



# GST-BERT-TTS: Prosody Prediction Without Accentual Labels For Multi-Speaker TTS Using BERT With Global Style Tokens

Tadashi Ogura<sup>1</sup>, Takuma Okamoto<sup>1</sup>, Yamato Ohtani<sup>1</sup>, Erica Cooper<sup>1</sup>, Tomoki Toda<sup>2,1</sup>, Hisashi Kawai<sup>1</sup>

<sup>1</sup>National Institute of Information and Communications Technology, Japan

<sup>2</sup>Nagoya University, Japan

{tadashi.ogura, okamoto, yamato.ohtani, ecooper, hisashi.kawai}@nict.go.jp,  
tomoki@icts.nagoya-u.ac.jp

## Abstract

Prosody prediction is crucial for pitch-accent languages like Japanese in text-to-speech (TTS) synthesis. Traditional methods rely on accent labels, which are often incomplete and do not generalize well. BERT-based models, such as  $f_o$ -BERT, enable fundamental frequency prediction without accent labels but have been limited to single-speaker TTS. We propose GST-BERT-TTS, a novel method for multi-speaker TTS that integrates speaker-specific style embeddings from global style tokens (GST) into the token embeddings in BERT. The proposed method can realize speaker-aware fundamental frequency ( $f_o$ ) prediction in an accent label-free setting. Additionally, we extend  $f_o$ -BERT to predict not only  $\log f_o$  but also energy and duration, improving speech expressiveness. Experiments using a Japanese multi-speaker TTS corpus demonstrate that GST-BERT-TTS improves the prosody prediction accuracy and synthesis quality compared with  $f_o$ -BERT.

**Index Terms:** global style tokens, multi-speaker-text-to-speech, BERT, fundamental frequency, pitch accent language, prosody prediction

## 1. Introduction

Recent neural text-to-speech (TTS) models have achieved high-fidelity speech synthesis [1–6]. Modern TTS systems typically convert text into phoneme sequences via grapheme-to-phoneme (G2P) conversion, followed by an acoustic model and a vocoder to generate speech waveforms. To improve text analysis and prosody prediction, BERT-based models have been incorporated into TTS [7–12].

For pitch-accent languages like Japanese, prosody is crucial for intelligibility and naturalness. Conventional methods rely on accent labels obtained from dictionaries [13, 14] or hand-crafted rules [15], but these approaches struggle with out-of-vocabulary words and require costly manual annotation. These dictionary-based methods also frequently produce accent errors when accents vary with context, making exhaustive manual curation impractical. To overcome this, data-driven accent prediction has been explored [16–19], yet these methods still require large amounts of labeled data.

A recent alternative,  $f_o$ -BERT [20], predicts mora-level fundamental frequency ( $f_o$  [23]) from text input without requiring accent labels, achieving superior accent correctness compared to conventional neural TTS models. However,  $f_o$ -BERT was designed for the single-speaker TTS case and lacks speaker-aware prosody modeling. This limitation is particularly problematic in multi-speaker TTS, where prosody prediction must account for speaker-dependent characteristics. While a large-scale single-speaker dataset covering all possible

Table 1: Comparison of existing methods and proposed approach in terms of prosody accentual label-free setting (Label-free) and multi-speaker adaptability (Multi-spk).

Method	Label-free	Multi-spk
Prosody symbol control [13]		
$f_o$ -BERT [20]	✓	
Speaker embedding TTS [21]		✓
GST-TTS [22]		✓
<b>GST-BERT-TTS (Proposed)</b>	✓	✓

