Mora-Level Prosody Prediction for Text-to-Speech Using Japanese BERT Without Accentual Labels

Tadashi Ogura[1], Takuma Okamoto[1], Yamato Ohtani[1], Erica Cooper[1], Tomoki Toda[2,1], Hisashi Kawai[1]
[1]*National Institute of Information and Communications Technology, Japan,* [2]*Nagoya University, Japan*
{tadashi.ogura, okamoto, yamato.ohtani, ecooper, hisashi.kawai}@nict.go.jp, tomoki@icts.nagoya-u.ac.jp

***Abstract*—In practical text-to-speech (TTS) for pitch accent languages, such as Japanese, high-fidelity synthesis with correct prosody requires not only a phoneme sequence but also accentual information. Although accentual information can be obtained from accent dictionaries, words not included in the dictionaries and accent sandhi are sometimes synthesized with incorrect prosody, and manual registration of huge amounts of accent data is costly. Additionally, previous machine learning-based data-driven accent information estimation approaches for TTS also require huge quantities of handcrafted accentual labels during training. This paper proposes a data-driven prosody prediction method for Japanese TTS that uses Japanese BERT and does not require any accentual labels during training. A Japanese TTS acoustic model with mora-level (katakana sequence) input is first trained and mora-level fundamental frequency values (f_o), which directly correspond to the prosody, are extracted for the training data using forced alignment. Then, a pre-trained Japanese BERT is finetuned for the mora-level f_o prediction task with word sequences including kanji and the corresponding katakana sequences as input and the mora-level f_o extracted using forced alignment as the prediction target. During TTS inference, the mora-level f_o sequence predicted by the finetuned Japanese BERT is input to the TTS acoustic model along with the katakana input, and correct prosodic synthesis can be realized thanks to this predicted f_o sequence. Experimental results demonstrate that the proposed method can realize the same synthesis quality and higher accent correctness compared with conventional neural TTS models with accentual labels.**

***Index Terms*—BERT, fundamental frequency, pitch accent language, prosody prediction, text-to-speech**

I. Introduction

Recent advancements in neural network technology have realized high-fidelity text-to-speech (TTS) synthesis [1]–[6]. Typically, input texts are converted into phoneme sequences by a text analyzer (G2P), and speech waveforms are then synthesized from the phoneme sequences. By introducing pre-trained Bidirectional Encoder Representations from Transformers [7] (BERT) [8] based on self-supervised learning using huge amounts of external text data into neural TTS, additional grapheme or word sequence input has been shown to improve the quality and prosody of synthetic speech [9]–[14].

In neural TTS for pitch accent languages, such as Japanese[1], not only the phoneme sequence but also accentual information is required for high-fidelity and prosodically-correct synthesis [5], [15], [16]. Although accentual information can be obtained from accent dictionaries, words not included in the dictionaries and accent sandhi [17] are sometimes synthesized with incorrect prosody. Moreover, manual registration of huge amounts of accent data is costly. To predict accentual information from input text for Japanese neural TTS, machine learning-based data-driven methods have been investigated [18]–[21]. However, these methods also require a huge amount of handcrafted accentual labels for training. Following the success of PnG BERT for English neural TTS [13], Japanese PnG BERT with tone prediction [22] has also been investigated. In this method, the mel-spectrogram decoder and tone predictor are jointly finetuned using a pre-trained Japanese PnG BERT with Japanese word and phoneme sequence input. However, this method does not outperform a standard neural TTS model with accentual labels in terms of synthesis quality and accent correctness because the predicted tones are not explicitly used during inference. This method also requires accentual labels for training the tone predictor.

