*INTERSPEECH 2021*
30 August – 3 September, 2021, Brno, Czechia

# Improve Cross-Lingual Text-To-Speech Synthesis on Monolingual Corpora with Pitch Contour Information

*Haoyue Zhan, Haitong Zhang, Wenjie Ou, Yue Lin*

NetEase Games AI Lab, China
{zhanhaoyue, zhanghaitong01, ouwenjie, gzlinyue}@corp.netease.com

## Abstract

Cross-lingual text-to-speech (TTS) synthesis on monolingual corpora is still a challenging task, especially when many kinds of languages are involved. In this paper, we improve the cross-lingual TTS model on monolingual corpora with pitch contour information. We propose a method to obtain pitch contour sequences for different languages without manual annotation, and extend the Tacotron-based TTS model with the proposed Pitch Contour Extraction (PCE) module. Our experimental results show that the proposed approach can effectively improve the naturalness and consistency of synthesized mixed-lingual utterances.

**Index Terms**: TTS, cross-lingual, monolingual corpora

## 1. Introduction

In recent years, researchers on end-to-end text-to-speech have gained success in generating natural speech [1, 2]. Most recently, there are some studies on controlling extended characteristics such as speaker identity, rhythm, expressiveness, and emotion [3–5]. Moreover, some researchers have attempted to extend the TTS model from a single language to multiple languages.

[6] proposes to synthesize speech in multilingual languages by using a multilingual corpus from the same speaker. However, the construction of the multilingual corpus is expensive and time-consuming, since it requires the voice talent to be proficient in multiple languages. Therefore, researchers have begun to build multilingual TTS models on monolingual corpora. [7] introduces a transfer learning scheme that converts source language symbols to target language symbols through Automatic Speech Recognition (ASR). [8] also generates code-switched speech with an ASR-based approach. [9] presents a mixed-lingual TTS system with only monolingual corpora by utilizing speaker embedding and phoneme embedding. Similarly, [10] adopts language embedding and separate encoders to synthesize code-switched speech. At the same time, there have been some works on exploring the representation of text. [11] uses International Phonetic Alphabet (IPA) to unify the input representation of different languages, and implements a speaker encoder network based on residual convolution to extract the characteristics of the target speaker. This model generates decent-quality speech for both seen and unseen speakers. [12] proposes a multilingual TTS system using a bytes representation for English, Mandarin, and Spanish text. [13] adds a residual module to improve the stability of the TTS model and uses adversarial training to disentangle the input representation from the speaker identity. The model can synthesize multilingual speech with native or foreign accents. The above works mainly focus on the text representation and speaker identity, while more and more speech synthesis models begin to introduce more various information to TTS models, such as duration, fundamental frequency, and other information [14–16]. These researches have demonstrated the effectiveness of introducing external features in speech synthesis.

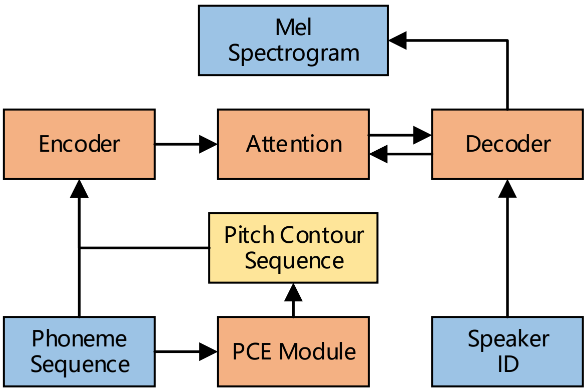

Figure 1: *Overview of the components of the proposed model.*

Although some progress has been made in cross-lingual TTS on monolingual corpora, most of the works concern tone languages and stress-accent languages [9, 11–13, 17], and exclude pitch-accent languages where the prominence of syllables is achieved by pitch [18]. Accentual-types (accent nucleus positions) in some pitch-accent languages, such as Japanese, may change the meanings of words but they are not shown in characters. Previous researchers use an external text analyzer including rules of pitch accent types and hand-written dictionaries to improve the performance of TTS model [19]. However, how to integrate pitch-accent languages into a cross-lingual speech synthesis system remains a challenge [20].