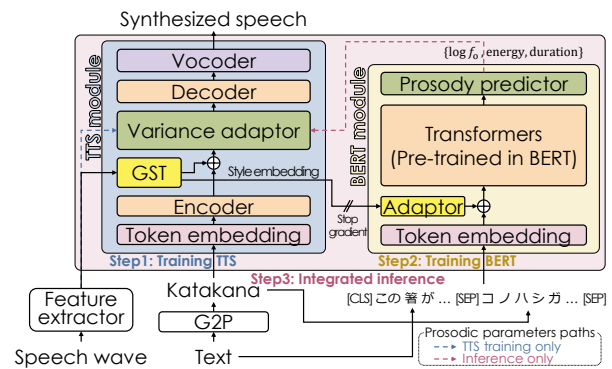


Figure 1: Proposed GST-BERT-TTS framework which integrates GST into BERT-based prosody prediction, allowing shared style embeddings across multiple speakers to enhance multi-speaker prosody modeling. Step 1: Train the TTS module. Step 2: Train the BERT module using features extracted from the TTS module. Step 3: Perform inference using both modules.

Japanese words with correct prosody could theoretically provide a solution, such a dataset is impractical to construct due to the immense data collection requirements. Multi-speaker TTS provides a more feasible solution by leveraging diverse speech data, but conventional approaches struggle to model speaker-specific prosodic variations effectively [21].

A common method for handling speaker variability is to use one-hot speaker ID embeddings [24, 25], but this approach does not generalize well to unseen speakers. Alternatively, global style tokens (GST) [22] offer a speaker-agnostic representation of speaking style, making them a promising solution for multi-speaker prosody modeling. However, our preliminary experiments revealed that if GST embeddings are not properly controlled, the synthesized speech may exhibit overly flat prosody, degrading prediction accuracy compared to models without GST. This suggests that a more structured integration of

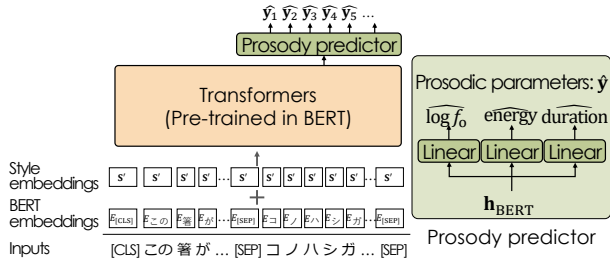


Figure 2: Overview of BERT module in GST-BERT-TTS. Style embeddings extracted from the GST module in TTS are added to BERT embeddings. Prosodic parameters ( $\log f_0$ , energy, duration) are predicted at the character level using a linear layer applied to the final BERT output.

GST into prosody prediction is necessary.

To address these issues, we propose GST-BERT-TTS, a novel method that integrates speaker-specific style embeddings from GST into the token embeddings in BERT. GST allows speaker-aware prosody prediction without explicit accent labels, extending  $f_0$ -BERT to a multi-speaker setting. Furthermore, the proposed method predicts not only  $\log f_0$  but also energy and duration, enhancing speech expressiveness. Figure 1 illustrates the proposed approach, where GST embeddings are shared across multiple speakers, allowing BERT to generate speaker-dependent prosodic features while maintaining an accentual label-free framework. Some of the speech samples used in the experiments are available on the demo page<sup>1</sup>.

Table 1 summarizes the differences between existing methods and the proposed approach. As shown, conventional methods either rely on explicit accent labels or lack multi-speaker adaptability. In contrast, GST-BERT-TTS is the only approach that achieves both label-free prosody prediction and multi-speaker compatibility.

Our contributions are summarized as follows:

- GST-BERT-TTS integrates GST-based style embeddings into BERT for speaker-aware prosody prediction.
- We extend  $f_0$ -BERT to a multi-speaker setting, enabling accentual label-free prosody prediction.
- We demonstrate that incorporating GST improves  $\log f_0$ , energy, and duration prediction accuracy, enhancing speech expressiveness.

## 2. Proposed Method

The proposed method, GST-BERT-TTS, extends the  $\log f_0$ -BERT framework by integrating speaker-specific style embeddings extracted from GSTs into the embedding layer in BERT. Figure 1 illustrates the overall framework, which consists of three main steps:

**Step 1) Training TTS module:** The TTS model is first trained using a multi-speaker speech dataset. The GST module extracts style embeddings from mel-spectrograms, which are then added to the hidden representations after the encoder. The variance adaptor predicts  $\log f_0$ , energy, and duration, which are then used to regulate the length of the hidden states before being processed by the decoder. By incorporating GST, the variance adaptor receives speaker-adapted prosodic information.

