Interspeech 2022
18-22 September 2022, Incheon, Korea

Decoupled Pronunciation and Prosody Modeling in Meta-Learning-Based Multilingual Speech Synthesis

Yukun Peng, Zhenhua Ling

National Engineering Research Center of Speech and Language Information Processing, University of Science and Technology of China, Hefei, China

pyk@mail.ustc.edu.cn, zhling@ustc.edu.cn

Abstract

This paper presents a method of decoupled pronunciation and prosody modeling to improve the performance of meta-learning-based multilingual speech synthesis. The baseline meta-learning synthesis method adopts a single text encoder with a parameter generator conditioned on language embeddings and a single decoder to predict mel-spectrograms for all languages. In contrast, our proposed method designs a two-stream model structure that contains two encoders and two decoders for pronunciation and prosody modeling, respectively, considering that the pronunciation knowledge and the prosody knowledge should be shared in different ways among languages. In our experiments, our proposed method effectively improved the intelligibility and naturalness of multilingual speech synthesis comparing with the baseline meta-learning synthesis method.

Index Terms: text-to-speech, speech synthesis, multilingual, meta-learning

1. Introduction

In recent years, neural text-to-speech (TTS) synthesis has achieved remarkable progress [1, 2], and the naturalness of synthetic speech has been improved significantly. The acoustic model is a key component in neural TTS systems, which predicts acoustic features from input texts. One challenge of building acoustic models for multilingual TTS is the difficulty of constructing large-scale speech corpora for all languages, especially for a lot of minor languages in the world. Therefore, instead of training separate acoustic models for different languages, some studies resorted to building a unified acoustic model using a multilingual training dataset and sharing some model parameters among different languages [3–7].

Considering the difficulties of sharing knowledge among the text encoders for different languages, Nekvinda et al. [8] proposed to replace the original LSTM-based text encoder of Tacotron2 with a meta-learning encoder. The parameters in the meta-learning encoder were not trained separately for different languages, but were estimated by a parameter generator [9] conditioned on language embeddings, therefore it can better capture the commonality among languages. This method achieved better performance than building a single model for each language and building a unified model but with separate encoders for different languages [8]. One issue with this meta-learning-based multilingual TTS method is that only a single parameter generator was used in the encoder. Thus, it ignored that the pronunciation knowledge and the prosody knowledge should be shared in different ways among languages.

Pronunciation and prosody are two important characteristics of languages. The pronunciation differences among languages can be described by their different but overlapped phoneme sets. Several studies [5, 10, 11] have shown that replacing characters with phonemes as input can significantly improve the pronunciation accuracy of multilingual TTS. On the other hand, it is necessary to consider the prosody properties of different languages when building multilingual TTS systems. For example, adding tone embeddings was proposed to improve the naturalness of synthesizing tones in Chinese [12]. Predicting a binary fundamental frequency profile for each phoneme was employed to enhance Japanese synthesis performance in a multilingual model [13]. Considering that some languages may have similar phoneme sets, while others may have similar prosodic properties, it is reasonable to share the pronunciation knowledge and the prosody knowledge among languages separately in a unified multilingual TTS model.

Therefore, this paper proposes to decouple the pronunciation and prosody modeling in meta-learning-based multilingual speech synthesis. First, unlike the baseline meta-learning TTS method [8], which used character input, our input sequence includes International Phonetic Alphabet (IPA) symbols with word boundaries and prosody labels to introduce explicitly pronunciation-related and prosody-related descriptions. Second, a two-stream meta-learning-based model structure is designed, which has two encoders and two decoders for pronunciation and prosody modeling, respectively, and a shared attention module. Third, instead of using Mel-spectrograms, spectral features (i.e., Mel-cepstra) and excitation features (i.e., energy, fundamental frequency and voiced/unvoiced flag) are used as the prediction targets of the pronunciation stream and the prosody stream, respectively. Experimental results show that our proposed method significantly improved the intelligibility and naturalness of the baseline meta-learning multilingual TTS method [8].

2. Proposed method

In this section, we introduce the input and output representations, the architecture and the training strategy of our proposed acoustic model. For vocoding, a HiFi-GAN [14] vocoder is used in our implementation.

