Simple and Effective Multi-sentence TTS with Expressive and Coherent Prosody

Peter Makarov[*†], *Ammar Abbas*[*], *Mateusz Lajszczak, Arnaud Joly, Sri Karlapati, Alexis Moinet, Thomas Drugman, Penny Karanasou*

Alexa AI, Amazon

syeabbs@amazon.com

Abstract

Generating expressive and contextually appropriate prosody remains a challenge for modern text-to-speech (TTS) systems. This is particularly evident for long, multi-sentence inputs. In this paper, we examine simple extensions to a Transformer-based FastSpeech-like system, with the goal of improving prosody for multi-sentence TTS. We find that long context, powerful text features, and training on multi-speaker data all improve prosody. More interestingly, they result in synergies. Long context disambiguates prosody, improves coherence, and plays to the strengths of Transformers. Fine-tuning word-level features from a powerful language model, such as BERT, appears to benefit from more training data, readily available in a multi-speaker setting. We look into objective metrics on pausing and pacing and perform thorough subjective evaluations for speech naturalness. Our main system, which incorporates all the extensions, achieves consistently strong results, including statistically significant improvements in speech naturalness over all its competitors.

Index Terms: neural text-to-speech, long-form TTS, multi-speaker TTS, contextual word embeddings, FastSpeech, BERT

1. Introduction

Recent advances in neural TTS [1, 2, 3, 4] have unlocked a range of applications in which human-like and coherent prosody is crucial for customer experience. This is particularly the case for applications involving complex multi-sentence input, such as conversational agents or systems reading out news or Wikipedia articles. Most research in TTS has focused on training and evaluating models on single, isolated sentences. While the advantage of this approach is that it applies to any speech dataset, it might not work as well in domains with multi-sentence input. For example, [5] demonstrate that the perceptual evaluation score of a paragraph cannot be reliably predicted from the scores of its sentences evaluated in isolation. Prosodic coherence and contextual appropriateness are thus highly important and should be modeled accordingly.

In this paper, we focus on TTS for domains with complex multi-sentence input, which require expressive and coherent prosody, e.g. long-form reading, dialogues, question-answering prompts, etc. We investigate what improvements to speech quality could be achieved through the following simple extensions to a single-sentence baseline model:

1. **long context**: training and synthesis on utterances containing multiple adjacent coherent input sentences
2. **contextual word embeddings**: conditioning on contextual word embeddings [6], which enable context disambiguation at a more coarse-grained level than frames or phonemes,
3. **multi-speaker modeling**, which facilitates transfer learning and helps avoid overfitting due to larger combined training dataset sizes.

Naturally, these extensions can be applied simultaneously to the same model. We base our work on a **Transformer**-based FastSpeech-like TTS system [3, 4], which achieves high speech quality, efficiency, and is generally expected to perform best on longer inputs. We examine how these extensions—both individually and in various combinations—impact overall prosody and its coherence on multi-sentence input. We look into objective metrics related to pausing and pacing, and perform thorough subjective evaluations for speech naturalness.

We summarize our contributions and findings below.

1. We examine three simple extensions to a state-of-the-art TTS system to achieve more expressive, contextually appropriate, and coherent prosody on multi-sentence input: long context, contextual word embeddings, and multi-speaker modeling. To the best of our knowledge, we are the first to consider them in this setting together.
2. We see gains from all three extensions. Contextual word embeddings and long context both lead to the largest improvements in naturalness and metric scores for pausing and pacing. On the other hand, synthesis on multi-sentence input using a model not trained with long context does not work well.
3. We observe synergy effects. In particular, the baseline model extended in all three ways at once outperforms other system combinations in naturalness MUSHRA [7] evaluations by a margin. For example, it achieves a statistically significant gap reduction[1] of 59% over multi-speaker Transformer-BERT.

2. Related work

Although most research effort in TTS has focused on models operating on isolated sentences, there has been growing interest in improving synthesis for longer, multi-sentence inputs. Most innovations focus on better modeling of surrounding textual content. This is achieved with generic paragraph-based features [8, 9], hand-crafted discourse relation labels [10, 11], and—more recently—textual embeddings of neighboring sentences [12, 13]. [14] study the impact of previous acoustic context. In this paper, we explore a most direct approach to long

[*] Equal contribution.

[†] Work done while at Amazon.

