INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Towards Universal Text-to-Speech

Jingzhou Yang and Lei He

Microsoft, China
{jingy,helei}@microsoft.com

Abstract

This paper studies a multilingual sequence-to-sequence text-to-speech framework towards universal modeling, that is able to synthesize speech for any speaker in any language using a single model. This framework consists of a transformer-based acoustic predictor and a WaveNet neural vocoder, with global conditions from speaker and language networks. It is examined on a massive TTS data set with around 1250 hours of data from 50 language locales, and the amount of data in different locales is highly unbalanced. Although the multilingual model exhibits the transfer learning ability to benefit the low-resource languages, data imbalance still undermines the model performance. A data balance training strategy is successfully applied and effectively improves the voice quality of the low-resource languages. Furthermore, this paper examines the modeling capacity of extending to new speakers and languages, as a key step towards universal modeling. Experiments show 20 seconds of data is feasible for a new speaker and 6 minutes for a new language.

Index Terms: multilingual, speech synthesis, neural text-to-speech, transfer learning

1. Introduction

The conventional text-to-speech (TTS) system employs different models to generate voices in different languages. Since these models are independent to each other, it is difficult to leverage the resources of other speakers and transfer one voice in a specific language to other languages. Moreover, when M voices in N languages need to be built, MN different models are required. Thus, it is costly to train, deploy and maintain a huge number of independent models. In statistical parametric and unit selection speech synthesis, multilingual systems are normally built with polyglot corpora. In addition, the data from voice conversion and phone mapping cross languages can also be used [1, 2]. Limited by the amount of the polyglot data, and the quality of voice conversion and phone mapping, it is difficult to build a high-quality multilingual voice. It is even more challenging to build a multilingual customized voice with limited amount of monolingual data.

In order to leverage the resources of other speakers and transfer the knowledge from one language to others, a single unified model is preferred [3]. In [4], a factorized multilingual deep neural network (DNN) model has been proposed, where speakers and languages are factorized by using language and speaker-specific layers in the DNN. In order to achieve cross-lingual synthesis, the training corpus is dominated by bi-lingual data. This limits the application of this framework. Similar framework based on long short-term memory (LSTM) recurrent neural network (RNN) has been studied in [5]. By using a mean tower and language bias towers to represent different languages, the model can be easily extended to a new language without changing the model structure. However, these models still rely on the heavily human designed frame-level linguistic features.

Recent years, end-to-end (E2E) models have been widely used in speech synthesis [6, 7, 8, 9, 10, 11, 12], where the models can be directly trained on text-speech pairs with minimal engineering efforts. Based on the encoder-decoder E2E framework, various multilingual TTS approaches have been proposed [13, 14, 15, 16, 17]. [13] investigates a multilingual model based on Deep Voice 3 [18] for Indian languages, but only mel cepstral distortion is used in evaluation. In [14], based on Tacotron 2 [10], UTF-8 Byte sequence is used as the input of the model, and a fixed speaker embedding for each speaker is used. This system may be suboptimal for the character-based languages, as different characters may share the same Bytes but have totally different pronunciations. This might cause problems especially when the amount of training data is limited. Previous work [15] extends a multi-speaker system [8] to multilingual by using multiple language-depended speaker encoders. This makes the speaker and language information coupled. Thus, it is difficult to control the accent in cross-lingual synthesis. Moreover, the number of encoders grows with the number of languages. This limits of flexibility of the system when extending to new languages. [16] presents a multilingual model based on Tacotron 2 with additional speaker and language embeddings, and incorporates an autoencoding input to stabilize training. Adversarial training is deployed to reduce the speaker information of the text encoder outputs. [17] studies a similar framework based on Tacotron 2, but the speaker embeddings are from a speaker verification system. However, all these previous work only examined very limited number of languages, i.e. no more than four. Thus, these conclusions could not be applied directly to universal TTS modeling.