In this paper, we investigate the multi-speaker cross-lingual speech synthesis system for three different kinds of languages based on monolingual corpora. Without loss of generality, Mandarin, English, and Japanese are used as our target languages in the cross-lingual TTS system. As we all know, Mandarin is a tone language, English is a stress-accent language, and Japanese is a typical pitch-accent language. We build upon the recent works of TTS model to design a new cross-lingual synthesis framework. In this framework, we propose a method to obtain binary pitch contour sequences for different languages without manual annotation, and extend Tacotron-based TTS model with pitch contour features by using our proposed Pitch Contour Extraction module.

The remaining parts of this paper are structured as follows. In section 2, we describe our enhanced cross-lingual TTS system with the proposed pitch contour extraction module. Section 3 shows the experimental results of our proposed approach. Section 4 concludes with our findings and future works.

http://dx.doi.org/10.21437/Interspeech.2021-474

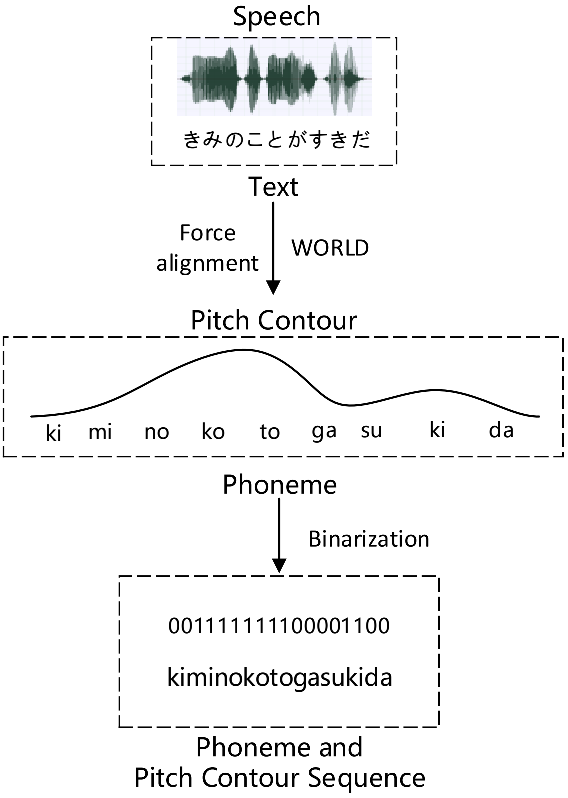

Figure 2: *Overview of the construction of pitch contour sequence dataset.*

# 2. Proposed approach

We adopt Tacotron [1] as our baseline end-to-end cross-lingual TTS model, which consists of an encoder-decoder-with-attention framework. We augment the baseline model with additional speaker embedding to synthesize speech from multiple speakers. We then propose a PCE module to predict binary pitch contour sequences. Meanwhile, we design an automatic procedure of constructing the dataset for training this module. The proposed model architecture in the inference phase is illustrated in Figure 1. The external features obtained by PCE module are concatenated with phoneme features and fed into the model.

## 2.1. Input Representations

Characters and phonemes are typically used as input representations in TTS model. Since characters cannot represent the actual pronunciation in many languages, the TTS model needs to learn the pronunciation from context, which increases the difficulty of model training. Most recently, some studies attempt to obtain a general representation for different kinds of languages, for example, Li et al. [12] proposes to model text via a sequence of Unicode bytes. However, according to [13], bytes representation may fail to synthesize intelligible Mandarin speech.

In our work, we employ two kinds of phonemes representation, language-dependent phonemes(LDP) and IPA, and evaluate the effects of using these representations for cross-lingual TTS.