<sup>1</sup>[https://ast-astrec.nict.go.jp/demo\\_samples/gst\\_bert\\_tts\\_interspeech2025/](https://ast-astrec.nict.go.jp/demo_samples/gst_bert_tts_interspeech2025/)

**Step 2) Training BERT module:** The BERT model is trained using text and katakana input, with the corresponding prosodic parameters ( $\log f_0$ , energy, duration) as output. The training dataset is constructed using G2P conversion for text and katakana, while prosodic parameters are extracted from the alignment module of the trained TTS model, ensuring alignment with katakana tokens [20]. To enable speaker-aware prosody prediction, GST style embeddings are incorporated into the token embeddings in BERT.

**Step 3) Integrated inference:** During inference, the trained BERT model predicts prosodic parameters, which are then fed into the variance adaptor of the TTS module, allowing multi-speaker prosody generation without requiring explicit accent labels.

### 2.1. Integration of GST into BERT Embeddings

To enable speaker-aware prosody prediction, GST-based style embeddings are incorporated into the token embeddings in BERT. Since GST embeddings and BERT embeddings differ in dimension, a linear layer (denoted as “Adaptor” in Figure 1) is used to transform the GST embedding  $s$  into a compatible dimension:

$$s' = W_s s, \quad s' \in \mathbb{R}^{d_e}, \quad (1)$$

where  $W_s \in \mathbb{R}^{d_e \times d_s}$  is a trainable weight matrix. This transformation ensures that GST embeddings are properly aligned with BERT’s token embeddings, allowing BERT to utilize speaker-specific prosody information effectively.

Figure 2 illustrates the detailed architecture of the BERT module. The transformed style embedding  $s'$  is added to each token embedding  $e_{\text{BERT}}$  before input to BERT:

$$e'_{\text{BERT}} = e_{\text{BERT}} + s'. \quad (2)$$

Unlike standard token embeddings, the style embedding  $s'$  is applied uniformly across the entire sequence, ensuring that all tokens in the input share the same speaker-adapted prosodic representation. As shown in Figure 2, the transformed GST embeddings are applied before BERT processes the text, ensuring that speaker-dependent features influence all subsequent predictions.

### 2.2. Prosody Parameter Prediction

As illustrated in Fig. 2, the BERT encoder outputs a hidden representation  $h_{\text{BERT}}$  for each token. Each katakana character in the input sequence corresponds to a set of predicted prosodic parameters  $\hat{y}$ , which includes  $\log f_0$ , energy, and duration. These predictions are obtained through dedicated linear layers:

$$\hat{y} = \mathbf{W} h_{\text{BERT}} \quad (3)$$

where  $\mathbf{W} \in \mathbb{R}^{d_p \times d_e}$  is a trainable weight matrix that maps the BERT hidden representation to the prosodic parameter space.

### 2.3. Loss Function

The proposed model is trained using a loss function combining token classification and prosody parameter prediction:

$$\mathcal{L} = \alpha_t \mathcal{L}_{\text{token}} + \alpha_f \mathcal{L}_{\log f_0} + \alpha_e \mathcal{L}_{\text{energy}} + \alpha_d \mathcal{L}_{\text{duration}}. \quad (4)$$

### 2.4. Style Embedding Extraction and Application

GST embeddings are pre-trained in the TTS module. For training and inference, speaker-specific embeddings are computed as the average of all utterance-level GST vectors from a

speaker:

$$\mathbf{s}^{(i)} = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbf{s}_j^{(i)}. \quad (5)$$

During inference, the predicted prosody parameters ( $\log f_o$ , energy, duration) are applied to the variance adaptor in the TTS module, replacing the default parameters generated within the TTS model. Unlike standard TTS models, which use variance adaptors trained on internal feature predictions, our method allows the variance adaptor to incorporate externally predicted prosodic parameters from BERT.