Towards universal TTS modeling, this paper studies a multilingual TTS framework based on transformer [11], and examines the model on a large scale, with around 1250 hours of data from 50 language locales. Normally, the amount of training data in different languages is highly unbalanced, thus the low-resource languages cannot be adequately trained in a massive multilingual model. This paper proposes an effective training strategy to balance the data from different language locales. By using this strategy, the overall performance of the multilingual model can be improved. In order to achieve universal modeling, any speaker in any language needs to be modeled, even for the unseen speakers and languages. Thus, speaker and language extensions are also studied. The shared model structure enables extension to new speakers and languages and achieve cross-lingual synthesis with very limited amount of data.

The paper is organized as follows. In section 2 the general framework of the multilingual transformer TTS model is introduced, and the possible issues towards universal modeling are discussed. Experiments and corresponding evaluation results are presented in section 3. Finally, conclusions and the future work are discussed in section 4.

http://dx.doi.org/10.21437/Interspeech.2020-1590

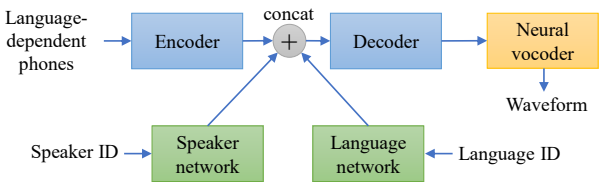

Figure 1: *The framework of the multilingual system.*

2. Multilingual transformer TTS

In this section, the general framework of the multilingual transformer TTS is introduced. To achieve universal modeling, the possible challenges and solutions are also discussed.

2.1. The general framework

The general framework of the multilingual transformer TTS system is illustrated in Figure 1. It is comprised of three main components: (1) The global conditions, i.e. the speaker and language conditions, that represent the speaker and language global characteristics; (2) The sequence-to-sequence synthesizer, which predicts a mel-scale spectrogram from a sequence of phone inputs, conditioned on the speaker and language networks; (3) The neural vocoder, that converts the predicted spectrogram into time domain waveforms.

The speaker and language conditions are used to control the synthesizer to generate speech of different speakers and languages. The vectors from trainable networks are used as global conditions in this work. The speaker and language conditions can be viewed as to characterize the voice of the speaker, and the global prosody of the language. In general, the speaker network could be any network that can provide discriminative information for speakers. For example, the network can be a pre-trained model from a speaker verification system [19]. In this work, the speaker network is a lookup table (LUT) followed by a mapping network that makes the generated features have the similar dynamic range. The input to the speaker network is a one-hot vector encoding the speaker identity. The language network has the same structure as the speaker network, and the input one-hot vector represents the language identity.

The sequence-to-sequence synthesizer is an encoder-decoder architecture based on transformer [11], which predicts a mel spectrogram directly from the input text. In this work, the text is mapped to a sequence of phones as inputs to reduce pronunciation errors. The input phones are represented by one hot vectors followed by a LUT in the encoder. To avoid the engineering efforts on clustering the phones with similar pronunciations in different languages, the input phones are language-dependent, i.e. they are not shared cross languages. Although different languages use different phone sets, the phones with similar pronunciations tend to be clustered together. This is illustrated in Figure 2, where the phone embeddings of zh-CN, en-US and en-GB[1] in the LUT are from a well-trained model. The speaker and language network outputs are concatenated with the encoder output and then passed to the attention layers. By using training data from different speakers and languages, the global characteristics of different speakers and languages can be represented by the speaker and language conditions. In synthesis, by using different combinations of input phones, speaker and language identities, cross-lingual synthesis can be achieved for any speaker in the training set.

[1]The abbreviation consists of the language code and the locale ID, e.g. zh is Chinese, CN represents China.

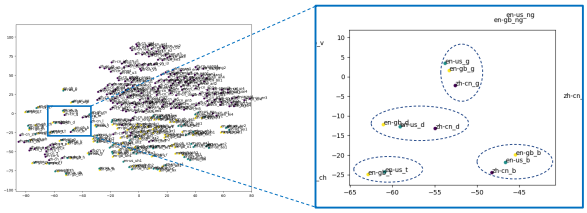

Figure 2: *The t-SNE visualization of the phone embeddings.*

The neural vocoder can be any vocoder that converts mel spectrograms to waveforms, e.g. WaveNet [20], WaveRNN [21] or LPCNet [22]. WaveNet is used in this paper.