Compared with these previous approaches [18]–[22], one ideal method would be to automatically predict accent nucleus from input Japanese text for multi-speaker Japanese neural TTS [23] using a huge amount of external text data (large language models [24]–[26]) without handcrafted accentual labels. As an initial investigation, we propose a mora-level fundamental frequency (f_o [27]) predictor, based on pre-trained Japanese BERT, for single-speaker TTS. In the proposed method, a Japanese TTS acoustic model with mora-level (katakana sequence) input is first trained, and mora-level f_o, which is a physical quantity analyzed by signal processing that directly corresponds to the prosody, is extracted for the training data using forced alignment. Then, a pre-trained Japanese BERT is finetuned on word sequences with kanji and their corresponding katakana sequences to learn to predict the mora-level f_o extracted using forced alignment. During inference, the mora-level f_o predicted by the finetuned Japanese BERT is input to the TTS acoustic model along with the katakana input. Compared with the conventional Japanese PnG BERT that requires accentual labels during training [22], the results of the experiments conducted in Sec. IV demonstrate that the proposed method can realize the same synthesis quality and higher accent correctness without requiring accentual labels. Speech samples from the experiments have been made available[2].

II. Problem statement

In Japanese, accent patterns affect speech prosody significantly, as shown in Fig. 1 where two sentences with identical phonemes have different prosodic patterns (indicated by "[" for initial rising and "]" for pitch drop) that distinguish their meanings ("chopsticks" vs "edge"). In Japanese neural TTS, input texts including kanji, hiragana, katakana, numbers, etc. are converted to phonetic representations as either (b) katakana or (c) phonemes as shown in Fig. 1 using G2P [16]. Additionally, accentual labels may be added by using accent dictionaries, and (b) or (c) are converted to (d) katakana + accent or (e) phoneme + accent in Fig. 1. Therefore, accentual labels are typically required to synthesize the speech of these two texts with correct prosody using the conventional Japanese neural TTS models [5], [15], [16]. In the proposed method, on the other hand, the speech of these two texts can be synthesized with correct prosody by

[1] (1) Many words in Japanese are typically written in kanji, a logographic script based on traditional Chinese characters, and written sentences typically contain a mix of kanji with hiragana and katakana (syllabary characters). Most kanji have several different possible readings, and the correct one including both the phonetic sequence and the pitch accent can usually be determined by the reader from context. (2) Each katakana character represents a mora, and the phonetic reading of a katakana sequence is unambiguous, but pitch accent information is not present. This paper addresses the latter problem.

[2] https://ast-astrec.nict.go.jp/demo_samples/bert_tts_icassp2025/

	Sentence 1	Sentence 2	Unit
(a) kanji + hiragana (Translated to English)	この 箸 を 持って ください (Please take these chopsticks)	この 端 を 持って ください (Please hold the edge)	Grapheme
(b) katakana	コ ノ ハ シ オ モ ッ テ ク ダ サ イ	コ ノ ハ シ オ モ ッ テ ク ダ サ イ	Mora
(c) phoneme	k o n o h a sh i o m o cl t e k u d a s a i	k o n o h a sh i o m o cl t e k u d a s a i	Phoneme
(d) katakana + accent	コ [ノ ハ] シ オ モ] ッ テ ク [ダ サ] イ	コ [ノ ハ [シ オ モ] ッ テ ク [ダ サ] イ	Mora
(e) phoneme + accent	k o [n o h a] sh i o m o] cl t e k u [d a s a] i	k o [n o h a [sh i o m o] cl t e k u [d a s a] i	Phoneme

Fig. 1. Example of Japanese texts with the same phonetic reading that differ only in prosody. “[” and “]” indicate initial rising and accent nucleus, respectively.