2.1. Input and output representations

In order to explicitly reflect the pronunciation similarity among different languages, we use IPA phonetic symbols as the basis of model input. Texts are converted to IPA symbol sequences by

This work was supported in part by the National Key R&D Program of China under Grant 2019YFF0303001, and in part by the National Nature Science Foundation of China under Grant 61871358.

10.21437/Interspeech.2022-831

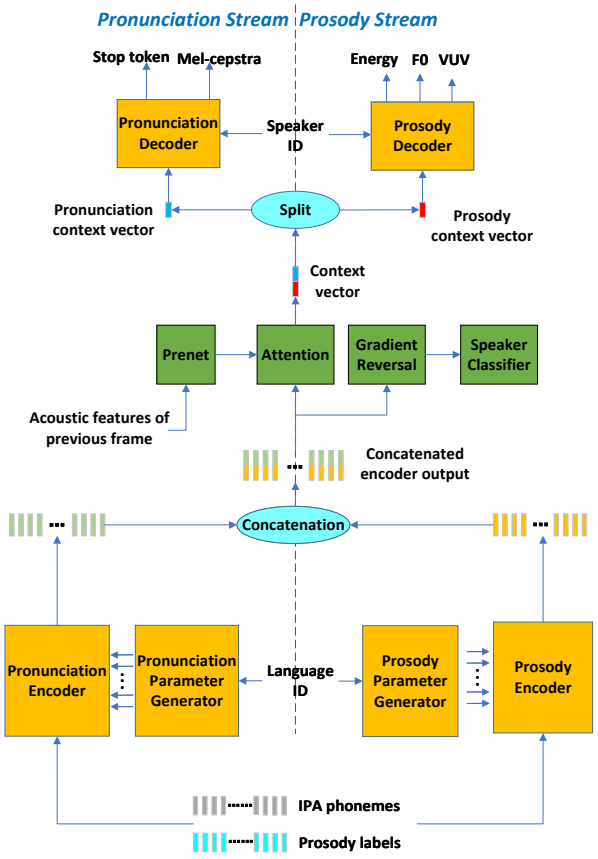


Figure 1: *The architecture of the proposed model.*

the Phonemizer tool[1]. To introduce prosody descriptions, a specific token is inserted at every word boundary in phoneme sequences. Besides, a prosody label is assigned to each phoneme to describe the tone or stress character of the phoneme. The prosody label is a one-hot vector with $M + N$ dimensions, where M corresponds to the number of tones of the tonal languages, and N corresponds to the number of stress categories of the non-tonal languages.

The previous meta-learning-based multilingual synthesis model [8] adopted Mel-spectrograms as model output, which mixes all pronunciation-related and prosody-related information. In our proposed method, different acoustic features are utilized to represent these two types of information. Specifically, spectral features, i.e., Mel-cepstra, are used as the output of the pronunciation stream in our model, while excitation features, i.e., energy, logarithmic F0 (logF0) and voiced/unvoiced (V/UV) flag are used as the output of the prosody stream in our model. These acoustic features are extracted by the STRAIGHT vocoder [15]. All features are normalized to zero mean and unit variance except the V/UV flag before acoustic modeling.

2.2. Model architecture

The architecture of our proposed model is shown in Fig. 1. It follows the attention-based sequence-to-sequence (seq2seq) framework for acoustic modeling and adopts Tacotron2 [2] as its backbone. As shown in Fig. 1, it contains a pronunciation stream and a prosody stream. These two streams have separate text encoders with separate parameter generators conditioned on language embeddings, and separate decoders for predicting different acoustic features. Two streams share an attention module to maintain the synchrony between the two streams.

In each encoder, a lookup table is employed to convert IPA phonemes and prosody labels into IPA embedding vectors and prosody embedding vectors, respectively. For each phoneme, the IPA embedding vector and the prosody embedding vector are concatenated and are then sent into the encoder. The pronunciation encoder and the prosody encoder contain two separate embedding tables, which allows the model to select necessary information from the input in a stream-dependent manner. Each encoder includes two 1-dimensional-convolutional (1D-Conv) layers and twelve highway 1D-Conv layers. The stride and size of the convolution kernel of each layer refer to DCTSS [16]. Each encoder relies on a parameter generator to obtain the weights and biases of its network. Following previous study [8], each parameter generator takes language embeddings as input. Let $\mathbf{X}_a \in \mathbf{R}^{D_a * L}$ and $\mathbf{X}_p \in \mathbf{R}^{D_p * L}$ denote the outputs of the pronunciation encoder and the prosody encoder, where D_a and D_p are the output dimensions of the two encoders and L is phoneme sequence length. Then, $\mathbf{X}_a$ and $\mathbf{X}_p$ are concatenated to get $\mathbf{X} \in \mathbf{R}^{(D_a+D_p)*L}$ for attention calculation.