[1] The gap reduction in mean MUSHRA scores between systems 1 and 2 given reference is computed as $100 * (\text{system1} - \text{system2})/(\text{reference} - \text{system2})$.

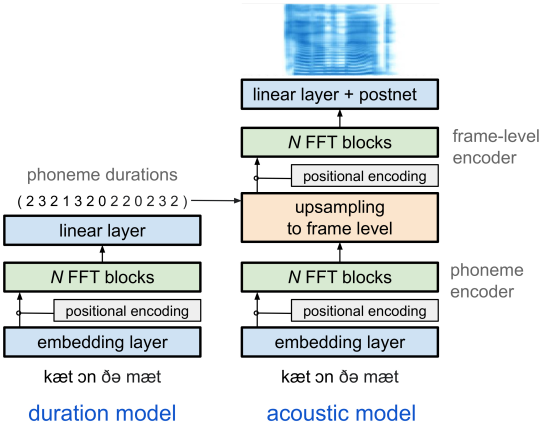

Figure 1: *FastSpeech-like Transformer baseline.*

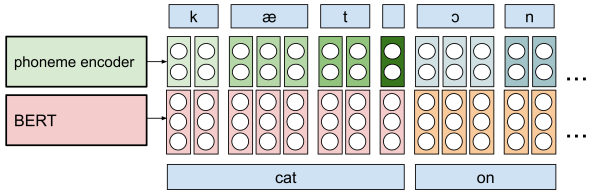

Figure 2: *Alignment and upsampling of phoneme encodings (top vectors, shares of green) and BERT word embeddings (bottom vectors, shades of red) in Transformer-BERT.*

context, which involves both text and mel-spectrogram context conditioning.

Transformers [15] are particularly well suited for modeling long-distance dependencies in the data as they do not suffer from the recency bias, unlike earlier models. This is advantageous for prosody modeling, particularly with multi-sentence inputs. [3, 4] apply non-autoregressive Transformers to TTS, and we build on their work.

There has been a lot of interest in multi-speaker TTS recently [16, 17, 18, 19, 20]. Increased scalability and improvements through transfer learning are the main advantages of such models. Unlike previous work, we examine to which extent transfer learning with multi-speaker modeling can help improve prosody and coherence of multi-sentence speech.

[21, 22, 23, 24, 25, 26] apply contextual word embeddings from large pre-trained language models in the TTS backend. Most closely related to our work is [24], who show that fine-tuned BERT word embeddings improve the perception of synthesized speech. Building on this work, we further investigate these representations in a multi-speaker setting and in the presence of longer context.

3. Methods

3.1. FastSpeech-like Transformer baseline

We take a Transformer-based FastSpeech-like system [3] as our baseline. The system is comprised of an acoustic model and a duration model (Figure 1). Both models use Feed-Forward Transformer (FFT) blocks as feature encoders. The duration model predicts phoneme durations (in frames) and is a regressor on top of the phoneme FFT encoder. The acoustic model predicts mel-spectrogram frames and consists of the phoneme FFT encoder and frame-level feature FFT encoder. The output embeddings of the phoneme FFT encoder are upsampled to the frame level using the phoneme durations predicted by the duration model and then fed into the frame-level feature FFT encoder. The acoustic and duration models use separate phoneme encoders. The architecture of the duration model and the fact that we use force-aligned phoneme durations as training targets are the two main differences from FastSpeech.

3.2. Extensions

We examine three extensions to this baseline with the goal of improving expressiveness and prosodic coherence on multi-sentence inputs.

3.2.1. Multi-speaker modeling

Multi-speaker TTS systems are commonly used as a data reduction technique, with the expectation of obtaining strong positive transfer from high-resource to low-resource speakers [16, 19]. In the context of large language models (LM) for TTS (see §3.2.2 next), the added advantage of multi-speaker modeling is that it reduces the risk of overfitting the LLM during fine-tuning simply due to the sheer amount of training data. To turn the baseline into a multi-speaker model, we concatenate pre-trained speaker embeddings to phoneme encodings. The resulting feature vectors are passed through a ReLU layer (to match the FFT block input dimension) and upsampled to the frame level. The speaker embeddings are obtained from a speaker verification model [27].