Since BERT is pre-trained on large-scale textual data, it captures rich linguistic context, which can be leveraged for prosody prediction. By integrating GST with BERT, our approach enables speaker-aware prosody prediction while benefiting from the contextual knowledge embedded in BERT’s language model. This allows the system to generalize better across diverse linguistic inputs, particularly for multi-speaker scenarios where prosodic variations are more complex.

### 3. Experiments

#### 3.1. Dataset

We conducted the experiments using an internal Japanese multi-speaker TTS corpus<sup>2</sup>. This corpus consisted of 170 speakers, including professional and amateur voice actors who recorded speech in child, adult, and elderly styles, covering diverse speaking rates and pitch variations. The dataset contained 51,510 utterances, divided into training (48,450), development (510), and test (2,550) sets, ensuring all speakers were represented in the test set. For additional evaluation, we used the Hi-Fi-CAPTAIN corpus [26]. While large-scale multi-speaker TTS corpora existed for English, such as LibriTTS-R [27] and VCTK [28], Japanese resources remained limited. Our corpus was designed to address this gap and facilitate speaker-adaptive prosody modeling.

#### 3.2. Model Setting

The proposed model followed the  $f_o$ -BERT [20] configuration (with ConvNeXt-based acoustic model [5, 29], monotonic alignment search [30, 31] and MS-FC-HiFi-GAN neural vocoder [32]), integrating GST-based style embeddings.  $\log f_o$  was analyzed by the Harvest algorithm [33]. The BERT model had a hidden size of 768, with GST embeddings of dimension 256 added to token embeddings. All layers of BERT were fine-tuned, and the token output layer shared weights with the embedding layer. The predicted prosody parameters ( $\log f_o$ , energy, duration) replaced those generated by the variance adaptor in the TTS module. We compared GST-BERT-TTS against two baseline methods. The first baseline was a standard TTS model trained without explicit accent labels. We also include a TTS+Accent baseline, which uses the identical multi-speaker TTS architecture but is trained with explicit accent labels extracted from a Japanese accent dictionary. The second was  $f_o$ -BERT [20], a BERT-based prosody prediction model that did not incorporate speaker-aware embeddings. The model was trained using ESPnet2-TTS [34] on an Nvidia A100 cluster with 4 GPUs (40GB each) for 50 epochs, using a batch size of 128. We employed ADOPT [35] as the optimizer, with the learning rate determined empirically.

<sup>2</sup>This corpus will soon be published by the authors.

Table 2: Mean Squared Error (MSE) of prosody parameter prediction. The values represent the prediction accuracy of fundamental frequency ( $\log f_o$ ), energy, and duration. The first row shows the values predicted by the TTS system, serving as a reference for comparison.

Method	$\log f_o$	Energy	Duration
Predicted by TTS	0.330	<b>0.132</b>	17.5
$f_o$ -BERT [20]	0.776	–	–
$f_o$ -BERT (full params.)	0.773	0.314	23.4
GST-BERT ( $\log f_o$ only)	0.302	–	–
<b>GST-BERT (full params.)</b>	<b>0.266</b>	0.133	<b>14.6</b>

#### 3.3. Evaluation Criteria

We evaluated the models based on prosody prediction accuracy, accent correctness, and speech quality. For prosody prediction accuracy, we calculated the MSE between predicted and ground-truth prosodic parameters ( $\log f_o$ , energy, duration) on the test set. The baseline methods included  $f_o$ -BERT (full params.), which predicted energy and duration in addition to pitch, and GST-BERT ( $\log f_o$  only), which predicted only pitch using GST embeddings. Our proposed GST-BERT (full params.) predicted all prosodic parameters. The TTS module’s variance adaptor served as a reference. Accent correctness was evaluated using a listening test following the four-level scoring method from  $f_o$ -BERT [20]: (4) no discernible accent issues, (3) one accent appeared slightly incorrect, (2) one accent was noticeably incorrect, and (1) multiple accents were incorrect. Since our corpus focused on speaker adaptation, we used Hi-Fi-CAPTAIN for this evaluation due to its broader text diversity, ensuring transparency and reproducibility. Synthesized speech was generated using the GST embeddings of one male and one female speaker from our corpus. Speech quality was assessed using a Mean Opinion Score (MOS) test [36], in which 20 subjects rated 120 samples (20 utterances  $\times$  3 conditions  $\times$  2 speakers). For synthesis, we used one male and one female speaker’s averaged GST embeddings. Hi-Fi-CAPTAIN’s test set was chosen for its diverse linguistic content.