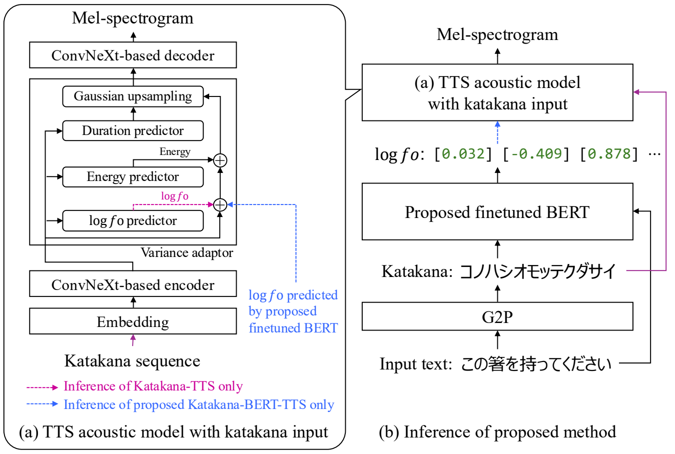
Mel-spectrogram
ConvNeXt-based decoder
Gaussian upsampling
Duration predictor
Energy
Energy predictor
log fo
log fo predictor
Variance adaptor
ConvNeXt-based encoder
Embedding
Katakana sequence
log fo predicted by proposed finetuned BERT
Inference of Katakana-TTS only
Inference of proposed Katakana-BERT-TTS only
(a) TTS acoustic model with katakana input
Mel-spectrogram
(a) TTS acoustic model with katakana input
log fo: [0.032] [-0.409] [0.878] ⋯
Proposed finetuned BERT
Katakana: コノハシオモッテクダサイ
G2P
Input text: この箸を持ってください
(b) Inference of proposed method

Fig. 2. (a) Network architecture of non-autoregressive Japanese neural TTS model with katakana input. During training, katakana token durations are obtained by monotonic alignment search [28], and mora-level $\log f_\text{o}$ is obtained through forced alignment. During inference, the model uses mora-level $\log f_\text{o}$ predicted by finetuned Japanese BERT. (b) Inference procedure of the proposed method.

using (a) kanji + hiragana (word sequence) and (b) katakana sequence input because these two kanji characters, “chopsticks” and “edge,” are included in the training data for TTS, as shown in the speech samples[2].

III. Proposed Method

A. Training Japanese TTS acoustic model with katakana sequence input and extracting mora-level $\log f_\text{o}$ of training data

In the proposed method, a Japanese neural TTS acoustic model with katakana sequence input (instead of phoneme sequences) without accentual labels is first trained because accent information for Japanese is defined in mora units that correspond well to katakana units [16]. The acoustic model is based on a non-autoregressive encoder-decoder model as in [29]–[31] (Fig. 2(a)). During training, we introduce an alignment training framework [32] that uses monotonic alignment search (MAS) [28]. Therefore, the loss function for training the acoustic model is the same used in [31]. After training, mora-level $\log f_\text{o}$ of the training data can be extracted using forced alignment, which is then used to finetune a pre-trained Japanese BERT for the mora-level $\log f_\text{o}$ prediction task as described in the next subsection. The mora-level $\log f_\text{o}$ is a physical quantity analyzed by signal processing which directly corresponds to the prosody. However, this acoustic model by itself cannot correctly synthesize the speech of the two texts in Fig. 1 from their katakana sequences because no accentual labels are used.

B. Finetuning of pre-trained Japanese BERT for predicting mora-level $\log f_\text{o}$ considering word sequence input

A Japanese BERT is pre-trained using a huge amount of Japanese text from an external corpus with the same standard multitask learning of next sentence prediction and masked language modeling as vanilla English BERT [8]. For finetuning, we augment the pre-trained Japanese BERT with an additional task-specific module for mora-level $\log f_\text{o}$ prediction. This module, which is trained together with BERT during finetuning, consists of a single fully connected layer with an output dimension of 1 for each time step. The complete model is then finetuned for multitask learning of token prediction and $\log f_\text{o}$ prediction (Fig. 3). The input to this model is a word sequence including kanji and its corresponding katakana sequence. The token prediction loss $\mathcal{L}_t$ is defined as the cross-entropy loss: $\mathcal{L}_t = -\sum_{t=1}^{T}\sum_{v=1}^{V}\tau_{(t,v)}\log(\hat{y}_{(t,v)})$, where T is the number of tokens in the sequence, V is the size of the vocabulary, $\tau_{(t,v)}$ is the ground truth token label (0 or 1), and $\hat{y}_{(t,v)}$ is the predicted probability for token v at position t. The $\log f_\text{o}$ prediction loss $\mathcal{L}_f$ is calculated using the mean squared error: $\mathcal{L}_f = \frac{1}{T}\sum_{t=1}^{T}(\phi_t - \hat{\phi}_t)^2$, where ϕ_t represents the ground truth $\log f_\text{o}$ obtained using forced alignment and $\hat{\phi}_t$ is the predicted $\log f_\text{o}$ for the token at position t. The total loss $\mathcal{L}_\text{total}$ for finetuning is a weighted sum of $\mathcal{L}_t$ and $\mathcal{L}_f$:

