11th ISCA Speech Synthesis Workshop (SSW 11)
26-28 August 2021, Budapest, Hungary

Audiobook Speech Synthesis Conditioned by Cross-Sentence Context-Aware Word Embeddings

Wataru Nakata[1], *Tomoki Koriyama*[2], *Shinnosuke Takamichi*[2], *Naoko Tanji*[2]
Yusuke Ijima[3], *Ryo Masumura*[3], *Hiroshi Saruwatari*[2]

[1]Faculty of Engineering, The University of Tokyo, Japan.
[2]Graduate School of Information Science and Technology, The University of Tokyo, Japan.
[3]Nippon Telegraph and Telephone Corporation, Japan.

nakata-wataru855@g.ecc.u-tokyo.ac.jp
t.koriyama@ieee.org

Abstract

This paper proposes an audiobook speech synthesis method that considers a wider range of contexts than a sentence level. The style of the audiobook speech depends not only on the current sentence to be synthesized but also on its neighboring sentences. Therefore, unlike conventional text-to-speech synthesis for isolated sentences, it is necessary to consider the context of the neighboring sentences. Our method utilizes cross-sentence context-aware word embedding, which is obtained by inputting the neighboring and current sentences into BERT. The speech synthesis model, Tacotron2, is conditioned by this word embedding in addition to the current sentence. Experimental results show that taking neighboring sentences into account significantly improves synthetic speech quality.
Index Terms: speech synthesis, cross-sentence context-aware word embedding, BERT, audiobook

1. Introduction

The quality of synthetic speech is getting closer to that of natural human speech [1]. This raises the opportunity of applying text-to-speech (TTS) to a wider range of applications. In this work, we focus on applying TTS for audiobooks. Audiobook speech synthesis is expected to reduce time and monetary requirements by replacing the recordings by professional speakers with automatic generation, and to broaden the selection of available audiobook titles. When applying TTS for audiobooks, we need to keep in mind a series of sentences, that is to be uttered fluently. Specifically, the prosody of human speech often varies on the basis of the neighboring sentences. For example, consider the following passage.

She whispered, "you won't believe it."

When humans read this passage aloud, the style of the second sentence is heavily affected by the first sentence. Considering contexts of neighboring sentences (hereinafter, "*cross-sentence context*") is one of the major challenges when it comes to achieving human-like speech in audiobook speech synthesis.

To model the cross-sentence context in speech, we consider using the techniques of natural language processing (NLP). Taking into account the cross-sentence context in a document is a common practice in NLP tasks, and deep neural networks have been proposed for this purpose. In particular, Bidirectional Encoder Representations from Transformers (BERT) [2] made breakthroughs in various downstream tasks in NLP, such as question answering, natural language inference, and document classification. A key advantage of BERT is that the model parameters of pre-trained BERT can be fine-tuned for a desired task because BERT itself is also a DNN. Moreover, the word embeddings of BERT are context-aware that is, the embedding vectors vary depending on the neighboring linguistic units. BERT can also handle multiple input sentences, which enables us to utilize cross-sentence context for modeling.

It has recently been reported that BERT is also effective for speech synthesis [3, 4, 5, 6]. Hayashi et al. [3] improved the quality of synthetic speech by using context-aware word embeddings from BERT. Fang et al. [4] tried to improve the quality of speech on a relatively small corpus and observed faster convergence during training. Kenter et al. [5] showed that the fine-tuning of BERT is pivotal to improve the quality of synthesized speech. They also demonstrated that a smaller model size of BERT works better. Recently, Jia et al. proposed PnG BERT which is an encoder model for speech synthesis models [6]. In PnG BERT, both phonemes and graphemes are used as the input.