3.2.2. BERT word embeddings

Since prosody closely tracks syntax [28, 29], linguistic features (especially syntactic information) are widely believed to be beneficial for TTS. In line with previous work [24], we extend both the duration and acoustic models with contextual word embeddings from a large pre-trained Transformer LM. Concretely, we concatenate the word embedding to the encodings of the phonemes making up this word (Figure 2). As with speaker embeddings, the resulting feature vectors pass through a ReLU layer before being upsampled to the frame-level. The word embedding is computed by taking the output embedding corresponding to the first sub-token of that word. We use the cased version of BERT-base [6] and fine-tune it on the end task (phoneme duration or mel-spectrogram frames prediction) as we train the rest of the model. We refer to this baseline extension as Transformer-BERT.

3.2.3. Long context

TTS systems are generally built to work with single-sentence inputs. Some advantages of this approach are that it applies to any dataset and that synthesis is trivially parallelizable across sentences. Yet, in case data consists of coherent multiple sentences (e.g. dialogues, long-form reading, question answering prompts), it may be advantageous to take longer context into account. Long context disambiguates prosody, which is particularly evident for utterances that would be shorter otherwise. It leads to a simpler learning problem, which is beneficial for TTS models without latent variables, and narrows the space of possible realizations at synthesis time.

Here, we take a simple approach to incorporating long context for domains with multi-sentence input. We propose training and synthesizing on utterances spanning multiple adjacent

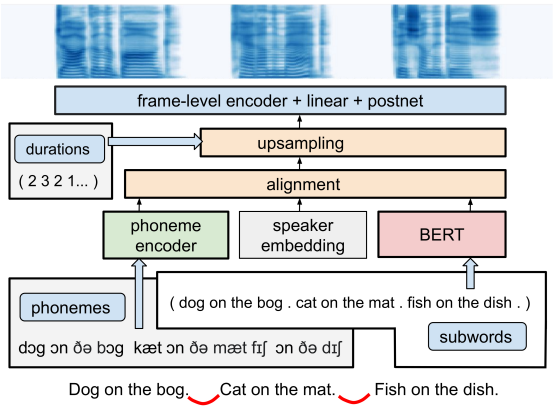

Figure 3: *Multi-speaker long-context Transformer-BERT.*

sentences. To this end, we concatenate sentences that come after each other in the original data. We highlight that this approach is viable for non-attention systems such as the baseline. Attention-based models are known to suffer from instabilities arising from attention such as mumbling, skipping, or repetitions in the audio [3, 30], which can deteriorate with input length.

We chunk the training data into consecutive chunks of concatenated sentences. A chunk is always composed of complete sentences, and there is a limit on the maximum length of the chunk. We experimented with multiple maximum lengths and found that approximately 24 seconds worked well with our data. This takes into account training efficiency and model convergence as longer utterances put a significant strain on batch size, particularly when combined with fine-tuned word embeddings. We decided against chunks of fixed length, which would lead to incomplete sentences, since this would potentially increase ambiguity in prosody. Figure 3 visualizes the incorporation of long context with the rest of the components.

During synthesis, we use a similar concatenation scheme. We highlight the importance of using a similar concatenation scheme during both training and synthesis, as this results in comparable data distributions. We examine this in more detail in §4.2.3.

4. Evaluations

We conducted experiments on internal datasets of three female English speakers referred hereon as speakers A, B, and C. By design, the data contain extensive contiguous chunks of text. We sampled the audio at 24 kHz and extracted 80-band mel-spectrograms with a frame shift of 12.5 ms. The training sets of speakers A, B, and C consist of recordings of multi-sentence chunks taken from books with total durations of about 24 h, 32 h, and 22 h, respectively.

Table 1 shows our hyperparameter and model selection choices. We train duration models for at most 2M steps and select the checkpoints with the lowest validation set mean absolute error.