#### 3.4. Experimental Results

Table 2 presents the MSE results for prosody prediction, comparing  $\log f_o$ , energy, and duration across models. The TTS system’s variance adaptor serves as a reference.

The results indicate that  $f_o$ -BERT [20] improves over label-free prosody prediction but performs worse than the TTS module. Incorporating GST significantly enhances accuracy, allowing GST-BERT-TTS to outperform the TTS module in both pitch and duration prediction. Furthermore, adding energy and duration prediction to  $f_o$ -BERT (full params.) had minimal impact, whereas GST-BERT (full params.) demonstrated a synergistic effect, improving overall accuracy.

Table 3 presents the accent correctness evaluation. GST-BERT-TTS outperforms the TTS-only system and  $f_o$ -BERT, confirming its effectiveness in prosody modeling. However, TTS + Accent achieved the highest scores overall, except in the 4+3 category, where GST-BERT-TTS slightly surpassed it. This discrepancy is likely due to errors in the accent dictionary used for TTS + Accent and the mismatch between the standardized dictionary-based accents and the natural accents produced by individual speakers.

Table 3: Accent correctness evaluation. The percentages represent the cumulative proportion of samples rated at each level [%]. Higher values indicate better performance.

Method	4 [%]	4+3 [%]	4+3+2 [%]
TTS only	34.0	75.0	97.0
TTS + Accent	<b>73.5</b>	88.0	<b>100.0</b>
$f_o$ -BERT (full params.)	47.5	85.0	97.5
<b>GST-BERT-TTS</b>	65.0	<b>94.0</b>	99.5

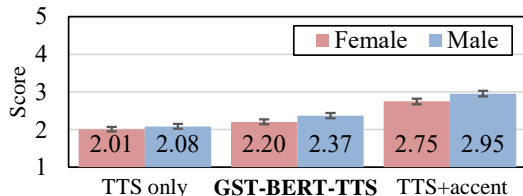


Figure 3: Mean Opinion Score (MOS) evaluation results. The error bars represent the 95% confidence intervals.

Interestingly, despite performing worse than TTS in MSE-based evaluations,  $f_o$ -BERT achieved better accent correctness. Listening tests revealed that while  $f_o$ -BERT often produced unnatural pitch contours, it accurately placed accents. Since the accent correctness evaluation focuses solely on accent placement and not pitch quality,  $f_o$ -BERT occasionally outperformed the TTS system in this regard.

Figure 3 presents the results of the MOS test. GST-BERT-TTS achieved a significantly higher MOS than TTS only, confirming its effectiveness in improving speech quality. However, its score remained noticeably lower than TTS + Accent.

The overall low MOS scores can be attributed to the characteristics of our dataset. Our corpus included professional and amateur voice actors who recorded speech in child, adult, and elderly styles, resulting in highly varied pitch dynamics and speaking rates. These variations posed challenges for the vocoder, which struggled to generalize effectively, leading to artifacts and distortions in the synthesized speech.

MOS evaluates the overall speech quality, including factors such as naturalness and audio clarity, rather than prosodic correctness. While GST-BERT-TTS demonstrated improved accent correctness compared to the baseline models, prosody accuracy alone does not necessarily translate into higher MOS scores. Listening tests revealed that TTS + Accent occasionally produced incorrect accent patterns but still achieved a higher MOS due to superior audio quality. Additionally, the relatively small number of training samples per speaker and the use of a vocoder trained on a multi-speaker corpus with high speaker variability may have contributed to the observed degradation in synthesis quality. Future improvements in model architecture, such as increasing the depth or width of the acoustic model (e.g., ConvNeXt), and enhancing vocoder training strategies could help mitigate these issues.