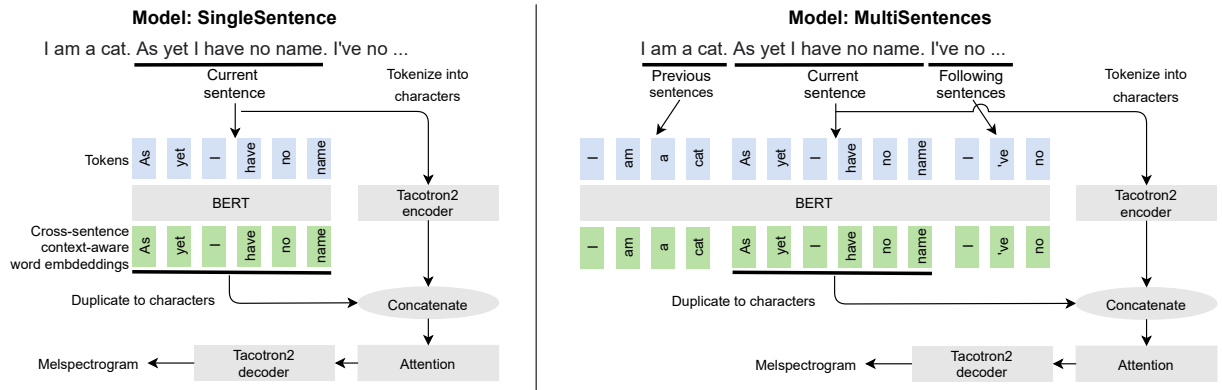
In this study, we propose an audiobook speech synthesis model that reflects the wider context by using the characteristics of BERT. Our proposed model utilizes cross-sentence context-aware word embeddings obtained by inputting multiple sentences to BERT, and Tacotron2 is conditioned by these embeddings. We performed experiments to examine the effectiveness of BERT for audiobook speech synthesis and when using the previous two sentences and current sentence as the BERT input. Subjective evaluation results showed that utilizing the cross-sentence context-aware word embeddings improved the synthetic speech quality. Synthetic speech samples are available online[1].

2. Proposed TTS synthesis model

The proposed model is based on Tacotron2 [1], a widely studied sequence-to-sequence TTS synthesis model. Our proposed model extends Tacotron2 by conditioning its encoder output with cross-sentence context-aware word

[1]https://wataru-nakata.github.io/posts/2021/05/01/ssw11

10.21437/SSW.2021-37



(a) *Model architecture of SingleSentence*

(b) *Model architecture of MultiSentences*

Figure 1: *Model architecture of proposed models. The only difference between SingleSentence and MultiSentences is that the latter also takes neighboring sentences as BERT input. Note that [CLS] and [SEP] tokens and the context-aware word embedding encoder are clipped out for simplification.*

embeddings. The proposed model only changes the structure of the encoder and not the decoder. Therefore, the structure of the decoder is identical in all compared models. We propose two models, SingleSentence and MultiSentences. SingleSentence only takes current sentence as input while MultiSentencse takes the current sentence and neighboring sentences as input.

2.1. SingleSentence

Figure 1a shows the model architecture of SingleSentence. SingleSentence takes the current sentence as input and outputs a melspectrogram. The current sentence is input to both BERT and the Tacotron2 encoder. The context-aware word embeddings from BERT then goes through a context-aware word embedding encoder. The context-aware word embedding encoder consists of two fully connected layers with ReLU activation. This is mainly used for dimensional reduction. This architecture of SingleSentence is similar to the subword-level model in [3]. However, SingleSentence concatenates the outputs of the context-aware word embedding encoder and Tacotron2 encoder, while the subword-level model has an attention mechanism for the BERT output. The word embedding is a word-level vector whereas the encoder output of Tacotron2 is a character-level one. To match the length of the vector sequences, we simply duplicated each context-aware word embedding output with their wordpiece character counts in a similar manner to [5]. With this model, we expect to synthesize speech while reflecting each word's meaning by using context-aware word embedding.

2.2. MultiSentences

Figure 1b shows the model architecture of MultiSentences. This model takes not only a text to be spoken but also neighboring sentences as input. In this study, we use the previous two sentences. This makes the model take the cross-sentence context into account. The Tacotron2 encoder only takes text to be spoken as input, while BERT takes the current sentence and its neighboring sentences as input. Except for taking multiple sentences as input, this model is identical to SingleSentence.

3. Experiments

We evaluated three models: Tacotron2, SingleSentence, and MultiSentences. For SingleSentence and MultiSentences, we evaluated on both before and after the fine-tuning of BERT.

3.1. Experimental conditions

We used the publicly available JSUT [7] and newly released J-KAC (see Appendix A for details) corpora for pretraining and fine-tuning, respectively. These are single-speaker corpora. The JSUT corpus consists of the reading-style speech of isolated sentences by a single female speaker. The J-KAC corpus consists of the very expressive continuous speech of audiobooks and kamishibai (picture stories) by a single male speaker. We downsampled the speech signals to 22.5 kHz in advance and segmented it into a sentence level. For pretraining, we split the JSUT corpus into 7496 and 100 utterances as training and development sets, respectively. For fine-tuning, we split the J-KAC corpus into 4117 (6 hours, 26 books), 100, and 97 (1 book) utterances as training, development, and test sets, respectively. The test set was open to others; no overlap existed in sentences and documents. The generated melspectrogram configurations were 80 dimensions, with the frame length of 1024 samples and frame shift of 256 samples. For input to the Tacotron2 encoder, we used katakana (i.e. Japanese pronunciation symbol) sequence.