2.2. Towards universal modeling

It is a challenging task to enable a single TTS model to synthesize any voice in any language. To achieve universal TTS, there are many issues need to be addressed: the lexicon and phonetic representation of massive number of languages, data imbalance cross speakers and languages, and the limit of the model capacity. In this work, the data imbalance issue is mainly discussed. The lexicon and phonetic representation for different languages are assumed to be ready.

A universal model not only saves training and maintenance cost, but also helps model generalization. However, as the number of languages increases, data imbalance becomes a severe problem in real applications. There might be hundreds of hours of data for some languages, but only hours of data (or even less) for low-resource languages. To overcome the data imbalance issue, this work applies a data sampling strategy [23] to balance the data of different languages in training. This sampling strategy controls the proportions of different languages feeding into a batch. For language i, the number of utterances is assumed to be N_i. In a naive strategy of sampling from the whole training set, the probability of the sample from language i is $c_i = N_i / \sum_j N_j$. To alleviate the data imbalance problem, a scaling factor α is introduced to control the sampling probabilities for different languages:

$$p_i \propto c_i^{\alpha} \tag{1}$$

where $\alpha \in [0, 1]$. When $\alpha = 0$, a uniform distribution is used. When $\alpha = 1$, the true data distribution is retrieved. There might be many speakers in one language, and the amount of data could be highly unbalanced among the speakers. Analogously, the balance strategy can be applied to the speakers in each language.

To achieve universal TTS, any speaker and language need to be modeled, even for the unseen speakers and languages. Thus, model extension is a key step to approach universal modeling. In our multilingual model, LUTs are used, and the followed networks are shared by all speakers, languages and phones. Thus, the model can be easily extended to new speakers and languages without changing the model structure. Moreover, the shared structure helps transfer learning from other speakers and languages. For example, the shared encoder makes the phones with similar pronunciations tend to have similar phone embeddings in the LUT as illustrated in Figure 2. This clustering property helps to learn the pronunciations in a new language from the existing ones. The modeling capacity of extending to new speakers and languages will be examined in the experimental section.

In terms of the vocoder, as the input mel spectrogram contains information about the speaker and language, a universal WaveNet vocoder can be trained to generate waveforms for any speaker and language. In this work, the universal vocoder is trained with around 100 hours of data, which is a subset of the

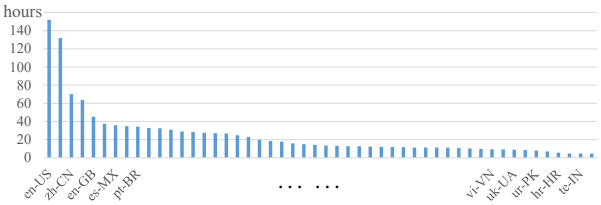

Figure 3: *The data distribution over 50 language locales.*

Table 1: *The naturalness MOS in different languages.*

Language Data size	en-US 20h/150h	de-DE 10h/30h	vi-VN 7h/7h	te-IN 5h/5h
Rec.	4.51±0.10	4.22±0.13	4.23±0.15	4.47±0.13
Single	4.34±0.08	4.19±0.08	4.14±0.09	3.40±0.13
Multi	**4.30±0.08**	4.07±0.08	3.83±0.10	3.59±0.12
+LgB	4.03±0.09	**4.08±0.08**	**4.03±0.09**	**3.89±0.11**
+SpkB	4.19±0.08	4.03±0.09	3.90±0.09	3.73±0.11

data to train the multilingual model. Once the universal vocoder is trained, it can be applied to the spectrogram from any speaker in any language without additional adaptation or fine-tuning. Thus, this universal vocoder is used throughout our experiments without further modification.

3. Experiments

The multilingual model is studied on a large scale. The training corpus is comprised of around 1250 hours professional recordings from 50 language locales. In this work, the same language from different locales is treated independently, i.e. different phone sets and language identities are used. The amount of training data is highly unbalanced for different locales, with range from several hours to hundreds of hours. The data distribution over 50 language locales is illustrated in Figure 3.