### *2.1.1. Language-dependent Phonemes*

Lexicon or grapheme-to-phoneme(G2P) conversion model is usually used to convert graphemes into phonemes. And there is common phoneme representation in each language, such as initials and finals in Mandarin, ARPABET [21] in English. There are pitch information indicators of phonemes in Mandarin and English, such as tone and stress, except for Japanese. We extend finals in Mandarin with tone symbols, vowels in English with stress symbols, and some unvoiced symbols in Japanese. A language-dependent combined phonemes set is used for our experiments.

### *2.1.2. International Phonetic Alphabet*

IPA can be used for phoneme transcription in many languages. Based on the International Phonetic Alphabet Handbook [22] and IPA library[1], we obtain corresponding phoneme sequences with the help of G2P library and convert LDP to IPA by using a custom dictionary across three languages. Then we use IPA as a unified input representation for different languages. Similar to [23], we perform some replace and delete operations on three kinds of languages while designing this mapping function. Tone and stress information are kept for Mandarin and English.

## 2.2. Speaker Embedding

To train a multi-speaker TTS model, we initialize a look-up table for speaker embedding. The position of speaker embedding has a great impact on speaker consistency in multilingual speech synthesis. For the attention-based encoder-decoder speech synthesis model, the pronunciation information is easily entangled with the speaker's characteristics if the speaker embedding is added to the encoder part. As a result, it is hard for cross-lingual TTS model to synthesize mixed-lingual utterances with the same voice [9]. Therefore, we add the speaker embedding to the decoder by concatenating it with the context features learned by the attention module.

## 2.3. Attention Mechanisms

In the sequence-to-sequence speech synthesis models, there are several common attention mechanisms, such as Location Sensitive Attention(LSA) [24], GMM attention [25], and Forward Attention(FA) [26]. Considering the differences of these mechanisms may affect the speaker consistency in the cross-lingual TTS model, we evaluate the effects of these attention mechanisms on speaker consistency on mixed-lingual utterances.

## 2.4. Pitch Contour Extraction Module

On one hand, unlike tone languages and stress languages, there is no explicit pitch-related information in pitch-accent language text, which makes it challenging in the cross-lingual synthesis model. On the other hand, external features like energy, fundamental frequency($F_0$), prosodic-context [27] prove to be effective for improving the quality of synthesized speech. Based on these, we learn from [28] and propose a PCE module to learn pitch-related features from different kinds of languages, like duration module in [15, 16]. We expect all the languages in the cross-lingual synthesis model will benefit from the external information.

We use semi-supervised training to get the annotation of pitch contour information. Firstly, we use the Kaldi toolkit [29]

[1] https://github.com/open-dsl-dict/ipa-dict-dsl

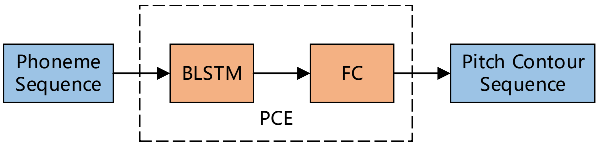

Figure 3: *Overview of the Pitch Contour Extraction(PCE) module in inference stage.*

to train an ASR system and perform a forced alignment for speech and text. We then use the WORLD [30] tool to extract pitch contours of the corresponding phonemes sequences. The phoneme in tone languages and stress-accent languages contain relevant pitch information, while pitch-accent languages are more concerned with the trend of pitch change rather than the specific value. At the same time, there is usually a large gap between male and female $F_0$ coverage, and we find that directly predict $F_0$ of the phoneme sequences cannot achieve a satisfactory result which is similar to [16]. So we simply convert the pitch contour sequences into binary sequences implied the increase or decrease of the pitch. This conversion can help to reduce the error involved by specific pitch values. Thereby obtaining the <phoneme, pitch contour > paired data from speech without manual annotation. Taking Japanese as an example, the construction of the pitch contour dataset is shown in Figure 2. After obtaining these paired data, we use them to train our proposed PCE model. The PCE module is illustrated in Figure 3. The phoneme and pitch contour sequence are used as the source and target sequences, respectively. We train the PCE model using the cross-entropy loss. Note that the trained PCE module is only used in the TTS inference phase because we can directly use the pitch contour sequence extracted from speech and text in training.