Phoneme embedding / FFT input block size	256
FFT block Conv1D filter size	1024
FFT Conv1D kernel size	9
Number of attention heads	2
Number of FFT blocks in an encoder	4
FFT block dropout	0.1
Batch size (long context)	60 (45)
Batch size, duration model (long context)	48 (24)
Optimizer, on defaults	Adam
Learning rate	2e-05
Learning rate, duration model	1e-05
Acoustic model checkpoint (long context)	200k (120k)

Table 1: *Hyperparameters and model selection.*

		mean MUSHRA scores		
		this	+BERT	reference
Baseline		71.6 ± 0.7	$\mathbf{74.1 \pm 0.7}$	75.9 ± 0.7
+ multi-speaker	MT	67.5 ± 0.9	$\mathbf{72.8 \pm 0.8}$	75.3 ± 0.8

Table 2: *Ablation results for BERT word embeddings. this=the system in the leftmost column, +BERT=the system in the leftmost column extended with BERT word embeddings, reference=the reference mel-spectrograms vocoded with a parallel student vocoder. In each row, all means are statistically significantly different under a paired two-sided t-test with* $\alpha = 0.01$.

		mean MUSHRA scores		
		this	MLTB	reference
Baseline		70.0 ± 0.8	$\mathbf{73.6 \pm 0.8}$	75.1 ± 0.8
+ multi-speaker	MT	66.9 ± 0.8	$\mathbf{72.9 \pm 0.7}$	73.7 ± 0.7
+ multi-speaker, BERT	MTB	72.1 ± 0.8	$\mathbf{73.4 \pm 0.7}$	74.3 ± 0.7
CC2		69.2 ± 0.8	$\mathbf{71.0 \pm 0.8}$	72.3 ± 0.8

Table 3: *Results for multi-speaker long-context Transformer-BERT (MLTB). In each row, all means are statistically significantly different under a paired two-sided t-test with* $\alpha = 0.01$.

4.1. Subjective evaluations with MUSHRA tests

We conduct MUSHRA evaluations with 24 testers. We evaluate on two voices (A and B), with 25 test samples per voice. The samples are recordings of contiguous multi-sentence excerpts from books with an average duration of approximately 19 seconds. We ask the testers to rate the naturalness of the voices reading the samples. We vocode with a parallel student vocoder [31]. To nullify the effect from the vocoder, we also apply it to the mel-spectrograms extracted from the reference audio. We use these samples as the upper anchor in the MUSHRAs, and we refer to them as "reference" in the tables above.

BERT word embeddings. Adding BERT word embeddings leads to more natural speech (cf. strong statistical gains in mean MUSHRA scores in Table 2). While this agrees with the literature on BERT for single-speaker models, this also appears to be the case for multi-speaker systems (the bottom row of Table 2, multi-speaker Transformer (MT) vs multi-speaker Transformer-BERT (MTB)). Interestingly, the linguistic information available in BERT cannot be compensated by an almost threefold increase in training data due to multiple speakers. We note, though, that the MT samples had some harshness (not reported for the other systems), which likely impacted the scores for this system.

Ablations on multi-speaker long-context Transformer-BERT. Next, we compare various model combinations to our main system with all three extensions, multi-speaker long-context Transformer-BERT (MLTB). We observe consistent statistically significant improvements in all ablations (Table 3). They are particularly large due to the length of the test samples, which comprise about 4 sentences on average and, thus, favor

		non-pauses	pauses		
			within	between	
		MSE	MSE	MSE	R^2
Baseline		8.8	57.3	1030.1	-0.19
+ multi-speaker	MT	8.3	50.4	923.0	-0.07
+ multi-speaker, BERT	MTB	8.3	47.3	755.9	0.12
+ multi-speaker, BERT, long context	MLTB	**7.2**	**44.3**	**680.8**	**0.21**

Table 4: *Comparison of objective metrics between predicted and reference durations for various model configurations. within=intra-sentence pauses, between=inter-sentence pauses, MSE=mean squared error, R^2=coefficient of determination.*

long-context systems. We highlight the fact that the addition of long context training and synthesis leads to a 59.0% gap reduction over the MTB system, featuring BERT word embeddings. Looking at per-voice results, this gain is much larger for voice B than voice A, which could be due to the textual differences in the data.

Comparison to a strong multi-speaker system. We also evaluate MLTB against CC2 (the bottom row of Table 3), a highly competitive autoencoder-based single-utterance multi-speaker system [32]. CC2 learns a word-level variational autoencoder to capture word prosody. At synthesis time, a separate model predicts autoencoder bottleneck features from fine-tuned BERT word embeddings. We observe a strong statistically significant improvement from MLTB (a gap reduction of 58.3%).