$$\mathcal{L}_\text{total} = (1-\alpha)\mathcal{L}_t + \alpha\mathcal{L}_f, \quad 0 \leq \alpha \leq 1, \tag{1}$$

where α is a weighting factor. When $\alpha = 0$, the model only conducts token prediction, while $\alpha = 1$ solely focuses on $\log f_\text{o}$ prediction.

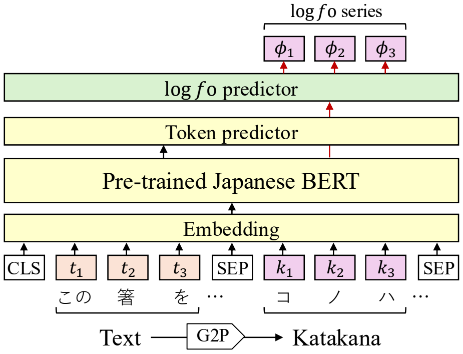

Fig. 3. Proposed Japanese BERT-based mora-level $\log f_\text{o}$ prediction model with word sequences including kanji and their corresponding katakana sequence input. CLS and SEP are special tokens for start and end of a sentence used in BERT.

C. Inference from word sequences including kanji and corresponding katakana sequences

During inference of the proposed method, an input Japanese text is converted into a word sequence using a word segmenter, and into a katakana sequence using a G2P. Then, the mora-level $\log f_\text{o}$ corresponding to each katakana token is predicted from the word sequence and katakana sequence input by the finetuned BERT. The katakana sequence and predicted mora-level $\log f_\text{o}$ are then input to the Japanese TTS acoustic model (Fig. 2(b)), and the output speech waveform can be obtained using a neural vocoder. Although no accentual labels are introduced throughout the training and inference of the proposed method, correct prosodic synthesis can be realized.

TABLE I
ACOUSTIC AND INTELLIGIBILITY METRICS FOR TEXT-TO-SPEECH METHODS: MCD, $\log f_\text{o}$ RMSE, AND CER FOR FEMALE AND MALE VOICES

Method name	**MCD [dB]**		$\log f_\text{o}$ **RMSE**		**CER [%]**	
	Female	***Male***	***Female***	***Male***	***Female***	***Male***
Original	-		-		0.4	1.2
Phoneme + accent	5.37 ± 0.52	**4.61 ± 0.68**	0.21 ± 0.06	0.20 ± 0.07	0.5	1.1
Katakana + accent	5.47 ± 0.66	4.68 ± 0.54	0.22 ± 0.06	**0.19 ± 0.05**	**0.4**	**1.0**
Katakana	5.58 ± 0.66	4.70 ± 0.64	0.20 ± 0.06	**0.19 ± 0.05**	1.9	2.2
Katakana-BERT (Proposed)	**5.30 ± 0.59**	4.69 ± 0.66	**0.19 ± 0.06**	**0.19 ± 0.05**	1.1	1.9

Note: Acoustic metrics - MCD: Mel-Cepstral Distortion, $\log f_\text{o}$ RMSE: Root Mean Square Error of $\log f_\text{o}$.
Intelligibility metric - CER: Character Error Rate from ASR. Lower values indicate better performance.
Values for MCD and $\log f_\text{o}$ RMSE are presented as mean ± standard deviation.