Our training procedure involved three steps. First, we pretrained Tacotron2, SingleSentence, and MultiSentences using JSUT with frozen BERT weights. Second, we trained all models using J-KAC with frozen BERT weights. Finally, we performed fine-tuning of BERT using J-KAC for SingleSentence and MultiSentences. During the fine-tuning, the weights unfrozen except for the embedding layer, as some wordpieces did not appear in J-KAC. We conducted evaluations on the following five models.

- Tacotron2
- SingleSentence (**without** fine-tuning of BERT)
- MultiSentences (**without** fine-tuning of BERT)

- SingleSentence (**with** fine-tuning of BERT)
- MultiSentences (**with** fine-tuning of BERT)

For the pretrained BERT model, we used the one provided by akirakubo[2]. Specifically, we used the model trained using AozoraBunko (6 million sentences) and Japanese Wikipedia (3 million sentences) tokenized by SudachiPy with SudachiDict_core-20191224 for 2 million steps.

For optimization, we used an Adam [8] optimizer with $\alpha = 1 \times 10^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$ with the L2 weight regularization of 1×10^{-6}. When performing fine-tuning of BERT, we used the small L2 weight regularization of 1×10^{-9} to avoid catastrophic forgetting. Batch size was 128 distributed across four NVIDA V100 GPUs except when fine-tuning BERT. During the fine-tuning of BERT, we used the batch size of 64. For the loss function of Tacotron2, SingleSentence, and MultiSentences, we used the mean squared error of melspectrograms. We also implemented the teacher forcing on the decoder to stabilize the training.

When generating the speech, we applied a temperature softmax function on location-sensitive attention mechanism between the encoder and decoder of Tacotron2 with T = 0.5 to stabilize the speech generation in the same way as [9], which used an expressive speech dataset. In fact, without using temperature softmax, we failed to synthesize speech in most cases. As a vocoder, we used WaveRNN [10] trained on the JVS [7] corpus.

The codes of the experiments were based on NVIDIA's Tacotron2[3] implementation.

3.2. Evaluation methods

We evaluated synthesized speech with two objective metrics: Mel-Cepstral Distortion (MCD) [11] and Gross Pitch Error (GPE) [12]. When calculating these metrics, the duration of synthetic and original speech samples were aligned using FastDTW [13].

Average Mel-Cepstral Distortion (MCD)

MCD is calculated as follows.

$$\mathrm{MCD}_k = \frac{1}{T} \sum_{t=0}^{T-1} \sqrt{\sum_{k=1}^{K} (c_{t,k} - \hat{c}_{t,k})^2} \qquad (1)$$

where $c_{t,k}$ and $\hat{c}_{t,k}$ denote the k-th mel-cepstral coefficients of the t-th frames of original and synthetic speech respectively. We used $K = 13$ in the same way as [4].

Gross Pitch Error (GPE)

Gross pitch error refers to the proportion of voiced frames that deviate by more than a given ratio in pitch signal of the synthesized speech compared to the original speech. In this work, we counted pitch errors of more than 20 % as gross pitch error.

[2] https://github.com/akirakubo/bert-japanese-aozora

[3] https://github.com/NVIDIA/tacotron2

Table 1: *Objective evaluation results.*

Model	MCD[dB]	GPE
Tacotron2	4.228	0.365
SingleSentence	4.161	0.359
SingleSentence (finetuned)	4.229	0.374
MultiSentences	4.250	0.310
MultiSentences (finetuned)	4.205	0.318

3.2.1. Subjective evaluation

We evaluated the five models on three tasks: two Naturalness Mean Opinion Score (MOS) tests and 1 AB test. The naturalness MOS test included the evaluation of speech samples of both one-sentence and five-sentence lengths to examine the naturalness of a series of sentences. On each MOS test, human raters were asked to rate how natural each speech was on a 5-point scale. The number of raters was 60 and each rator evaluated 15 samples in total.

On the AB tests, raters were asked to select which of the five-sentence speech samples was more preferable for reading of a picture book. The number of raters was 40, and each rator evaluated 10 pairs.

For five-sentence speech, we generated speech for each sentence individually and concatenated them with 400 ms of silence between each sentences.