4.2. Further analysis

4.2.1. Impact on phoneme durations and pausing

In this section, we investigate the impact of the proposed extensions on phoneme durations predicted by the duration model. We examine how well predicted durations match the reference test-set distribution, looking separately at non-pause phonemes, inter- and intra-sentence pauses. The correct placement and durations of pauses are particularly important for the overall intelligibility and coherence of speech [33]. We quantify duration error using mean squared error (MSE) and the coefficient of determination (R^2).

Generally, we see progressive improvement in durations of both pauses and non-pauses as the extensions get incorporated progressively into the baseline (Table 4). We highlight a large improvement in pause durations from a multi-speaker model over the single-speaker baseline.

We also note the large absolute improvement in pauses between sentences from the addition of word embeddings. The negative R^2 values for models without word embeddings or long context indicate they fit inter-sentence pauses worse than the test set mean. Long-context models are better at explaining variance of inter-sentence pause durations. Figure 4 shows distributions of inter-sentence pause durations for various model combinations. The models without word embeddings (the baseline and MT) produce dominant unimodal distributions. A significant improvement comes from long context as these models come closest to the reference distribution.

4.2.2. Manual error inspection

We examine test samples with large MUSHRA score gaps between any two systems. We highlight the following case where long context leads to strong gains: The sample is a sequence of

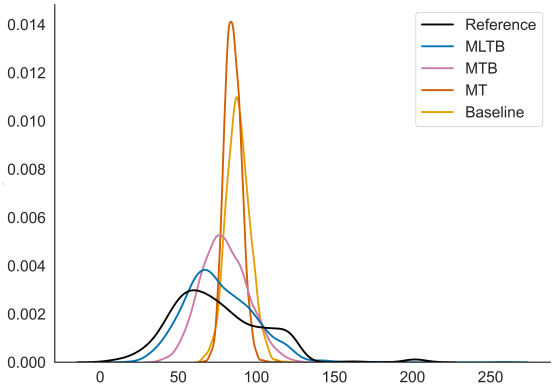

Figure 4: *Distribution of inter-sentence pauses in frames.*

many shorter sentences; or the sample contains a discontinuity which requires special prosody (e.g. a topic change).

We also look into MUSHRA score differences attributable to BERT word embeddings. BERT-powered systems have fewer problems with general vs wh-questions, and do better on phonetically ambiguous input (e.g. the sentence-final "too") and sentences with complex syntax, improving overall comprehensibility with appropriate pausing and a slower pace.

4.2.3. Effect of longer context during training and synthesis

We also examine whether a system trained on short, single-sentence input can be directly used to synthesize multi-sentence input. To this end, we conducted a preference test between MLTB and MTB with which we ran synthesis on chunks of concatenated sentences (MTB-synL). The preference test consisted of 50 utterances, 25 each from speakers A and B, of approximately 19 seconds each on average, which were rated by 120 crowdsourcing platform testers. We found that MLTB performs statistically significantly better than MTB-synL ($\alpha = 0.05$). Upon further inspection, we found that some of the samples from MTB-synL were suffering from glitches around sentence boundaries. We attribute this to the mismatch in the distributions of inputs to the MLTB-synL system at training and synthesis.

5. Conclusion

This paper looks at TTS for domains with multi-sentence input. To increase the coherence and contextual appropriateness of prosody on this type of input, we examine three simple and effective extensions to a Transformer-based FastSpeech-like baseline: long context, contextual word embeddings, and multi-speaker modeling. We show that all these extensions contribute to large improvements in objective metrics relating to pausing and phoneme durations, as well as in subjective evaluations for speech naturalness. We report strong synergies. Our best system, which combines all three extensions, leads to statistically significant improvements in speech naturalness over all its competitors.

6. Acknowledgement

We would like to thank Simon Slangen, Bajibabu Bollepalli, Arent van Korlaar, and Ray Li for their help with this work.

7. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Interspeech*, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning Wavenet on MEL spectrogram predictions," in *ICASSP*, 2018.

[3] Y. Ren, Y. Ruan, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech: Fast, robust and controllable text to speech," in *NeurIPS*, 2019.

[4] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, and T.-Y. Liu, "FastSpeech 2: Fast and high-quality end-to-end text to speech," in *ICLR*, 2021.

[5] R. Clark, H. Silen, T. Kenter, and R. Leith, "Evaluating long-form text-to-speech: Comparing the ratings of sentences and paragraphs," in *SSW10*, 2019.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional Transformers for language understanding," in *NAACL-HLT*, 2019.

