TTS model [11]. Since the original method of Nakata et al. didn't use FastSpeech2 to synthesize mel-spectrograms, we adapted it to the proposed method by (1) changing the context length to one sentence and (2) only inputting the target sentence embedding to the TTS model.

4.2. Objective evaluations

4.2.1. Metrics

In the objective evaluations we used several metrics to evaluate the synthetic speech:

- **Character error rate (CER)** computed using Vosk Japanese speech recognition API[5].
- **Mel-cepstral distortion (MCD)** computed with dynamic time warping (DTW).
- **F0 root mean square error (F0-RMSE)** computed with DTW.
- **Gross Pitch Error (GPE)** represents the proportion of voiced frames whose relative pitch error is higher than a certain threshold (20% in this work).
- **Accuracy of speaker classification (ACC)** computed by a speaker classifier trained on the training set. We used ResCNN [28] as the speaker classifier, which is a powerful architecture based on residual neural networks.

Here CER and MCD measure the general speech quality, ACC measures the speaker similarity, F0-RMSE and GPE measure the performance on speech prosody.

4.2.2. Textual context length

We first evaluated model performances with different textual context length k. We trained ATCE-bi with k in $\{16, 32, 64, 128\}$. The result is shown in Table 1. First, all models have similar CER, MCD, and ACC, which is natural since the proposed method only focuses on prosody. Second, the best performance is obtained when $k = 64$, which demonstrates that increasing textual context length can improve speech prosody, but when the length is too long ($k = 128$), the performance degrades. We suppose this is because textual context with a long distance to the target sentence is not relevant for predicting the target prosody. To verify this hypothesis, we also set $k = 64$ in the $k = 128$ model for inference. As expected, the performance increases slightly in this case ($128 \rightarrow 64$) in Table 1), which implies the correctness of the hypothesis. Given that the average sentence length of the corpus is 27 characters, our results suggest that the best textual context length is about 2-3 sentences from either lateral of the target sentence.

Table 1. Results of objective evaluation for the proposed ATCE-bi model with different context length k. **Bold** indicates the best score.

k	CER(↓)	MCD(↓)	F0-RMSE(↓)	GPE(↓)	ACC(↑)
16	0.206	7.69	28.60	13.49	99.18
32	0.206	7.68	28.33	13.21	99.32
64	0.208	7.72	**28.19**	**12.90**	98.77
128	0.207	7.64	28.56	13.51	98.77
$128 \rightarrow 64$	0.206	7.65	28.43	13.23	98.91

Table 2. Results of objective evaluation for models with different types of context. **Bold** indicates the best score. * represents significant improvement over the baseline methods with p-value < 0.05.

Model	CER(↓)	MCD(↓)	F0-RMSE(↓)	GPE(↓)	ACC(↑)
GT	0.192	N/A	N/A	N/A	96.19
HiFi-GAN	0.208	4.25	14.03	2.2	98.64
Baseline	0.210	7.76	29.02	13.59	98.91
ACE	0.207	7.85	29.05	14.16	98.37
Nakata et al.	0.206	7.70	28.89	13.21	98.64
TCE-pre.	0.206	7.71	28.19	13.00	98.64
TCE-suc.	0.206	7.70	28.82	13.18	99.05
TCE-bi	0.207	7.69	28.84	12.94	98.50
ATCE-pre.	0.205	7.68	28.86	13.11	99.18
ATCE-suc.	0.209	**7.66**	**27.93***	**12.57***	98.91
ATCE-bi	0.208	7.72	28.19	12.90	98.77

[2] https://huggingface.co/cl-tohoku/bert-base-japanese-v2

[3] https://open-jtalk.sp.nitech.ac.jp/

[4] https://github.com/jik876/hifi-gan

[5] https://github.com/alphacep/vosk-api

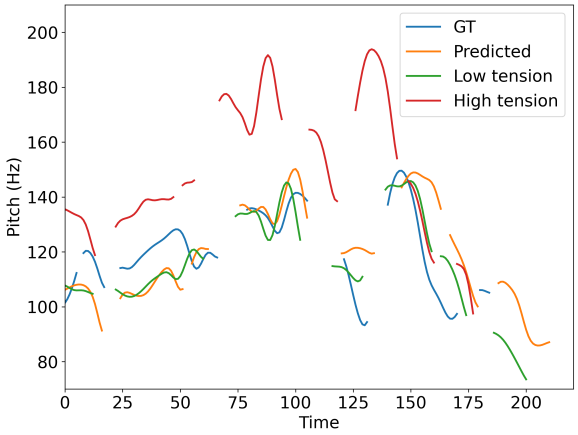

Fig. 3. Pitch contours of the same utterance synthesized using the ATCE-bi model with 2 random contexts. "Predicted" represents the one synthesized with the correct context.