TABLE II
COMPARISON OF ACCENT EVALUATION SCORES FOR DIFFERENT METHODS ACROSS IN-DOMAIN AND OUT-OF-DOMAIN DATASETS

Method	**Score**	
	Hi-Fi-CAPTAIN (in-domain)	***JVS-Parallel (out-of-domain)***
Phoneme + accent	3.67 ± 0.14	3.16 ± 0.16
Katakana + accent	3.62 ± 0.15	3.17 ± 0.16
Katakana	3.17 ± 0.20	2.16 ± 0.19
Katakana-BERT (Proposed)	**3.80 ± 0.10**	**3.25 ± 0.15**

Note: Scores are presented as mean ± 95% confidence interval.

IV. EXPERIMENTS

To evaluate the proposed method, experiments were conducted with a sampling frequency of 24 kHz. All the neural network models were implemented by modifying ESPnet2-TTS [33] on PyTorch [34] and trained using an NVIDIA Tesla A100 GPU with 40 GB of memory.

A. Experimental conditions

Dataset: The experiments were conducted using the Japanese speech dataset (one female and one male speaker) of the Hi-Fi-CAPTAIN corpus released by NICT [35], with over 22 hours per speaker (one female, one male). The training set consisted of 18,655 parallel utterances each for the female and male speakers, with an additional 201 and 203 non-parallel utterances, respectively. 100 utterances were used for the validation set and the same number was used for the test set, as specified in [35]. 80-dimensional mel-spectrograms bandlimited to 7600 Hz were used. The STFT length and shift length were 1024 and 256 samples, respectively. For the pre-trained Japanese BERT, we used the open-sourced NICT Japanese BERT model with a vocabulary size of 100,000 tokens, trained on Japanese Wikipedia articles[3]. To ensure complete coverage of all single-character katakana tokens, which is crucial for the mora-level processing, we added four katakana characters that were not included in the original vocabulary. These additional katakana characters were inserted by replacing unused tokens in the pre-trained vocabulary. This modification was necessary to accommodate the unique approach of single-character katakana tokenization. Notably, there were no out-of-vocabulary words in the Hi-Fi-CAPTAIN corpus text with respect to the pre-trained BERT vocabulary.

For finetuning of the pre-trained Japanese BERT, we created a dataset using the trained TTS acoustic model with katakana input. We performed forced alignment on the training data to obtain mora-level alignments and extracted $\log f_\text{o}$ values for each mora. The resulting dataset comprises input text (with kanji), katakana representation, and corresponding mora-level $\log f_\text{o}$ values and was used to finetune the pre-trained Japanese BERT model.

Model setting: All models were based on a modified version of the Fastspeech 2-based acoustic model implemented in ESPnet2-TTS [33]. Instead of a Transformer-based encoder and decoder, ConvNeXt-based ones with the same parameters as used in [5] were introduced. Additionally, MAS [28] was introduced as in [31]. The Harvest algorithm [36] was used for f_o analysis. The following four Japanese neural TTS acoustic models were investigated:

(1) Phoneme + accent: As a baseline model, the neural TTS model with phoneme and accentual sequence input (Fig. 1(e)) was trained. The G2P function was based on pyopenjtalk (OpenJtalk [37]) and enhanced with prosody symbols [16] as used in [33].

(2) Katakana + accent: As another baseline model, the neural TTS model with katakana and accentual sequence input (Fig. 1(d)) was also trained with OpenJtalk-based G2P.

(3) Katakana: To investigate the effectiveness of the proposed method, the neural TTS model with katakana sequence input (Fig. 1(b)) was trained with OpenJtalk-based G2P.

(4) Katakana-BERT (proposed): The proposed model with word sequences and their corresponding katakana sequence input (Fig. 1(a)+(b)) was trained.