## 4. Discussion

### 4.1. Limitations and Future Directions

Traditional dictionary-based prosody prediction suffers from high error rates due to context-dependent accent shifts and

the impracticality of exhaustive manual labeling (Section 1; Table 2). Our data-driven GST-BERT-TTS framework is designed as a critical first step toward a scalable, accent-label-free TTS solution by leveraging abundant multi-speaker resources.

While GST enables adaptation to unseen speakers, our experiments showed that prosody prediction for unknown speakers remains challenging. When the style embedding of an unseen speaker was used, the synthesized speech often exhibited flat and monotonous prosody, suggesting that GST’s latent space does not adequately represent unseen speaker characteristics. This may stem from the limited number of speakers in our training data. Increasing speaker diversity could improve GST’s ability to generalize across styles and lead to more natural prosody predictions. A promising direction for future research is integrating our approach with large language models (LLMs). LLMs possess extensive linguistic knowledge and can leverage contextual information, potentially improving prosody prediction. However, LLMs are generally not well suited for constrained predictions, such as assigning prosodic parameters at the level of individual katakana characters. Developing methods to bridge this gap will be crucial in adapting LLMs for prosody modeling.

### 4.2. Adaptability to Speaker Diversity and Dialects

One major challenge in prosody prediction is adapting to diverse speaker characteristics. Our dataset includes both professional and amateur voice actors performing in varied vocal styles. Despite this variability, GST-BERT-TTS successfully captured speaker-dependent prosodic features, demonstrating its robustness. Since GST encodes speaker-specific style information, it may also be capable of modeling prosodic variations across dialects and speaking styles. If trained with speakers who use different dialects or unique rhythmic patterns, GST could enable prosody prediction consistent with such variations. Furthermore, our approach is not language-specific; it could potentially model regional prosodic variations in languages like English, where rhythmic and intonational differences exist across dialects.

### 4.3. Towards Automatic Prosody Label Generation

A long-term goal of this research is automatic prosody label generation. Existing accent dictionaries are incomplete, and manually annotating prosody at scale is impractical. By leveraging our method, it may be possible to extract prosody labels from large-scale speech corpora in an automated manner. This would provide a data-driven approach to prosody annotation, facilitating more comprehensive modeling without the need for handcrafted labels.

## 5. Conclusion

We proposed GST-BERT-TTS for multi-speaker TTS without accent labels, which integrates GST-based speaker embeddings into BERT, enabling speaker-aware prosody prediction while retaining  $f_o$ -BERT’s label-free approach. We further extended  $f_o$ -BERT to predict energy and duration, improving speech expressiveness. The experimental results demonstrated that GST-BERT-TTS improves the prosody prediction accuracy and synthesis quality compared with  $f_o$ -BERT. Specifically, it achieved lower MSE in prosody parameter prediction and higher accent correctness scores while maintaining competitive MOS ratings.