3.3. Results

Table 1 shows the results for objective evaluation using GPE and MCD. The difference of MCD was marginal in all compared models. On the other hand, the GPEs of MultiSentences and MultiSentences (fine-tuned) were significantly smaller than the other methods. This result suggests that the pitch of synthetic speech gets closer to that of natural human speech by using cross-sentence context-aware word embeddings.

Table 2 shows the subjective test results for the naturalness MOS test on one-sentence speech. In all cases, the scores were lower when we incorporated BERT into Tacotron2. One possible reason for this is that generated speech got expressive when we incorporated BERT, which resulted in a lower naturalness MOS score. Table 3 shows the naturalness MOS test results on five-sentence speech. In contrast to the results for one-sentence speech, MultiSentences (fine-tuned) outperformed the other models. This was most likely due to the speaker consistency. Specifically, Tacotron2 often made mistakes when selecting an appropriate speech style. When speech samples were concatenated to make five-sentence speech, this phenomenon became more apparent because the style of speech changes drastically among the sentences. This would result in a lower MOS. In fact, Tacotron2 had a low MOS score for the dialog speech that switches styles among sentences, but MultiSentences (fine-tuned) improved it.

Table 4 shows the results for the AB test. We can see here that MultiSentences was preferable for reading a picture book both before and after the fine-tuning of BERT. Moreover, MultiSentences was preferable to Tacotron2 when BERT was fine-tuned. Even though SingleSentence was able to utilize linguistic information from BERT, we

Table 2: *Naturallness MOS evaluation on one-sentence speech with 95% confidence intervals.*

Model	MOS
Tacotron2	3.450 ±0.141
SingleSentence	2.710 ±0.142
SingleSentence (fine-tuned)	2.676 ±0.162
MultiSentences	2.933 ±0.167
MultiSentences (fine-tuned)	2.900 ±0.164

Table 3: *Naturalness MOS evaluation on five-sentence speech with 95% confidence intervals. The model in **bold text** shows a significantly better result than Tacotron2.*

Model	MOS
Tacotron2	2.844 ±0.138
SingleSentence	2.628 ±0.144
SingleSentence (fine-tuned)	2.750 ±0.130
MultiSentences	2.767 ±0.130
MultiSentences (fine-tuned)	3.144 ±0.128

did not observe any improvement in either the naturalness MOS or AB test in our settings. These findings are different from what was reported in [3]. However, note that we used different model configurations and trained with a different language from [3].

4. Effect of modifying the previous sentences

We analyzed how the pitch of synthetic speech from MultiSentences (fine-tuned) changes by modifying previous sentences. The original input was as follows:

> ありたちが、ゾロゾロゾロゾロえさをさがしてあるいています。いちばんまえのありくんがいいました。「このあいだは、チョコレートにおせんべい、アイスクリームもおちてたね。」

which means,

> A group of ants were walking around, looking for food. The foremost ant said, *"The other day I found chocolate, rice crackers and ice cream on the ground."*

After the modifications, the previous sentences changed as follows:

> ありたちが、ゾロゾロゾロゾロえさをさがしてあるいています。いちばんまえのありくんが**大声で**いいました。「このあいだは、チョコレートにおせんべい、アイスクリームもおちてたね。」

which means,

> A group of ants were walking around, looking for food. The foremost ant said **loudly**, *"The other day I found chocolate, rice crackers and ice cream on the ground."*

We also prepared the previous sentences with antonymous modification:

> ありたちが、ゾロゾロゾロゾロえさをさがしてあるいています。いちばんまえのありくんが**小**

Table 4: *Results for AB test. Raters were asked to choose which speech was preferable for a picture book speech. Conf. shows 95% confidence intervals. **Bold text** shows results with significant difference.*

Method A	Scores	Conf.	Method B
Tacotron2	0.533 vs. 0.466	0.048	SingleSentence
Tacotron2	0.423 vs. **0.578**	0.049	MultiSentences
SingleSentence	0.400 vs. **0.600**	0.048	MultiSentences
Tacotron2	0.512 vs. 0.483	0.049	SingleSentence (fine-tuned)
Tacotron2	0.307 vs. **0.693**	0.045	MultiSentences (fine-tuned)
SingleSentence (fine-tuned)	0.302 vs. **0.698**	0.045	MultiSentences (fine-tuned)