In training, all audios are down-sampled to 16 kHz, and the beginning and ending silences are trimmed to a fixed length, i.e. 30 ms. Except the neural vocoder (which is pre-trained), the whole network is trained jointly on 4 Tesla V100 GPUs. The Adam optimizer is used with initial learning rate 10^{-3}, and exponential decay after 100k steps. The minimum learning rate is set to 10^{-5}. In experiments, crowd-sourced subjective listening tests are used to evaluate the quality of the synthesized speech. The mean opinion score (MOS) is used to rate the naturalness and similarity to the target speakers, with range from 1 to 5.

3.1. Data balance training strategies

It is a challenging task to train a massive multilingual model with a large amount of unbalanced data. The trained model may bias to the high-resource languages and yield poor performance for the low-resource languages. To overcome this issue, the data balance training strategy discussed in section 2.2 is implemented. In this section, two specifications of the data balance strategy are studied. In training, the sampling probabilities over 50 language locales are given by equation (1), and for each locale, the sampling probabilities over the speakers from this locale are also give by equation (1). This is referred to as language-balanced training strategy, and it is denoted as "LgB" in Table 1. When the scaling factors α are set to be 0, both the languages and the speakers in each language are equally distributed. Another strategy is called speaker-balanced training strategy, where the sampling probability of each language is set to be proportional to the number of speakers in that language, and the sampling probabilities of the speakers in each language are given by equation (1). This training strategy is indicated

Table 2: *The naturalness MOS to the de-DE speaker.*

Language	en-US	vi-VN	te-IN
Rec.	4.55±0.09*	4.50±0.11*	4.59±0.14*
Multi	**3.97±0.10**	3.78±0.09	3.54±0.13
+LgB	3.86±0.09	**3.79±0.07**	**3.79±0.11**

Table 3: *The similarity MOS to the de-DE speaker.*

Language	en-US	vi-VN	te-IN
Rec.	1.27±0.08*	1.12±0.07*	1.52±0.12*
Multi	2.93±0.19	2.69±0.17	2.70±0.17
+LgB	2.98±0.19	2.50±0.18	2.47±0.16

by "SpkB" in Table 1. When the scaling factor is set to 0, the speakers in the whole training corpus are equally distributed. To avoid too much impact on the high-resource languages and speakers, in training all scaling factors are set to be 0.2.

The naturalness MOS results for intra-lingual synthesis are tabulated in Table 1. In this table, the naive multilingual system and the systems using language-balanced and speaker-balanced training are compared in 4 languages, i.e. en-US (English), de-DE (German), vi-VN (Vietnamese) and te-IN (Telugu). The second row of the table gives the amount of data for the target speakers and target languages. For comparison, the third and fourth rows also give the scores of the recordings and single speaker transformer TTS model. The single speaker model outperforms the multilingual systems for the speakers with data no less than 7 hours. This might be limited by the model capacity of the multilingual model, where a single model needs to handle a large number of speakers and languages. The multilingual models outperform the single speaker model on low-resource language te-IN, as transfer learning from other languages benefits the low-resource languages. In terms of the multilingual models, the model using language-balanced training has the best overall performance, where the low-resource languages are improved, and the performance degradation on high-resource language en-US is as expected.

The cross-lingual synthesis experiments are conducted on the de-DE speaker. The naturalness and similarity MOS to this speaker are tabulated in Table 2 and Table 3 respectively. Only the de-DE recordings of the target speaker are available, thus the recordings from other speakers are used for comparison, and the corresponding scores are marked with "$*$" in the tables. In cross-lingual synthesis, compared with the naive multilingual model, the model using language-balanced training yields good overall performance, with improvements on low-resource languages and reasonable regression on high-resource language en-US. The experiments also show that, without using bilingual data, cross-lingual synthesis can be achieved with naturalness MOS around 3.8. If the cross-lingual experiments were carried on the vi-VN or te-IN speaker, more improvements on low-resource languages can be observed. The cross-lingual speaker similarity is evaluated in Table 3. In evaluation, the de-DE recordings of the de-DE speaker are used as reference. The synthesized cross-lingual speech of the de-DE speaker and the recordings of other speakers are compared. As shown in Table 3, the similarity scores of the synthesized speech range from 2.5 to 3, and are much higher than the scores for other speakers. However, the overall similarity scores are relatively low. The possible reason is that different languages have different pitch ranges and variations [24, 25, 26]. These differences may hinder identification of a person in cross-lingual scenario.