4.2.3. *Context modalities and laterals*

We then evaluated model performances with different context modalities and laterals. All of the models were trained with $k = 64$ obtained in the previous section. The result is shown in Table 2. First, the proposed ATCE-* models outperform all the baseline models, which demonstrates the effectiveness of the proposed method. Second, all TCE-* models outperform ACE, which demonstrates the effectiveness of introducing textual context. Third, all ATCE-* models except ATCE-pre. have better performance than the corresponding TCE-* models, which demonstrates the necessity of combining both acoustic and textual context information for synthesis. Finally, to our surprise, ATCE-suc. using succeeding textual context obtains the best performance and has better performance than ATCE-bi using bilateral textual context. We suppose this is because the information overlapping between the $(N-1)$-th mel-spectrogram and the preceding texts makes succeeding textual context more informative for the model. Therefore in such case using preceding acoustic and bilateral textual contexts together is probably not beneficial and can even confuse the TTS model. This hypothesis can also be verified by comparing the performances of ATCE-pre. and TCE-pre..

We also notice that the accuracy of speaker classification of the GT model is the worst among all models. We believe this is because the speakers usually change their voices to act different characters in the audiobooks, which makes it difficult to recognize the speaker identity.

4.2.4. *Prosody under different contexts*

Finally, we verified whether the proposed method actually learned to predict context-dependent prosody. We selected an dialogic utterance from the test set and synthesized it using the ATCE-bi model with random contexts of other dialogic utterances. This is because we suppose prosody variations usually exist in dialogue. We observed that the model could synthesize the same text with different tensions. We selected typical examples and visualized their pitch contours in Figure 3. It can be seen that the prosody varies a lot with different contexts, which proves that the proposed method can predict context-dependent speech prosody.

4.3. Subjective evaluations

4.3.1. *Multi-sentence speech naturalness MOS test*

In the subjective evaluations, we first conducted a standard 5-scale mean opinion score (MOS) test. We fine-tuned the HiFi-GAN vocoder on the training set for 3000 epochs with an initial learning rate 10^{-5}. Following previous work [23], we conducted a five-sentence MOS test, in which the listeners rate the naturalness of an audio including five consecutive utterances. In this test we selected the two baselines and ATCE-{suc., bi.} with the best objective performance obtained in the previous section to evaluate. For each test speaker we synthesized 10 five-sentence audios, in which we inserted a 0.5 second pause after each sentence. The duration of each audio ranges from 40 seconds to 1 minute. We used Lancers[6], a Japanese crowd-sourcing platform, to conduct the test. 32 listeners joined in the test. Each listener rated 30 audios with 5 dummy samples at the beginning whose ratings were not counted in the final result. Each audio had 3 answers on average.

Table 3. Results of MOS evaluation. **Bold** indicates the best method without overlapping 95% confidence interval.

Model	MOS
ACE	3.26
Nakata et al.	3.30
ATCE-suc.	**3.38**
ATCE-bi	3.35

Table 4. Results of AB preference evaluation. **Bold** indicates the best method with p-value < 0.05.

Method A	Score A	Score B	Method B
ATCE-suc.	**0.615**	0.385	ACE
ATCE-suc.	**0.545**	0.455	Nakata et al.

The result is shown in Table 3. It can be seen that all proposed models outperform the two baseline models, and ATCE-suc. obtains the best performance, which is consistent with the results of the objective evaluations. This again demonstrates the effectiveness of the proposed method.

4.3.2. *Preference AB test*

Next we conducted a preference AB test using the same five-sentence audios synthesized in the previous section. We selected two AB pairs: (ATCE-suc., ACE), (ATCE-suc, Nakata et al.). 40 listeners participated in the test. Each listener rated 10 pairs, in which 5 pairs have the same audios but different orders from the rest 5 audios. Each pair had 3 answers on average. The result is shown in Table 4. It can be seen that the proposed ATCE-suc. method significantly outperforms the two baselines, which is consistent with the results obtained in the previous section. All in all, the proposed method utilizing informative acoustic and textual contexts obtains the best performance in all evaluations.

5. CONCLUSIONS

This paper presented a Japanese multi-speaker audiobook TTS system that fully and explicitly utilized preceding acoustic context and bilateral textual context to improve the prosody of the synthetic speech. Experimental results demonstrated that the proposed method significantly outperformed two previous work in both objective and subjective evaluations. We also found it was helpful to use multimodal contexts and the optimal textual length was about 2-3 sentences. These results can potentially shed light on future researches in this field. Future work could be extending ACE and TCE to frame-level resolution.

[6] https://www.lancers.jp/

6. REFERENCES

[1] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *Proc. ICASSP*. IEEE, 2018, pp. 4779–4783.