Therefore, the only difference between (3) Katakana-TTS and (4) Katakana-BERT-TTS is the mora-level $\log f_\text{o}$ used during inference (Fig. 1(a)). MS-FC-HiFi-GAN [38], [39] was used as the neural vocoder. The acoustic models and neural vocoder for each speaker were separately trained and jointly finetuned as in [33].

For the BERT component of the proposed model, we finetuned pre-trained Japanese BERT with 12 layers and 768-dimensional hidden states, as illustrated in Figure 3. The finetuning process used the AdamW optimizer [40] with an initial learning rate of 1×10^{-4}, incorporating a warm-up schedule that gradually decreased the learning rate. We trained the model for 500 epochs, selecting the best model based on the lowest training loss. The tokenization process followed the NICT-BERT approach, ensuring consistency with the pre-trained model. In the experiments, α in Eq. (1) was set to 0.5. Preliminary experiments with $\alpha = 1$ (discarding token prediction) resulted in lower f_o prediction accuracy, indicating that joint optimization of token and $\log f_\text{o}$ prediction enhances the model's performance by capturing contextual information more effectively.

Evaluation criteria: Mel-cepstral distortion (MCD), $\log f_\text{o}$ root-mean-square error (RMSE), and character error rate (CER) of automatic speech recognition (ASR) were used as objective evaluation criteria, following [4], [5], [33]. The MCD and $\log f_\text{o}$ RMSE were calculated by the ESPnet2-TTS toolkit [4], [33]. The CER was calculated by the Transformer-based Japanese ASR model trained using the CSJ corpus [41] used in [5], [33]. To evaluate the synthesized speech subjectively, mean opinion score (MOS) tests [42] were conducted. Each subject evaluated 400 samples: 40 utterances × 5 conditions × 2 speakers (female and male). The naturalness of each sample was rated on a five-point scale: (1) bad, (2) poor, (3) fair, (4) good, and (5) excellent. Twenty adult Japanese native speakers without hearing loss, who can judge correct Japanese accent, participated using

[3] https://alaginrc.nict.go.jp/nict-bert/NICT_BERT-base_JapaneseWikipedia_100K.zip

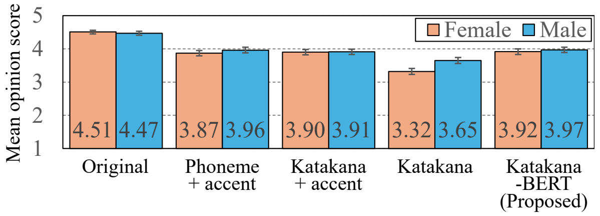

Fig. 4. MOS comparison of text-to-speech methods for female and male voices. Error bars indicate 95% confidence intervals.

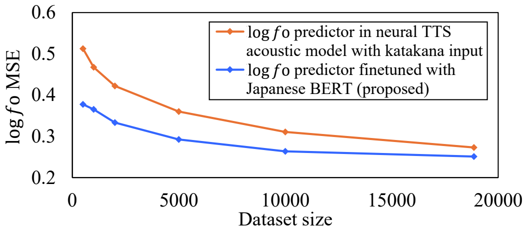

Fig. 5. Relationship between dataset size and $\log f_o$ prediction accuracy: Comparison of proposed BERT-based $\log f_o$ predictor (Fig. 3) and $\log f_o$ predictor in Japanese neural TTS acoustic model with katakana input (Fig. 2(a)).

headphones. To assess the accuracy of the synthesized accents, a Japanese native speaker from the standard dialect region (Tokyo), who can also judge correct Japanese accent, conducted expert annotation of the synthesized samples using a four-point scale: (4) no discernible accent issues, (3) one accent appears slightly incorrect, (2) one accent is noticeably incorrect, and (1) multiple accents are incorrect. This evaluation used both in-domain (Hi-Fi-CAPTAIN) and out-of-domain (JVS Parallel [43]) datasets. The JVS Parallel test set contains 100 samples without ground truth audio. There were also no out-of-vocabulary words in the JVS parallel text with respect to the pre-trained BERT vocabulary. Additionally, to assess the impact of dataset size, we evaluated $\log f_o$ prediction accuracy for both the Katakana-TTS and proposed BERT-based models trained on various subset sizes (0.5k, 1k, 2k, 5k, 10k, and all utterances) of the training data.