In our proposed method, we can factor mel-spectrogram synthesis into the following variables: text representation $t$, speaker embedding $s$, pitch contour sequence $p$. Therefore, our generative model $G$ is parameterized by trainable weights $W$, and the model is optimized by a loss function $L$, which penalizes the difference between the generated and ground truth mel-spectrogram in training phase:

$$\min_{W} L\left(G\left(t, s, p\right), mel_{GT}\right) \tag{1}$$

while in inference stage, the pitch contour sequence can be obtained by the PCE module, we obtain predicted mel-spectrogram as follows:

$$mel_{predicted} \sim G\left(t, s, PCE(t)\right) \tag{2}$$

# 3. Experiments

## 3.1. Data Setup

The cross-lingual TTS model is prone to having speaker consistency problems as people do not consider synthesized utterances in different languages are from the same person. To show the effectiveness of our method in maintaining speaker consistency, our experimental datasets include both male and female speakers. We train our model in three languages. Mandarin(CN) and English(EN) data is from Biaobei Technology [31] and LJSpeech [32], respectively. Japanese(JP) data is a subdataset of CSS10 [33]. All datasets contain only monolingual utterances. The first two speakers are female while the last one is male. To reduce the impact of data imbalance, we only use about 10-hour data from each dataset. The phonemes of each language are obtained based on the existing annotations and open-source G2P libraries. Finally, we use the method proposed in Section 2.4 to get phonemes and pitch contour sequences.

Table 1: *Mean Opinion Score (MOS) of ground truth audio.*

| MOS | Language | | |
|---|---|---|---|
| | CN | EN | JP |
| Naturalness | 4.94±0.08 | 4.95±0.07 | 4.74±0.11 |
| Consistency | 4.99±0.03 | 4.99±0.02 | 4.99±0.01 |

## 3.2. Experimental Setup

We follow [2] to add the stop token and residual modules to Tacotron model and use a look-up table speaker embedding as described in Section 2.2. All audios are resampled to 16 kHz. The audio features are represented as a sequence of 80-dim log-mel spectrogram frames, computed from 50ms windows shifted by 12.5ms. The PCE module consists of one layer of bidirectional LSTM and a fully connected layer. The dimension of phoneme embeddings and the hidden size of BLSTM layer are 200 and 512, respectively. The phoneme sequence and pitch contour sequence are passed through look-up tables and get 256-dim text features and 16-dim external features. We use Xavier Initialization to the weight parameters. To train our proposed TTS model, we take a batch size of 32 and use Adam optimizer with an initial learning rate of 0.001. We use a learning rate decay with warm-up strategy, the learning rate increases to 0.001 in the first 4000 steps, and then drops exponentially as the training continues. Our multilingual TTS model is optimized with mean square error (MSE) loss. We use WaveRNN to generate speech. To reduce the deviation caused by the quality of different datasets, we trained a universal vocoder using all datasets.

## 3.3. Evaluation

We use the Mean Opinion Score(MOS) to evaluate the naturalness and speaker consistency of the synthesized speech. The evaluation score is from 1 to 5 points in 0.5 point steps. For speaker consistency, 5 points indicate that the voices of the utterances are definitely the same speaker, 1 point indicates that the voices of the utterances are definitely not the same speaker. Similar to [13], we use 50 samples for each evaluation and each sample is rated by 6 raters. Naturalness and consistency MOS of ground truth audios from different languages are shown in Table 1. Related synthetic speech samples can be found at this website [2].