In order to better understand the similarity score in cross-lingual synthesis, the cross-lingual similarity MOS test is conducted on a bilingual female speaker with fr-CA and en-CA recordings. Only fr-CA data of this speaker are used in training.

Table 4: *The MOS to the new zh-CN speaker.*

	Naturalness		Similarity	
Language	zh-CN	en-US	zh-CN	en-US
Rec.	3.78±0.13	3.37±0.20	4.32±0.12	3.77±0.12
20s	3.61±0.07	3.72±0.08	4.21±0.12	3.43±0.12
1m	3.62±0.07	3.76±0.08	4.32±0.10	3.49±0.11
5m	3.68±0.07	3.71±0.08	4.20±0.12	3.35±0.12
10m	3.63±0.07	3.61±0.09	4.27±0.11	3.25±0.14

Table 5: *The MOS to the new en-GB speaker.*

	Naturalness		Similarity	
Language	en-GB	zh-CN	en-GB	zh-CN
Rec.	4.56±0.11	–	4.49±0.11	–
20s	4.08±0.08	3.61±0.07	4.36±0.12	2.60±0.24
1m	4.16±0.09	3.57±0.08	4.42±0.12	2.26±0.23
5m	4.24±0.08	3.34±0.07	4.47±0.12	2.30±0.23
10m	4.24±0.08	3.19±0.08	4.36±0.13	2.36±0.23

In similarity evaluation, her fr-CA recordings are used as reference, her en-CA recordings and the synthesized en-CA (cross-lingual) speech are compared. Because of the variations of the pitch and intonation in different languages [26], the similarity score of the recordings is 2.29, which is even less than the score of the synthesized cross-lingual speech 2.97.

3.2. Extend to new speakers and languages

A massive multilingual model is trained with a large number of speakers and languages, but the model still cannot cover all speakers and languages. To approach universal TTS, the model needs to be easily extended to new speakers and languages with limited amount of data. In this subsection, the modeling capacity of extending to new speakers and languages is examined. The multilingual model using language-balanced training, which gives good overall performance, is used in all extension experiments. There are many ways to extend the multilingual model to a new speaker. As discussed in section 2.1, a LUT is used in the speaker network. Thus in speaker extension, we can only update the LUT or the whole speaker network. In addition, we can also refine the whole multilingual model. Our initial experiments show that refining the whole model gives better speaker similarity. Thus, the whole model is updated in the extension experiments with learning rate 10^{-5}.

When extending to a new speaker, if only the data of the new speaker are used, the model might overfit to the target speaker. To achieve better generalization and cross-lingual synthesis, the data from the new speaker and existing speakers are used in model refining. Normally, the amount of data for the new speaker is very limited, whereas the data amount for other speakers is very large. Then language-balanced training discussed in the previous subsection is used.

To study the extension capacity of the model for different types of speakers, a non-professional zh-CN female speaker and a professional en-GB male speaker are examined. Table 4 tabulates the naturalness and similarity MOS to the zh-CN speaker. This speaker is bilingual, but only her zh-CN recordings are used in training. In intra-lingual synthesis, more training data yields better model performance, but when the amount of data exceeds 5 minutes, the model performance is degraded. This might be limited by the data quality of this non-professional speaker, e.g. some phones are not accurately pronounced. Thus more data introduces more cumulative errors. It is worth noting that in cross-lingual synthesis all the synthesized speech have better naturalness than her en-US recordings, as the zh-CN speaker's English is accented. In terms of the intra-lingual speaker similarity, all the models trained with different amounts of data can yield good similarity close to her zh-CN recordings. In cross-lingual similarity evaluation, her zh-CN recordings are used as references. The synthesized speech has a small gap to her English recordings. Similar experiment is carried on the en-GB male speaker, and the evaluation results are tabulated in Table 5. In intra-lingual synthesis, generally more data yields better naturalness, and all the models can achieve good speaker similarity close to his recordings. However, in cross-lingual synthesis, more data of the en-GB speaker degrades the model performance. The reason is that the vocal characteristics of the target speaker can be learned with very limited amount of data, say 20 seconds, thus more intra-lingual data does not help cross-lingual synthesis but introduces more data inconsistencies.