B. Experimental results

Table I presents acoustic and intelligibility metrics for various TTS methods including the proposed Katakana-BERT approach. The proposed method demonstrates competitive or superior performance in acoustic metrics in this experiment, particularly for the female voice used in this study. Notably, the proposed approach, which does not use explicit accentual labels, achieves performance comparable to or surpassing methods that utilize such information, underscoring the effectiveness of the proposed method in implicitly learning and reproducing proper Japanese accentuation through the BERT-based and mora-level f_o prediction.

The MOS results in Figure 4, evaluated by Japanese native speakers, show statistically significant improvement of the proposed Katakana-BERT method over the basic Katakana-TTS model. The proposed method performs comparably to approaches using explicit accentual information, further validating its effectiveness in reproducing natural Japanese prosody without explicit accentual labels.

The accent scores in Table II show that the proposed method outperforms all others, including those using explicit accent information. This is because some labels from OpenJtalk are incorrect. While all methods show worse performance in the out-of-domain setting, likely due to JVS's longer sentences and numerous foreign proper nouns, the proposed approach exhibits more robust performance across domains, suggesting effective implicit learning of accentuation patterns.

Fig. 5 illustrates the relationship between dataset size and $\log f_o$ prediction accuracy. The proposed method consistently outperforms the conventional model across all dataset sizes, maintaining higher accuracy even with smaller datasets. This suggests effective leveraging of pre-trained BERT knowledge. The performance curves indicate potential for further improvement with increased data, highlighting scalability for future enhancements in prosody prediction.

Consequently, the effectiveness of the proposed method is validated in the single-speaker TTS case, demonstrating competitive or superior performance in acoustic metrics and accent evaluation compared to the conventional approaches using explicit accentual labels.

C. Discussion

The main limitation of our approach is that words missing from the BERT vocabulary and TTS training data cannot be synthesized with correct prosody; however, this limitation highlights potential areas for improvement. Future work includes investigating multi-speaker TTS and integrating large language models to expand the vocabulary and voice types that can be synthesized with accurate prosody.

While we explored other pre-trained language models such as BART [44], BERT proved to be particularly well-suited for our mora-level f_o prediction task. The key advantage of BERT lies in its architecture, which maintains the same sequence length for input and output, aligning perfectly with our requirement of predicting f_o values for each mora in the input sequence. This one-to-one correspondence between input and output elements facilitates more straightforward and accurate predictions, making BERT an ideal choice for mora-level prosody prediction in Japanese TTS.

Despite its focus on Japanese, this approach has significant potential for broader applications. The method of predicting mora-level $\log f_o$ without explicit accentual labels could be readily adapted to phoneme-level prediction for other languages. This adaptability suggests that our framework for prosody prediction is not inherently limited to Japanese, but can be tailored to various linguistic contexts.

Furthermore, the methodology could be extended beyond f_o to predict other crucial speech parameters such as energy and duration. This indicates potential applicability to a wide range of prosodic features across different language families, not limited to pitch-accent languages like Japanese. By predicting these additional parameters, the method could potentially enhance the naturalness and expressiveness of synthesized speech in diverse languages and TTS applications.

V. Conclusion

This paper proposes a data-driven prosody prediction method for Japanese TTS using Japanese BERT without the use of accentual labels during training. A Japanese TTS acoustic model with katakana sequence input is first trained and mora-level $\log f_o$, which directly corresponds to the prosody, is extracted using forced alignment. Then, a pre-trained Japanese BERT is finetuned on word sequences including kanji and the corresponding katakana sequence input for the task of predicting the mora-level $\log f_o$ extracted using forced alignment. During inference, the mora-level $\log f_o$ predicted by the finetuned Japanese BERT is used in the TTS acoustic model with katakana input, and correct prosodic synthesis can be realized thanks to the word sequence with kanji input. Experimental results demonstrated that the proposed method can realize the same synthesis quality and higher accent correctness compared with conventional neural TTS models that require accentual labels.