The attention module is shared by both streams to keep inter-stream synchrony. Following previous work [13], location sensitive attention mechanism [17] is adopted here to align text encoder outputs to acoustic feature sequences. The predicted mel-cepstra, logF0 and energy at the previous frame are passed to the prenet. Following Tacotron2 [2], a long short-term memory (LSTM) layer is applied to obtain query vectors from the prenet output and the context vector of previous frame. The concatenated encoder outputs act as keys and values of attention calculation. Then, the derived context vector at each frame is further split into two parts according to D_a and D_p, which are sent into the decoders of two streams, respectively.

Each decoder contains an LSTM structure. In the prosody decoder, a lookup table converts the speaker ID into a speaker embedding vector, concatenated with context vectors as the input to LSTM. The LSTM output is projected through two separate linear transformations to predict energy and logF0, respectively. Meanwhile, the output is also projected into a scalar by a linear layer with sigmoid activation to predict a V/UV flag. The pronunciation decoder uses the same structure to predict a mel-cepstrum vector and a stop flag, respectively.

Following the previous work [13], an adversarial speaker classifier [18] with a gradient reversal layer is applied to the concatenated encoder output. It follows the principle of domain adversarial training [19] to remove the residual speaker information in the encoder output.

2.3. Model training

All model parameters are estimated simultaneously with a multilingual training corpus. The training loss consists of two parts. One part is the loss of reconstructing acoustic features, i.e., $Loss_{Rec}$. Mean squared error (MSE) loss function is adopted for mel-cepstra, energy and LogF0 prediction, while binary cross entropy (BCE) loss function is used for V/UV flag and stop flag prediction. These loss functions are added to get $Loss_{Rec}$. The other part is the speaker classifier loss, i.e., $Loss_{Spk}$, following Zhang et al.'s work [3]. The final loss function of our proposed model can be written as

$$Loss_{Total} = Loss_{Rec} - \lambda Loss_{Spk}, \quad (1)$$

where λ is a weight tuned manually in our implementation.

[1] https://github.com/bootphon/phonemizer

Table 1: *The data amount of the CSS/Common Voice datasets used in our experiments.*

Language	ZH	DE	FR	NL	RU
hours	6.5/1.0	16.1/4.8	19.2/3.0	14.1/1.3	21.4/3.4
speakers	1/6	1/39	1/22	1/11	1/8

3. Experiments

3.1. Dataset

Similar to previous study [8], our experiments used a subset of the multilingual single speaker dataset CSS10 [20] and selected clear speakers from the multilingual multi-speaker dataset Common Voice [21] to enhance CSS10. There are 10 languages in the original CSS10 dataset, and we used five of them in our experiments, including Mandarin (ZH), German (DE), French (FR), Dutch (NL) and Russian (RU). We removed too long and too short sentences in the dataset by setting the maximum and minimum sentence durations to 10s and 1s. Table 1 shows the data amount used for experiment. Then, the data of each language was split into a training set, a development set and a test set with a ratio of 8:1:1. All audios were sampled at 22.05 kHz. Our Hi-FiGAN vocoder was trained on the training sets of all languages.

3.2. Model implementation

STRAIGHT [15] was applied to extract acoustic features, which included 40-dimensional mel-cepstra, an energy, an F0 and a V/UV flag for each frame. The frame length was 25ms, and the frame shift was 10ms. These total 43-dimensional features were used as the input of our HiFi-GAN vocoder.