Table 6: *The naturalness MOS to the new id-ID speaker.*

	Single	Multi.
Rec.	4.50±0.11	
6m	failed	3.63±0.11
1h	failed	4.22±0.08
3h	failed	4.19±0.09

Table 7: *The naturalness MOS to the new ru-RU speaker.*

	Single	Multi.
Rec.	4.57±0.13	
6m	failed	3.08±0.12
1h	failed	4.14±0.09
3h	2.26±0.17	4.14±0.09

In language extension, two languages with different average utterance lengths are examined. These languages are id-ID (Indonesian) and ru-RU (Russian), and the average length for id-ID is longer. To reduce the impact of similar languages, a multilingual model trained with 14 languages is used in language extension. In experiments, only the data of the target language are used to refine the multilingual model. The single speaker transformer models trained from scratch are compared with the refined multilingual models using different amounts of data. The first experiment is carried on id-ID with around 3 hours of data. As shown in Table 6, the single speaker model is failed to be trained, whereas 6 minutes of data is feasible for multilingual model language extension. Generally, more data yields better model performance. The 1-hour multilingual model is much better than the 6-minute model. Another experiment is conducted on ru-RU with around 3 hours of data, and the naturalness scores are tabulated in Table 7. Similar conclusions can be drawn, but the single speaker model can be trained with 3 hours of data, as a shorter average utterance length facilitates model training. However, the 3-hour single speaker model is much worse than the 6-minute multilingual model. In these experiments, language extension can be achieved with very limited amount of data, given that the shared model structure helps transfer learning from other languages as discussed in section 2.

4. Conclusions and the future work

This paper studies a massive multilingual framework towards universal modeling. To address the data imbalance issue, an effective data balance strategy has been examined. A massive model still cannot cover all speakers and languages, thus extension to unseen speakers and languages is also studied. Transfer learning from other speakers and languages helps model extension. Experiments show that 20 seconds of data is feasible for a new speaker and 6 minutes for a new language. Future work will study combination or joint training with a universal front-end model in language extension, and scale the multilingual framework to more non-TTS data.

5. References

[1] C. Traber, K. Huber, K. Nedir, B. Pfister, E. Keller, and B. Zellner, "From multilingual to polyglot speech synthesis," in *Sixth European Conference on Speech Communication and Technology*, 1999.

[2] A. W. Black and K. A. Lenzo, "Multilingual text-to-speech synthesis," in *Proceedings of ICASSP*, vol. 3. IEEE, 2004, pp. iii–761.

[3] I. Demirsahin, M. Jansche, and A. Gutkin, "A unified phonological representation of south Asian languages for multilingual text-to-speech," in *Proceedings of The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*, August 2018, pp. 80–84.

[4] Y. Fan, Y. Qian, F. K. Soong, and L. He, "Speaker and language factorization in DNN-based TTS synthesis," in *Proceedings of ICASSP*. IEEE, 2016, pp. 5540–5544.

[5] B. Li and H. Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *Proceedings of Interspeech*, 2016, pp. 2468–2472.

[6] J. Sotelo, S. Mehri, K. Kumar, J. F. Santos, K. Kastner, A. Courville, and Y. Bengio, "Char2Wav: End-to-end speech synthesis," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2017.