[2] Yi Ren, Yangjun Ruan, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech: Fast, robust and controllable text to speech," *Proc. NeurIPS*, vol. 32, 2019.

[3] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu, "Fastspeech 2: Fast and high-quality end-to-end text to speech," *arXiv preprint arXiv:2006.04558*, 2020.

[4] Paul Taylor, *Text-to-speech synthesis*, Cambridge university press, 2009.

[5] Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian MüLler, and Shrikanth Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013.

[6] Jennifer Cole, "Prosody in context: A review," *Language, Cognition and Neuroscience*, vol. 30, no. 1-2, pp. 1–31, 2015.

[7] King Simon, Crumlish Jane, Martin Amy, and Wihlborg Lovisa, "The blizzard challenge 2018," in *Proc. Blizzard Challenge workshop*, 2018.

[8] Pilar Oplustil-Gallegos and Simon King, "Using previous acoustic context to improve text-to-speech synthesis," *arXiv preprint arXiv:2012.03763*, 2020.

[9] Pilar Oplustil Gallegos, Johannah O'Mahony, and Simon King, "Comparing acoustic and textual representations of previous linguistic context for improving text-to-speech," in *The 11th ISCA Speech Synthesis Workshop (SSW11)*. ISCA, 2021, pp. 205–210.

[10] Guanghui Xu, Wei Song, Zhengchen Zhang, Chao Zhang, Xiaodong He, and Bowen Zhou, "Improving prosody modelling with cross-utterance bert embeddings for end-to-end speech synthesis," in *Proc. ICASSP*. IEEE, 2021, pp. 6079–6083.

[11] Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, Naoko Tanji, Yusuke Ijima, Ryo Masumura, and Hiroshi Saruwatari, "Audiobook speech synthesis conditioned by cross-sentence context-aware word embeddings," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 211–215.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] Junjie Pan, Lin Wu, Xiang Yin, Pengfei Wu, Chenchang Xu, and Zejun Ma, "A chapter-wise understanding system for text-to-speech in chinese novels," in *Proc. ICASSP*. IEEE, 2021, pp. 6069–6073.

[14] Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, Yuki Saito, Yusuke Ijima, Ryo Masumura, and Hiroshi Saruwatari, "Predicting VQVAE-based Character Acting Style from Quotation-Annotated Text for Audiobook Speech Synthesis," *Proc. Interspeech*, pp. 4551–4555, 2022.

[15] Ning-Qian Wu, Zhao-Ci Liu, and Zhen-Hua Ling, "Discourse-level prosody modeling with a variational autoencoder for non-autoregressive expressive speech synthesis," in *Proc. ICASSP*. IEEE, 2022, pp. 7592–7596.

[16] Daisy Stanton, Yuxuan Wang, and RJ Skerry-Ryan, "Predicting expressive speaking style from text in end-to-end speech synthesis," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 595–602.

[17] Peter Makarov, Ammar Abbas, Mateusz Łajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman, and Penny Karanasou, "Simple and effective multi-sentence tts with expressive and coherent prosody," *arXiv preprint arXiv:2206.14643*, 2022.

[18] Liumeng Xue, Frank K Soong, Shaofei Zhang, and Lei Xie, "Paratts: Learning linguistic and prosodic cross-sentence information in paragraph-based tts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2854–2864, 2022.

[19] Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, et al., "Text-free prosody-aware generative spoken language modeling," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022, pp. 8666–8681.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Proc. NeurIPS*, vol. 30, 2017.

[21] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani, "Self-attention with relative position representations," *arXiv preprint arXiv:1803.02155*, 2018.

[22] Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," in *NIPS 2014 Workshop on Deep Learning*, 2014.

[23] Shinnosuke Takamichi, Wataru Nakata, Naoko Tanji, and Hiroshi Saruwatari, "J-mac: Japanese multi-speaker audiobook corpus for speech synthesis," *arXiv preprint arXiv:2201.10896*, 2022.

[24] Akinobu Lee, Tatsuya Kawahara, and Kiyohiro Shikano, "Julius—an open source real-time large vocabulary recognition engine," 2001.

[25] Masanori Morise, Fumiya Yokomori, and Kenji Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE TRANSACTIONS on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[26] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ-Skerry Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Ye Jia, Fei Ren, and Rif A Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proc. ICML*. PMLR, 2018, pp. 5180–5189.

[27] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[28] Chao Li, Xiaokong Ma, Bing Jiang, Xiangang Li, Xuewei Zhang, Xiao Liu, Ying Cao, Ajay Kannan, and Zhenyao Zhu, "Deep speaker: an end-to-end neural speaker embedding system," *arXiv preprint arXiv:1705.02304*, May 2017.