[7] B. Series, "Method for the subjective assessment of intermediate quality level of audio systems," *International Telecommunication Union Radiocommunication Assembly*, 2014.

[8] K. Prahallad, A. R. Toth, and A. W. Black, "Automatic building of synthetic voices from large multi-paragraph speech databases," in *ISCA*, 2007.

[9] À. Peiró Lilja and M. Farrús, "Paragraph prosodic patterns to enhance text-to-speech naturalness," in *International Conference on Speech Prosody*, 2018.

[10] A. Aubin, A. Cervone, O. Watts, and S. King, "Improving speech synthesis with discourse relations." in *Interspeech*, 2019.

[11] N. Hu, P. Shao, Y. Zu, Z. Wang, W. Huang, and S. Wang, "Discourse prosody and its application to speech synthesis," in *ISCSLP*, 2016.

[12] G. Xu, W. Song, Z. Zhang, C. Zhang, X. He, and B. Zhou, "Improving prosody modelling with cross-utterance BERT embeddings for end-to-end speech synthesis," in *ICASSP*, 2021.

[13] X. Wang, H. Ming, L. He, and F. K. Soong, "s-transformer: Segment-transformer for robust neural speech synthesis," *arXiv preprint arXiv:2011.08480*, 2020.

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

[15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *NeurIPS*, 2017.

[16] A. Gibiansky, S. Arik, G. Diamos, J. Miller, K. Peng, W. Ping, J. Raiman, and Y. Zhou, "Deep voice 2: Multi-speaker neural text-to-speech," in *NeurIPS*, 2017.

[17] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, 2018.

[18] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding." in *Interspeech*, 2019.

[19] M. Chen, X. Tan, Y. Ren, J. Xu, H. Sun, S. Zhao, and T. Qin, "MultiSpeech: Multi-speaker text to speech with Transformer," in *Interspeech*, 2020.

[20] E. Cooper, C.-I. Lai, Y. Yasuda, F. Fang, X. Wang, N. Chen, and J. Yamagishi, "Zero-shot multi-speaker text-to-speech with state-of-the-art neural speaker embeddings," in *ICASSP*, 2020.

[21] 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.

[22] W. Fang, Y.-A. Chung, and J. Glass, "Towards transfer learning for end-to-end speech synthesis from deep pre-trained language models," *arXiv preprint arXiv:1906.07307*, 2019.

[23] N. Prateek, M. Łajszczak, R. Barra-Chicote, T. Drugman, J. Lorenzo-Trueba, T. Merritt, S. Ronanki, and T. Wood, "In other news: A bi-style text-to-speech model for synthesizing newscaster voice with limited data," in *NAACL-HLT*, 2019.

[24] 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.

[25] Y. Xiao, L. He, H. Ming, and F. K. Soong, "Improving prosody with linguistic and BERT derived features in multi-speaker based Mandarin Chinese neural TTS," in *ICASSP*, 2020.

[26] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, "PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS," in *Interspeech*, 2021.

[27] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *ICASSP*, 2018.

[28] R. Bennett and E. Elfner, "The syntax–prosody interface," *Annual Review of Linguistics*, 2019.

[29] A. Köhn, T. Baumann, and O. Dörfler, "An empirical analysis of the correlation of syntax and prosody," in *Interspeech*, 2018.

[30] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei *et al.*, "Durian: Duration informed attention network for multimodal synthesis," in *Interspeech*, 2020.

[31] Y. Jiao, A. Gabryś, G. Tinchev, B. Putrycz, D. Korzekwa, and V. Klimkov, "Universal neural vocoding with parallel WaveNet," in *ICASSP*, 2021.

[32] S. Karlapati, P. Karanasou, M. Lajszczak, A. Abbas, A. Moinet, P. Makarov, R. Li, A. van Kolaar, S. Slangen, and T. Drugman, "CopyCat2: A single model for multi-speaker TTS and many-to-many fine-grained prosody transfer," *Under review*, 2022.

[33] V. Klimkov, A. Nadolski, A. Moinet, B. Putrycz, R. Barra-Chicote, T. Merritt, and T. Drugman, "Phrase break prediction for long-form reading TTS: Exploiting text structure information," in *Interspeech*, 2017.