### *3.3.1. Comparison of Language-dependent Phonemes and IPA*

We first evaluate the performance of two input representations, LDP and IPA, on CN-EN-JP mixed-lingual utterances. The results are shown in Table 2. It shows that a cross-lingual TTS model with IPA representation achieves a higher MOS score in terms of naturalness and consistency in our experiments. We also find that the consistency of the Japanese male speaker is obviously lower than other female speakers while using the LDP representation. This may be due to the entanglement between speaker embedding and LDP representation, especially

[2] https://hyzhan.github.io/Interspeech2021/

Table 2: *Comparison of Language-dependent Phonemes and IPA on mixed-lingual utterances.*

| Input | utterances | CN speaker | | EN Speaker | | JP Speaker | |
|---|---|---|---|---|---|---|---|
| | | Naturalness | Consistency | Naturalness | Consistency | Naturalness | Consistency |
| LDP | CN-EN-JP | 2.77±0.28 | 4.44±0.38 | 3.04±0.26 | 4.62±0.31 | 3.24±0.27 | 1.64±0.47 |
| IPA | CN-EN-JP | **2.83±0.41** | **4.57±0.31** | **3.33±0.33** | **4.86±0.16** | **3.63±0.34** | **3.80 ±0.63** |

Table 3: *Impact of Pitch Contour Extraction module on all kinds of mixed-lingual utterances.*

| Model | utterances | CN speaker | | EN Speaker | | JP Speaker | |
|---|---|---|---|---|---|---|---|
| | | Naturalness | Consistency | Naturalness | Consistency | Naturalness | Consistency |
| baseline | CN-EN | 4.35±0.40 | 4.60±0.37 | 4.56±0.38 | 4.77±0.28 | 4.02±0.39 | 3.32±0.74 |
| | CN-JP | 2.75±0.33 | 4.39±0.37 | 2.70±0.43 | 4.30±0.36 | 3.16±0.40 | 4.17±0.59 |
| | EN-JP | 2.77±0.31 | 4.88±0.19 | 2.68±0.44 | 4.70±0.28 | 3.00±0.45 | 4.13±0.59 |
| | CN-EN-JP | 2.83±0.41 | 4.57±0.31 | **3.33±0.33** | 4.86±0.16 | 3.63±0.34 | 3.80±0.63 |
| proposed | CN-EN | **4.61±0.48** | **4.66±0.34** | **4.86±0.11** | **4.84±0.16** | **4.20±0.43** | **3.91±0.59** |
| | CN-JP | **3.20±0.19** | **4.55±0.28** | **3.20±0.24** | **4.48±0.30** | **3.69±0.20** | **4.70±0.34** |
| | EN-JP | **3.01±0.21** | **4.95±0.09** | **3.23±0.24** | **4.94±0.13** | **3.82±0.18** | **4.85±0.25** |
| | CN-EN-JP | **2.94±0.29** | **4.75±0.16** | 3.29±0.31 | **4.87±0.12** | **3.66±0.32** | **4.81±0.20** |

when there are both female and male utterances in training data. Based on the results, we use IPA as input representation in the following experiments.

### *3.3.2. Comparison of attention mechanisms on speaker consistency*

Since the cross-lingual TTS model may suffer from speaker consistency problems, we evaluate the effects of different attention mechanisms, including LSA, GMM, and FA, on speaker consistency while synthesizing all kinds of mixed-lingual utterances. Table 4 shows the results on mixed-lingual utterances with our proposed PCE module. From the table, we see that LSA is the most robust attention mechanism for different speakers. We also attempt to include Step-wise Monotonic Attention(SMA) [34], but fail to obtain a satisfying result. We infer that since IPA representation is similar to characters representation and it also needs to learn the actual pronunciation of IPA combinations, using soft attention could be better than hard attention. We use LSA mechanism in the following experiments.