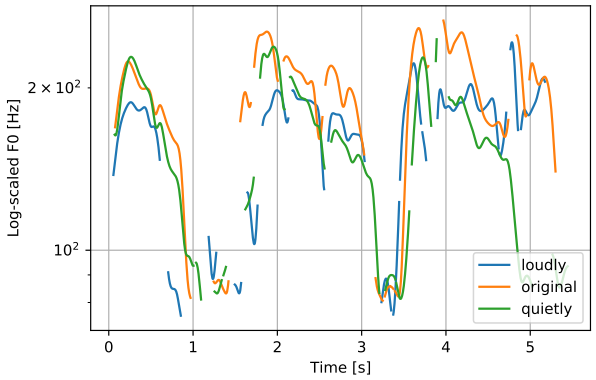


Figure 2: *Pitch change by modifying previous sentences. Note that y axis is shown in log scale. We applied trajectory smoothing [14] to the original F0 for better visualization.*

> **声で**いいました。「このあいだは、チョコレートにおせんべい、アイスクリームもおちてたね。」

which means,

> A group of ants were walking around, looking for food. The foremost ant said **quietly**, *"The other day I found chocolate, rice crackers and ice cream on the ground."*

The difference from the original sentence is shown in **bold** text, and in English translations, the current sentence is shown in *italics*. Figure 2 shows the F0 plot for before and after the modification. From this result, we can see that the generated speech was influenced by the previous sentences. We also tried to control the generated speech's emotion by modifying the previous sentences, but there was no meaningful change. The modified speech is available for listening on our speech sample page[1].

5. Conclusion

In this work, we proposed an audiobook speech synthesis model that utilizes both the current sentence and

the neighboring sentences as input to use cross-sentence context-aware word embeddings from BERT. Subjective evaluation results with generated five-sentence speech samples showed that the quality of speech improved by using the neighboring sentences. We also found that fine-tuning BERT further improved the generated speech quality.

Potential future work includes applying the proposed model for longer sentences, utilizing non-textual information such as paragraph number, and using different BERT configurations or analysis on how the neighboring sentences affect the synthetic speech.

6. References

[1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *2018 ICASSP*, 2018, pp. 4779–4783.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186.

[3] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, "Pre-trained text embeddings for enhanced text-to-speech synthesis," in *Interspeech 2019*, 2019.

[4] W. Fang, Y. Chung, and J. R. Glass, "Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models," *CoRR*, vol. abs/1906.07307, 2019.

[5] T. Kenter, M. Sharma, and R. Clark, "Improving the prosody of rnn-based english text-to-speech synthesis by incorporating a bert model," in *INTERSPEECH 2020*, 2020, pp. 4412–4416.

[6] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on Phonemes and Graphemes for Neural TTS," 2021.

[7] S. Takamichi, R. Sonobe, K. Mitsui, Y. Saito, T. Koriyama, N. Tanji, and H. Saruwatari, "JSUT and JVS: Free Japanese voice corpora for accelerating speech synthesis research," *Acoustical Science and Technology*, vol. 41, no. 5, pp. 761–768, 2020.

[8] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations*, 2014.

[9] X. Cai, D. Dai, Z. Wu, X. Li, J. Li, and H. Meng, "Emotion controllable speech synthesis using emotion-unlabeled dataset with the assistance of cross-domain speech emotion recognition," *ICASSP 2021*, 2021.

[10] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. van den Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 2410–2419.

[11] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[12] L. Rabiner, M. Cheng, A. Rosenberg, and C. McGonegal, "A comparative performance study of several pitch detection algorithms," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 5, pp. 399–418, 1976.

[13] S. Salvador and P. Chan, "Toward accurate dynamic time warping in linear time and space," *Intelligent Data Analysis*, vol. 11, no. 5, pp. 561–580, 2007.

[14] S. Takamichi, K. Kobayashi, K. Tanaka, T. Toda, and S. Nakamura, "The NAIST text-to-speech system for the Blizzard Challenge 2015," in *Proc. Blizzard Challenge workshop*, vol. 2. Berlin, Germany, 2015.

A. Japanese audiobook corpus J-KAC

We developed a very expressive corpus for Japanese audiobook speech, named J-KAC (Japanese kamishibai and audiobook corpus). This corpus includes nine hours (26 audiobooks and 17 kamishibai) of studio-quality 48-kHz sampled speech uttered by a single male professional speaker. The audio files are stored for each book, and the documents are structured in chapter, paragraph, style, and sentence levels. The "style" level has a binary label of inner sentences: "narrative style" or "character-acting style." The "sentence" level has a temporal alignment to audio. In addition to audio and documents, the corpus includes illustrations obtained by scanning products, which was done with permission from the book authors and publishers. The illustrations have various characters and background images, etc. The corpus is available for only research purposes. More information is available on our project page[4].

[4] `https://sites.google.com/site/shinnosuketakamichi/research-topics/j-kac_corpus`