[7] S. Ö. Arik, M. Chrzanowski, A. Coates, G. Diamos, A. Gibiansky, Y. Kang, X. Li, J. Miller, A. Ng, J. Raiman *et al.*, "Deep voice: Real-time neural text-to-speech," in *Proceedings of International Conference on Machine Learning (ICML)*, 2017, pp. 195–204.

[8] Y. Taigman, L. Wolf, A. Polyak, and E. Nachmani, "Voiceloop: Voice fitting and synthesis via a phonological loop," in *Proceedings of International Conference on Learning Representations (ICLR)*, 2018.

[9] Y. Wang, R. Skerry-Ryan, D. Stanton, Y. Wu, R. J. Weiss, N. Jaitly, Z. Yang, Y. Xiao, Z. Chen, S. Bengio *et al.*, "Tacotron: Towards end-to-end speech synthesis," in *Proceedings of Interspeech*, 2018, pp. 4006–4010.

[10] 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 *Proceedings of ICASSP*. IEEE, 2018, pp. 4779–4783.

[11] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6706–6713.

[12] R. Valle, J. Li, R. Prenger, and B. Catanzaro, "Mellotron: Multispeaker expressive voice synthesis by conditioning on rhythm, pitch and global style tokens," in *Proceedings of ICASSP*. IEEE, 2020, pp. 6189–6193.

[13] P. Baljekar, S. K. Rallabandi, and A. W. Black, "An investigation of convolution attention based models for multilingual speech synthesis of Indian languages." in *Proceedings of Interspeech*, 2018, pp. 2474–2478.

[14] 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 *Proceedings of ICASSP*. IEEE, 2019, pp. 5621–5625.

[15] E. Nachmani and L. Wolf, "Unsupervised polyglot text-to-speech," in *Proceedings of ICASSP*. IEEE, 2019, pp. 7055–7059.

[16] 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," *arXiv preprint arXiv:1907.04448*, 2019.

[17] Z. Liu and B. Mak, "Cross-lingual multi-speaker text-to-speech synthesis for voice cloning without using parallel corpus for unseen speakers," 2019.

[18] W. Ping, K. Peng, A. Gibiansky, S. O. Arik, A. Kannan, S. Narang, J. Raiman, and J. Miller, "Deep voice 3: Scaling text-to-speech with convolutional sequence learning," 2017.

[19] Y. Jia, Y. Zhang, R. Weiss, Q. Wang, J. Shen, F. Ren, P. Nguyen, R. Pang, I. L. Moreno, Y. Wu *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *Advances in Neural Information Processing Systems (NIPS)*, 2018, pp. 4480–4490.

[20] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, "WaveNet: A generative model for raw audio," in *9th ISCA Speech Synthesis Workshop*, 2016, pp. 125–125.

[21] N. Kalchbrenner, E. Elsen, K. Simonyan, S. Noury, N. Casagrande, E. Lockhart, F. Stimberg, A. Oord, S. Dieleman, and K. Kavukcuoglu, "Efficient neural audio synthesis," in *Proceedings of International Conference on Machine Learning (ICML)*, 2018, pp. 2415–2424.

[22] J.-M. Valin and J. Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *Proceedings of ICASSP*. IEEE, 2019, pp. 5891–5895.

[23] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry *et al.*, "Massively multilingual neural machine translation in the wild: Findings and challenges," *arXiv preprint arXiv:1907.05019*, 2019.

[24] I. Mennen, F. Schaeffler, and G. Docherty, "Cross-language differences in fundamental frequency range: A comparison of English and German," *The Journal of the Acoustical Society of America*, vol. 131, pp. 2249–2260, March 2012.

[25] P. Keating and G. Kuo, "Comparison of speaking fundamental frequency in English and Mandarin," *The Journal of the Acoustical Society of America*, vol. 132, no. 2, pp. 1050–1060, May 2012.

[26] B. Andreeva, G. Demenko, M. Wolska, B. Möbius, F. Zimmerer, J. Jügler, M. Oleskowicz-Popiel, and J. Trouvain, "Comparison of pitch range and pitch variation in Slavic and Germanic languages," in *Proceedings to the 7th Speech Prosody Conference.* ISCA, 2014, pp. 776–780.