For the prosody labels, we had M=5 according to the five tones in Mandarin and N=3 according to the stress categories of non-tonal languages. Here, the stress categories included primary stressed vowel, secondary stressed vowel and non-stressed phoneme. The dimensions of IPA embeddings and prosody label embeddings were 512 and 16. The pronunciation encoder model had 14 1D-Conv layers with $D_a = 256$ in each layer, like DCTTS [16], while the prosody encoder model set $D_p = 128$ in each layer due to the low dimensionality of prosody features. The hidden unit numbers in the pronunciation decoder and the prosody decoder were 1024 and 256. We set the batch size to 50 and considered the language balance when composing each batch. In the prosody encoder and the prosody decoder, half of the initial learning rate was used for parameter updating to reduce overfitting. The learning rate of the remaining model parameters was initialized to 10^{-3}. The Adam optimizer was adopted and the learning rate decayed to half every 15000 steps. The weight λ in Eq. (1) was set to 0.05.

3.3. Baseline models

To verify the effectiveness of our proposed method, three baseline models were built for comparison. For a fair comparison, the output acoustic features of baseline models were the same as the 43-dimensional ones used in our proposed model and were optimized by the same loss function. Preliminary experimental results showed that using 43-dimensional acoustic features or 80-dimensional Mel-spectrograms in this model led to similar performance of synthetic speech.

Tacotron2: This model followed the original Tacotron2 architecture [2]. To be compatible with multilingual speech synthesis, it had a fully shared encoder with characters and a language ID as input. Similar to our proposed model, an adversarial speaker classifier was added to remove the speaker information contained in the encoder output, and a speaker embedding was connected to the input state of the LSTM decoder layer. Its hyperparameters were consistent with the ones in our proposed model.

Meta-char: This model was built following the baseline meta-learning-based multilingual TTS method [8]. including meta-learning encoder and character input.

Meta-IPA: This model was the same as Meta-char, and the only difference was that IPA phonemes and prosody labels were used as model input instead of characters, just like our proposed model. The difference between Meta-IPA and our proposed model was that the two-stream modeling was adopted in our proposed model. It should be noticed that our proposed model had more model parameters than Meta-char due to the additional encoder and decoder. However, we have conducted some preliminary experiments to confirm that increasing the number of parameters in Meta-IPA accordingly can't improve the performance of this model.

Table 2: *Objective evaluation results of different models.*

(a) *Mel-cepstrum distortion (MCD) (dB)*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	2.770	2.563	2.460	**2.445**
DE	3.240	3.111	2.992	**2.788**
FR	3.184	2.984	2.814	**2.784**
NL	3.074	2.953	2.905	**2.872**
RU	4.074	3.719	**3.449**	3.509

(b) *Root mean square error of F0 (F0-RMSE) (Hz)*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	44.591	37.876	35.006	**33.268**
DE	44.880	43.776	38.722	**37.220**
FR	23.418	23.686	20.260	**18.677**
NL	38.374	38.267	32.385	**31.743**
RU	51.631	48.365	43.850	**41.624**

(c) *Pearson correlation coefficient of F0 (F0-CORR)*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	0.424	0.604	0.654	**0.683**
DE	0.275	0.357	0.479	**0.522**
FR	0.309	0.370	0.523	**0.566**
NL	0.284	0.349	0.479	**0.514**
RU	0.094	0.226	0.354	**0.430**

(d) *Root mean square error of energy (EN-RMSE) (Hz)*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	0.234	0.201	0.174	**0.166**
DE	0.261	0.254	0.238	**0.188**
FR	0.260	0.250	0.221	**0.218**
NL	0.204	0.202	**0.180**	0.195
RU	0.362	0.324	0.280	**0.270**

(e) *V/UV flag error rate (V/UV-ERR) (%)*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	10.665	8.854	8.351	**7.696**
DE	10.248	10.589	9.093	**7.139**
FR	14.770	14.313	10.316	**8.884**
NL	17.108	16.223	16.785	**16.161**
RU	19.911	17.294	14.234	**13.304**

3.4. Objective evaluation

100 sentences were randomly selected from the test set of each language and were synthesized by our proposed model and baseline models. An objective evaluation was conducted on the synthetic speech. The evaluation metrics and results are presented in Table 2.