Acknowledgment

We would like to thank Professor Katsuhito Sudoh of Nara Women's University, Japan, for his helpful comments on this work.

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. Agiomyrgiannakis, 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] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. NIPS*, Dec. 2017, pp. 5998–6008.

[8] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, June 2019, pp. 4171–4186.

[9] 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.

[10] 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.

[11] 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.

[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 *Proc. ICASSP*, June 2021, pp. 2958–1796.

[13] 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.

[14] 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.

[15] Y. Yasuda, X. Wang, S. Takaki, and J. Yamagishi, "Investigation of enhanced Tacotron text-to-speech synthesis systems with self-attention for pitch accent language," in *Proc. ICASSP*, May 2019, pp. 6905–6909.

[16] 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.

[17] H. Kubozono, "Japanese dialects and general linguistics," *J. Linguist. Soc. Jpn.*, vol. 148, pp. 1–31, 2015.

[18] 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.

[19] 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.

[20] 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.

[21] ——, "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.

[22] Y. Yasuda and T. Toda, "Investigation of Japanese PnG BERT language model in text-to-speech synthesis for pitch accent language," *IEEE J. Sel. Top. Signal Process.*, vol. 16, no. 6, pp. 1319–1328, Oct. 2022.

[23] 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.

[24] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," in *Proc. ACL*, July 2023, pp. 1049–1065.

[25] S. Qiao, Y. Ou, N. Zhang, X. Chen, Y. Yao, S. Deng, C. Tan, F. Huang, and H. Chen, "Reasoning with language model prompting: A survey," in *Proc. ACL*, July 2023, p. 5368–5393.

[26] Y. Yao, P. Wang, B. Tian, S. Cheng, Z. Li, S. Deng, H. Chen, and N. Zhang, "Editing large language models: Problems, methods, and opportunities," in *Proc. ENMLP*, Dec. 2023, p. 10222–10240.

[27] 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.

[28] 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.

[29] A. Łańcucki, "FastPitch: Parallel text-to-speech with pitch prediction," in *Proc. ICASSP*, June 2021, pp. 6573–6577.

[30] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, and Y. Zhang, "Recent developments on ESPnet toolkit boosted by Conformer," in *Proc. ICASSP*, June 2021, pp. 5874–5878.

[31] T. Okamoto, Y. Ohtani, S. Shimizu, T. Toda, and H. Kawai, "Challenge of singing voice synthesis using only text-to-speech corpus with FIRNet source-filter neural vocoder," in *Proc. Interspeech*, Sept. 2024, pp. 1870–1874.

[32] 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.

[33] 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.

[34] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. NeurIPS*, Dec. 2019, pp. 8024–8035.

[35] 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.

[36] M. Morise, "Harvest: A high-performance fundamental frequency estimator from speech signals," in *Proc. Interspeech*, Aug. 2017, pp. 2321–2325.

[37] A. Lee, K. Oura, and K. Tokuda, "MMDAgent—A fully open-source toolkit for voice tnteraction systems," in *Proc. ICASSP*, May 2013, pp. 8382–8385.

[38] 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.

[39] 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.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. ICLR*, May 2019.

[41] K. Maekawa, "Corpus of spontaneous Japanese: Its design and evaluation," in *Proc. SSPR*, Apr. 2003, pp. 7–12.

[42] ITU-T Recommendation P. 800, *Methods for subjective determination of transmission quality*, 1996.

[43] 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," *Acoust. Sci. Tech.*, vol. 41, no. 5, pp. 761–768, Sept. 2020.

[44] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. ACL*, July 2020, pp. 7871–7880.