Table 4: *Speaker Consistency of different attention mechanisms on all kinds of mixed-lingual utterances with PCE module.*

| Target Speaker | Attention | | |
|---|---|---|---|
| | LSA | GMM | FA |
| CN | 4.76±0.22 | 4.77±0.22 | **4.92±0.13** |
| EN | 4.71±0.25 | **4.84±0.22** | 4.83±0.15 |
| JP | **4.77±0.30** | 3.24±0.71 | 3.43±0.73 |
| Mean | **4.75±0.26** | 4.29±0.54 | 4.40±0.51 |

### *3.3.3. Impact of Pitch Contour Extraction module*

Finally, we evaluate the effects of the PCE module in the multi-speaker multilingual TTS model. The results are shown in Table 3. Overall speaking, our proposed model achieves a higher MOS score in terms of naturalness and consistency in almost all mixed-lingual utterances. This suggests that external pitch contour information can effectively improve cross-lingual TTS models. In particular, we notice that the speaker consistency of EN-JP utterances is higher than other mixed-lingual utterances for all three speakers in our proposed model. This contradicts with the intuition that speakers should perform better in utterances involving their native language. We speculate that compared to tone language, both stress language and pitch-accent language have similar sensitivity of $F_0$ [35], which is essential to distinguish different speakers. Therefore, the pitch information extracted by the proposed PCE module provides more effective features when synthesizing utterances in stress and pitch-accent language, which is the reason for the highest speaker consistency score on EN-JP utterances. We could also clearly see the naturalness of utterances largely decreases while JP text is involved. We believe this is caused by the difficulty of modeling the pitch-accent language and the data distribution. It may be helpful to achieve better synthetic speech quality by further optimizing the PCE module, which we leave for future works.

## 4. Conclusions

In this paper, we implement a multi-speaker multilingual TTS model with our proposed PCE module. We summarize the results as follows:

- IPA representation can help to learn disentangled information from different languages, and it is better than the language-dependent phonemes representation in our experiments;
- Location Sensitive Attention is a more robust attention mechanism in terms of speaker consistency in our proposed model.
- The Pitch Contour Extraction module proposed in this paper is helpful to multilingual TTS model and it greatly improves the naturalness and speaker consistency of mixed-lingual utterances;

For future work, we plan to continue investigating more robust and effective approaches to improve the performance further.

# 5. References

[1] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio, Q. Le, Y. Agiomyrgiannakis, R. Clark, and R. A. Saurous, "Tacotron: Towards end-to-end speech synthesis," in *Proc. Interspeech*, 2017, pp. 4006–4010.

[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 *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with tacotron," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[4] Y. Wang, D. Stanton, Y. Zhang, R.-S. Ryan, E. Battenberg, J. Shor, Y. Xiao, Y. Jia, F. Ren, and R. A. Saurous, "Style tokens: Unsupervised style modeling, control and transfer in end-to-end speech synthesis," in *Proceedings of the 35th International Conference on Machine Learning*. PMLR, 2018, pp. 5180–5189.

[5] S. Y. Um, S. Oh, K. Byun, I. Jang, C. Ahn, and H. G. Kang, "Emotional speech synthesis with rich and granularized control," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7254–7258.

[6] H. Ming, Y. Lu, Z. Zhang, and M. Dong, "A light-weight method of building an lstm-rnn-based bilingual tts system," in *2017 International Conference on Asian Language Processing (IALP)*. IEEE, 2017, pp. 201–205.

[7] Y.-J. Chen, T. Tu, C. chieh Yeh, and H.-Y. Lee, "End-to-End Text-to-Speech for Low-Resource Languages by Cross-Lingual Transfer Learning," in *Proc. Interspeech*, 2019, pp. 2075–2079.

[8] Y. Cao, S. Liu, X. Wu, S. Kang, P. Liu, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "Code-switched speech synthesis using bilingual phonetic posteriorgram with only monolingual corpora," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7619–7623.

[9] L. Xue, W. Song, G. Xu, L. Xie, and Z. Wu, "Building a Mixed-Lingual Neural TTS System with Only Monolingual Data," in *Proc. Interspeech*, 2019, pp. 2060–2064.

[10] Y. Cao, X. Wu, S. Liu, J. Yu, X. Li, Z. Wu, X. Liu, and H. Meng, "End-to-end code-switched tts with mix of monolingual recordings," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6935–6939.