Table 3: *Character error rates (CER) (%) of different models.*

Language	Tacotron2	Meta-char	Meta-IPA	Proposed
ZH	38.4	28.7	25.9	**24.9**
DE	16.6	10.1	7.5	**6.4**
FR	31.3	21.2	18.8	**17.7**
NL	26.3	17.7	16.8	**15.0**
RU	39.2	24.4	15.6	**13.3**

Table 4: *Naturalness mean opinion scores (MOS) of different models with 95% confidence intervals, where "GT" means ground truth.*

	Tacotron2	Meta-char	Meta-IPA	Proposed	GT
ZH	2.25±0.15	3.08±0.15	3.25±0.15	3.49±0.14	3.69±0.15
DE	2.14±0.13	2.50±0.15	3.1±0.13	3.30±0.13	3.66±0.13
FR	2.53±0.15	3.54±0.14	3.24±0.15	3.86±0.12	3.95±0.11
NL	3.02±0.10	3.36±0.07	3.92±0.06	3.96±0.05	4.31±0.07
RU	2.8±0.17	3.52±0.13	3.79±0.13	3.96±0.11	4.36±0.11

From this table, we can see that the two meta-learning-based baselines performed better than Tacotron2, and Meta-IPA outperformed Meta-char. Our proposed model achieved the best performance on all metrics and all languages, except that Meta-IPA slightly outperformed our proposed method on the MCD metric of Russian and the EN-RMSE metric of Dutch. These results demonstrated the effectiveness of meta-learning-based acoustic model, the strategy of composing IPA phonemes and prosody labels as model input, and the two-stream model structure proposed in this paper on improving the accuracy of acoustic feature prediction.

Furthermore, we evaluated the intelligibility of synthetic utterances by sending them into the speech recognition engine of Google cloud platform[2]. The character error rate (CER) of speech recognition was used as the evaluation metric and the results are shown in Table 3. From this table, we can see that Tacotron2 had the highest CERs. Meta-IPA performed better than Meta-char, and our proposed model achieved the lowest CERs for all five languages. This indicates that in addition to meta-learning and using IPAs, the proposed method of decoupled pronunciation and prosody modeling also benefited the accurate pronunciation of synthetic speech.

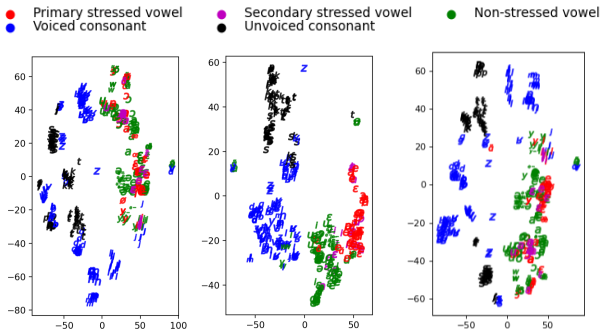
3.5. Subjective evaluation

A group of subjective listening tests were conducted to evaluate the naturalness mean opinion scores (MOS) of different models. The score range was from 1 (completely unnatural) to 5 (completely natural). 20 utterances[3] were used for each model and each language. For Mandarin, Russian and Dutch, 11, 7, and 8 native listeners were recruited offline, respectively. For German and French, the tests were conducted by crowdsourcing on Amazon Mechanical Turk[4], with 14 and 10 native listeners, respectively. The utterances recovered from ground truth acoustic features using the same HiFi-GAN vocoder were also included for comparison. The results are summarized in Table 4. We can see that the subjective evaluation results were consistent with the objective ones in Section 3.4. The Tacotron2 model had the lowest naturalness scores, while Meta-IPA performed better than Meta-char. Our proposed model achieved the highest naturalness among the four models for all five languages. The MOS differences between our proposed model and baseline models were significant according to the confidence intervals, except the difference between our model and Meta-IPA on Dutch. This confirms the effectiveness of our proposed method on improving the naturalness of multilingual speech synthesis.

[2]https://cloud.google.com/speech-to-text

[3]Audio samples can be found at https://pengyuk.github.io/dppmttsdemo.