## 6. References

- [1] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerry-Ryan, R. A. Saurous, Y. Agiomyriannakis, and Y. Wu, “Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions,” in *Proc. ICASSP*, Apr. 2018, pp. 4779–4783.
- [2] 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 *Proc. ICLR*, May 2021.
- [3] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *Proc. ICML*, July 2021, pp. 5530–5540.
- [4] D. Lim, S. Jung, and E. Kim, “JETS: Jointly training FastSpeech2 and HiFi-GAN for end to end text to speech,” in *Proc. Interspeech*, Sept. 2022, pp. 21–25.
- [5] T. Okamoto, Y. Ohtani, T. Toda, and H. Kawai, “ConvNeXt-TTS and ConvNeXt-VC: ConvNeXt-based fast end-to-end sequence-to-sequence text-to-speech and voice conversion,” in *Proc. ICASSP*, Apr. 2024, pp. 12 456–12 460.
- [6] K. Shen, Z. Ju, X. Tan, E. Liu, Y. Leng, L. He, T. Qin, S. Zhao, and J. Bian, “NaturalSpeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers,” in *Proc. ICLR*, May 2024.
- [7] T. Hayashi, S. Watanabe, T. Toda, K. Takeda, S. Toshniwal, and K. Livescu, “Pre-trained text embeddings for enhanced text-to-speech synthesis,” in *Proc. Interspeech*, Sept. 2019, pp. 4430–4434.
- [8] 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 *Proc. ICASSP*, May 2020, pp. 6704–6708.
- [9] T. Kenter, M. K. Sharma, and R. Clark, “Improving prosody of RNN-based English text-to-speech synthesis by incorporating a BERT model,” in *Proc. Interspeech*, Oct. 2020, pp. 2958–1796.
- [10] 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 *Proc. ICASSP*, June 2021, pp. 2958–1796.
- [11] Y. Jia, H. Zen, J. Shen, Y. Zhang, and Y. Wu, “PnG BERT: Augmented BERT on phonemes and graphemes for neural TTS,” in *Proc. Interspeech*, Aug. 2021, pp. 151–155.
- [12] R. Liu, Y. Hu, H. Zuo, Z. Luo, L. Wang, and G. Gao, “Text-to-speech for low-resource agglutinative language with morphology-aware language model pre-training,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 32, pp. 1075–1087, 2024.
- [13] K. Kurihara, N. Seiyama, and T. Kumano, “Prosodic features control by symbols as input of sequence-to-sequence acoustic modeling for neural TTS,” *IEICE trans. Inf. Syst.*, vol. E104-D, no. 2, pp. 302–311, Feb. 2021.
- [14] T. Okamoto, T. Toda, Y. Shiga, and H. Kawai, “Real-time neural text-to-speech with sequence-to-sequence acoustic model and WaveGlow or single Gaussian WaveRNN vocoders,” in *Proc. Interspeech*, Sept. 2019, pp. 1308–1312.
- [15] H. Kubozono, “Japanese dialects and general linguistics,” *J. Linguist. Soc. Jpn.*, vol. 148, pp. 1–31, 2015.
- [16] H. Tachibana and Y. Katayama, “Accent estimation of Japanese words from their surfaces and romanizations for building large vocabulary accent dictionaries,” in *Proc. ICASSP*, May 2020, pp. 8059–8063.
- [17] N. Kakegawa, S. Hara, M. Abe, and Y. Ijima, “Phonetic and prosodic information estimation from texts for genuine Japanese end-to-end text-to-speech,” in *Proc. Interspeech*, Aug. 2021, pp. 3606–3610.
- [18] K. Kurihara and M. Sano, “Low-resourced phonetic and prosodic feature estimation with self-supervised-learning-based acoustic modeling,” in *Proc. ICASSPW*, Apr. 2024, pp. 640–644.
- [19] —, “Enhancing Japanese text-to-speech accuracy with a novel combination Transformer-BERT-based G2P: Integrating pronunciation dictionaries and accent sandhi,” in *Proc. Interspeech*, Sept. 2024, pp. 2790–2794.
- [20] T. Ogura, T. Okamoto, Y. Ohtani, E. Cooper, T. Toda, and H. Kawai, “Mora-level prosody prediction for text-to-speech using Japanese BERT without accentual labels,” in *Proc. ICASSP*, 2025.