[11] 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 *Proc. Interspeech*, 2019, pp. 2105–2109.

[12] B. Li, Y. Zhang, T. Sainath, Y. Wu, and W. Chan, "Bytes are all you need: End-to-end multilingual speech recognition and synthesis with bytes," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5621–5625.

[13] Y. Zhang, R. J. Weiss, H. Zen, Y. Wu, Z. Chen, R. Skerry-Ryan, Y. Jia, A. Rosenberg, and B. Ramabhadran, "Learning to Speak Fluently in a Foreign Language: Multilingual Speech Synthesis and Cross-Language Voice Cloning," in *Proc. Interspeech*, 2019, pp. 2080–2084.

[14] 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 *Advances in Neural Information Processing Systems(NeurIPS)*, 2019, pp. 3171–3180.

[15] C. Yu, H. Lu, N. Hu, M. Yu, C. Weng, K. Xu, P. Liu, D. Tuo, S. Kang, G. Lei, D. Su, and D. Yu, "DurIAN: Duration Informed Attention Network for Speech Synthesis," in *Proc. Interspeech*, 2020, pp. 2027–2031.

[16] 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 *International Conference on Learning Representations(ICLR)*, 2021.

[17] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 7055–7059.

[18] H. van der Hulst, R. Goedemans, and E. van Zanten, Eds., *A Survey of Word Accentual Patterns in the Languages of the World*. De Gruyter Mouton, 2010.

[19] 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 *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6905–6909.

[20] T. Nekvinda and O. Dušek, "One Model, Many Languages: Meta-Learning for Multilingual Text-to-Speech," in *Proc. Interspeech*, 2020, pp. 2972–2976.

[21] A. Klautau, "Arpabet and the timit alphabet," 2001.

[22] D. Horga, "Handbook of the international phonetic association. a guide to the use of the international phonetic alphabet cambridge: Cambridge university press (1999),(204 stranice)," *Govor*, vol. 16, no. 2, pp. 181–188, 1999.

[23] J. Donahue, S. Dieleman, M. Binkowski, E. Elsen, and K. Simonyan, "End-to-end adversarial text-to-speech," in *International Conference on Learning Representations(ICLR)*, 2021.

[24] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems(NIPS)*, 2015, p. 577–585.

[25] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[26] J.-X. Zhang, Z.-H. Ling, and L.-R. Dai, "Forward attention in sequence-to-sequence acoustic modeling for speech synthesis," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4789–4793.

[27] M. Aso, S. Takamichi, N. Takamune, and H. Saruwatari, "Acoustic model-based subword tokenization and prosodic-context extraction without language knowledge for text-to-speech synthesis," *Speech Communication*, vol. 125, pp. 53–60, 2020.

[28] T. Akiyama, S. Takamichi, and H. Saruwatari, "Prosody-aware subword embedding considering japanese intonation systems and its application to dnn-based multi-dialect speech synthesis," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 659–664.

[29] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011.

[30] M. Morise, F. Yokomori, and K. Ozawa, "World: a vocoder-based high-quality speech synthesis system for real-time applications," *IEICE Transactions on Information and Systems*, vol. 99, no. 7, pp. 1877–1884, 2016.

[31] D. T. Co., "The biaobei dataset," https://www.databaker.com/open_source.html.

[32] K. Ito, "The lj speech dataset," https://keithito.com/LJ-Speech-Dataset/, 2017.

[33] K. Park and T. Mulc, "Css10: A collection of single speaker speech datasets for 10 languages," 2019.

[34] M. He, Y. Deng, and L. He, "Robust Sequence-to-Sequence Acoustic Modeling with Stepwise Monotonic Attention for Neural TTS," in *Proc. Interspeech*, 2019, pp. 1293–1297.

[35] M. Gordon, "Disentangling stress and pitch-accent: a typology of prominence at different prosodic levels," *Word stress: Theoretical and typological issues*, pp. 83–118, 2014.