[4]https://www.mturk.com



(a) (b) (c)

Figure 2: *t-SNE visualization of the outputs of (a) the pronunciation encoder of the proposed model, (b) the prosody encoder of the proposed model, and (c) the encoder of Meta-IPA for French. Each point corresponds to a phoneme with its phonetic symbol.*

3.6. Encoder analysis

To verify whether our model can decouple pronunciation and prosody representations effectively, the encoder outputs calculated from the test utterances of French were visualized by the t-SNE algorithm, as shown in Fig. 2. Different colors discriminate the stress categories. For better visualization, we divided the category of non-stressed phoneme into three sub-categories, i.e., non-stressed vowel, voiced consonant and unvoiced consonant.

We can see that the outputs of the pronunciation encoder in our model had similar distribution to that of the encoder in Meta-IPA, i.e., the phonemes with the same phonetic symbol were close to each other. In Fig. 2 (a), it can be further noticed that the instances of the consonants that have the same place and manner of articulation but only differ on voicing, e.g. /t/ and /d/, were also close to each other. On the other hand, French is a non-tonal language with stresses on vowels. In Fig. 2 (b), it can be observed that the outputs of the prosody vocoder in our model were clustered according to phoneme categories, but secondary stressed vowel are difficult to distinguish from primary stressed vowel. Similar patterns can also be observed in the outputs of the prosody encoder for Mandarin with tone categories which can't be shown in this paper due to limited space. All these results show that the outputs of the pronunciation encoder and the prosody encoder in our proposed model can capture the pronunciation and prosody characteristics of different languages separately, which indicates the effectiveness of decoupled modeling.

4. Conclusions

This paper proposed a two-stream model structure under the meta-learning framework to achieve decoupled pronunciation and prosody modeling for multilingual TTS. IPA symbols and prosody labels are employed as model input, and spectral features and excitation features are used as the prediction targets of the two streams, respectively. Experimental results have shown that our proposed model significantly outperformed the baseline models in both objective and subjective evaluations.

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 *et al.*, "Tacotron: Towards End-to-End Speech Synthesis," *Proc. Interspeech 2017*, pp. 4006–4010, 2017.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[3] 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," *Proc. Interspeech 2019*, pp. 2080–2084, 2019.

[4] S. Sitaram, S. K. Rallabandi, S. Rijhwani, and A. W. Black, "Experiments with Cross-lingual Systems for Synthesis of Code-Mixed Text." in *SSW*, 2016, pp. 76–81.

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

[6] B. Li and H. Zen, "Multi-Language Multi-Speaker Acoustic Modeling for LSTM-RNN Based Statistical Parametric Speech Synthesis," *Interspeech 2016*, pp. 2468–2472, 2016.

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

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

[9] E. A. Platanios, M. Sachan, G. Neubig, and T. Mitchell, "Contextual parameter generation for universal neural machine translation," *arXiv preprint arXiv:1808.08493*, 2018.

[10] M. Chen, M. Chen, S. Liang, J. Ma, L. Chen, S. Wang, and J. Xiao, "Cross-lingual, multi-speaker text-to-speech synthesis using neural speaker embedding." in *Interspeech*, 2019, pp. 2105–2109.

[11] 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 *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6935–6939.

[12] R. Liu, X. Wen, C. Lu, and X. Chen, "Tone learning in Low-Resource Bilingual TTS." in *INTERSPEECH*, 2020, pp. 2952–2956.

[13] H. Zhan, H. Zhang, W. Ou, and Y. Lin, "Improve Cross-Lingual Text-To-Speech Synthesis on Monolingual Corpora with Pitch Contour Information," *Proc. Interspeech 2021*, pp. 1599–1603, 2021.

[14] J. Kong, J. Kim, and J. Bae, "Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 022–17 033, 2020.

[15] H. Kawahara, I. Masuda-Katsuse, and A. De Cheveigne, "Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech communication*, vol. 27, no. 3-4, pp. 187–207, 1999.

[16] H. Tachibana, K. Uenoyama, and S. Aihara, "Efficiently trainable text-to-speech system based on deep convolutional networks with guided attention," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4784–4788.

[17] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," *Advances in neural information processing systems*, vol. 28, 2015.

[18] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. Lopez Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," *Advances in neural information processing systems*, vol. 31, 2018.

[19] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[20] K. Park and T. Mulc, "CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages," *Proc. Interspeech 2019*, pp. 1566–1570, 2019.

[21] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common Voice: A Massively-Multilingual Speech Corpus," in *LREC*, 2020.