- [21] S. Ueno, M. Mimura, S. Sakai, and T. Kawahara, “Multi-speaker sequence-to-sequence speech synthesis for data augmentation in acoustic-to-word speech recognition,” in *Proc. ICASSP*, May 2019, pp. 6161–6165.
- [22] Y. Wang, D. Stanton, Y. Zhang, R. Skerry-Ryan, E. Battenberg, J. Shor, Y. Xiao, F. Ren, Y. Jia, and R. A. Saurous, “Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis,” in *Proc. ICML*, July 2018, pp. 5167–5176.
- [23] I. R. Titze, R. J. Baken, K. W. Bozeman, S. Granqvist, N. Henrich, C. T. Herbst, D. M. Howard, E. J. Hunter, D. Kaelin, R. D. Kent, J. Kreiman, M. Kob, A. Löfqvist, S. McCoy, D. G. Miller, H. Noé, R. C. Scherer, J. R. Smith, B. H. Story, J. G. Švec, S. Ternström, and J. Wolfe, “Toward a consensus on symbolic notation of harmonics, resonances, and formants in vocalization,” *J. Acoust. Soc. Am.*, vol. 137, no. 5, pp. 3005–3007, May 2015.
- [24] R. Fu, J. Tao, Z. Wen, and Y. Zheng, “Phoneme dependent speaker embedding and model factorization for multi-speaker speech synthesis and adaptation,” in *Proc. ICASSP*, May 2019, pp. 6930–6934.
- [25] E. Casanova, J. Weber, C. D. Shulby, A. C. Junior, E. Gölge, and M. A. Ponti, “YourTTS: Towards zero-shot multi-speaker TTS and zero-shot voice conversion for everyone,” in *Proc. ICML*, July 2022, pp. 2709–2720.
- [26] T. Okamoto, Y. Shiga, and H. Kawai, “Hi-Fi-CAPTAIN: High-fidelity and high-capacity conversational speech synthesis corpus developed by NICT,” <https://ast-astrec.nict.go.jp/en/release/hi-fi-captain/>, 2023.
- [27] Y. Koizumi, H. Zen, S. Karita, Y. Ding, K. Yatabe, N. Morioka, M. Bacchiani, Y. Zhang, W. Han, and A. Bapna, “LibriTTS-R: A restored multi-speaker text-to-speech corpus,” in *Proc. Interspeech*, Aug. 2023, pp. 5496–5500.
- [28] J. Yamagishi, C. Veaux, and K. MacDonald, “CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92),” Nov. 2019. [Online]. Available: <https://doi.org/10.7488/ds/2645>
- [29] T. Okamoto, Y. Ohtani, and H. Kawai, “Mobile PresenTra: NICT fast neural text-to-speech system on smartphones with incremental inference of MS-FC-HiFi-GAN for low-latency synthesis,” in *Proc. Interspeech*, Sept. 2024, pp. 997–998.
- [30] J. Kim, S. Kim, J. Kong, and S. Yoon, “Glow-TTS: A generative flow for text-to-speech via monotonic alignment search,” in *Proc. NeurIPS*, Dec. 2020, pp. 8067–8077.
- [31] R. Badlani, A. Łańcucki, K. J. Shih, R. Valle, W. Ping, and B. Catanzaro, “One TTS alignment to rule them all,” in *Proc. ICASSP*, May 2022, pp. 6092–6096.
- [32] H. Yamashita, T. Okamoto, R. Takashima, Y. Ohtani, T. Takiguchi, T. Toda, and H. Kawai, “Fast neural speech waveform generative models with fully-connected layer-based upsampling,” *IEEE Access*, vol. 12, pp. 31 409–31 421, 2024.
- [33] M. Morise, “Harvest: A high-performance fundamental frequency estimator from speech signals,” in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.
- [34] T. Hayashi, R. Yamamoto, T. Yoshimura, P. Wu, J. Shi, T. Saeki, Y. Ju, Y. Yasuda, S. Takamichi, and S. Watanabe, “ESPnet2-TTS: Extending the edge of TTS research,” *arXiv:2110.07840*, 2021.
- [35] S. Taniguchi, K. Harada, G. Minegishi, Y. Oshima, S. C. Jeong, G. Nagahara, T. Iiyama, M. Suzuki, Y. Iwasawa, and Y. Matsuo, “ADOPT: Modified Adam can converge with any  $\beta_2$  with the optimal rate,” in *Proc. NeurIPS*, Dec. 2024.
- [36] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality*, 1996.