ZERO-SHOT CROSS-LINGUAL TRANSFER USING MULTI-STREAM ENCODER AND EFFICIENT SPEAKER REPRESENTATION

Yibin Zheng, Zewang Zhang, Xinhui Li, Wenchao Su, and Li Lu

Tencent Inc, China

{ybzheng,hiccupli}@tencent.com

ABSTRACT

We propose a novel method for zero-shot cross-lingual TTS task by using multi-stream text encoder and efficient speaker representation. Specifically, a unified multi-stream text encoder that takes both advantages of Transformer and CBHG is proposed to retain multiple hypotheses about input representations. For Transformer based stream, a multi-stream Transformer is further proposed to strengthen these hypotheses. Then the speaker representations are extracted from audio signals by a speaker encoder with a random sampling mechanism and a language adversarial loss, aiming to extract speaker embedding features that are independent of both content information and language identity. Meanwhile, we propose an efficient zero-shot cross-lingual transfer strategy with the help of other target lingual speakers' data and a language-balanced sampling strategy. The Experimental results show the proposed method not only could achieve higher speech quality and speaker similarity (with an average absolute improvement of 0.38 and 0.27 in MOS respectively) for zero-shot cross-lingual transfer, but also helpful for few-shot cross-lingual transfer in which has multi-lingual data. [1]

***Index Terms*—** zero-shot, cross-lingual transfer, multi-stream, speaker representations, end-to-end neural TTS

1. INTRODUCTION

End-to-end neural text-to-speech (TTS) models [1, 2, 3] have been developed to produce highly natural monolingual speech in a single speaker's voice. And these models have been extended to enable control of speaker identity and speaking styles [4, 5, 6, 7, 8]. However, it is still a major challenge to extend such models to support cross-lingual transfer [9], which is observed as growing demands in practical scenarios when only one lingual data is available. Such needs call for zero-shot cross-lingual transfer techniques to build a system which can synthesize speech in a specific language not spoken by the target speaker.

Researches on cross-lingual TTS have shifted from conventional statistical parameter speech synthesis [10] to end-to-end based methods [11, 12, 13]. And from previous studies, the challenges for zero-shot cross-lingual transfer mainly consist of three key components: unified text encoder design, speaker representation methods, and zero-shot cross-lingual transfer strategies. For unified text encoder design, previous works are committed to design a unified text representation that can be shared by all languages and speakers, like phoneme, byte or International Phonetic Alphabet (IPA) [14]. In [15], the authors proposes a code-switched TTS models by using a shared encoder with two separate language-dependent text encoder and language embeddings. However, the performance is not that good when synthesizing a Mandarin speaker's English speech. Recently, [14] presents a multi-lingual TTS model by using a unified phoneme input representation. However, their model heavily relies on data from a large number of speakers per language to achieve good performance in cross-lingual transfer. Most importantly, since TTS could be viewed as a one-to-many mapping process, where every processing step usually ends with selection from multiple alternative hypotheses. For examples, the generation process might be affected by styles, prosody and etc. However, all of these models employ only one single-stream for the unified text encoder, thus there is no storage for alternative hypotheses. Inspired by general idea as stacking ensemble models leveraging multiple hypotheses [16], we propose a multi-stream text encoder to retain multiple hypotheses about input representations. Meanwhile, for Transformer-based stream, a multi-stream Transformer [17] is proposed to further strengthen these multiple hypotheses.

As for speaker representation methods for multi-lingual TTS, most of previous works mainly rely on "one-hot" representation to obtain speaker characteristic control [18, 19]. However, the learned pronunciation from other speakers could not be well transferred by using the "one-hot" representation, leading to poor performance during zero-shot cross transfer. In [12, 15], the authors replaced the "one-hot" representation with a separately trained speaker embedding network based on VoiceLoop [20] or Tacotorn2 [2]. Except for the complex procedure of training another network to obtain the speaker embedding, the speaker similarity for unseen speakers cloning is not that satisfactory. To alleviate this problem, we extend our previous works [21] on multi-speaker TTS models to cross-lingual TTS modeling. Specifically, a speaker encoder method with a random sampling mechanism is employed and a language adversarial loss [22] is further added to ensure the speaker embedding to be independent of languages. Another advantage of using the proposed speaker representation methods lies in helpful and easier to obtain cross-lingual transfer since the pronunciations are learned and shared by all speakers in a unified architecture.

As for the zero-shot cross-lingual transfer strategies, most previous works either simply update the parameters of the entire models or only the speaker embedding during zero-shot cross-lingual speaker adaptation [23]. In [18], a naive sampling strategy is proposed to deal with the data imbalance problem for different languages during training the "average" multi-lingual TTS models. We further extend it to zero-shot cross-lingual adaptation and adopt an efficient strategy with the help of other target lingual speakers' data and similar language-balanced sampling strategy, with the consideration of speaker identity could be obtained by source-lingual data and the pronunciations of target lingual could be transferred from other target lingual speakers' data easily.

In this paper, we investigate the zero-shot cross-lingual neural TTS in three different aspects that has not been fully explored in

[1] https://ybz-ops.github.io/icassp/icassp2022.html

previous works. (1) We propose a multi-stream text encoder to build a more robust and efficient unified text representation which is able to retain multiple hypotheses about input sequence. (2) We employ a novel speaker encoder method with a random sampling mechanism and further add a language adversarial loss to ensure the speaker representation to be independent of languages. (3) We further discuss how to use limited amount of source lingual data to achieve zero-shot cross-lingual transfer.

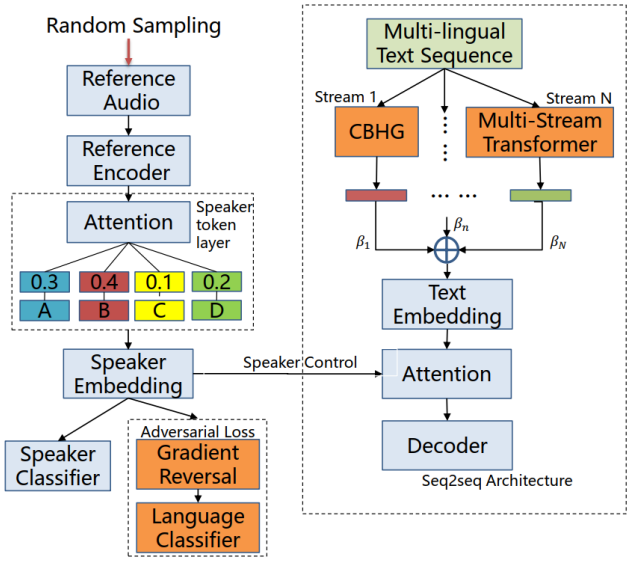

Fig. 1. System architecture of the proposed zero-shot cross-lingual TTS models using multi-stream text encoder and efficient speaker representation.

2. THE PROPOSED METHODS

Fig.1 illustrates the architecture of the proposed zero-shot cross-lingual TTS models using multi-stream text encoder and efficient speaker representation. It mainly consists of a multi-stream text encoder, a speaker encoder with a random sampling mechanism and a language adversarial classifier, and a decoder with attention mechanism to decode corresponding acoustic features.

2.1. Multi-stream Text Encoder

Different from [14], we remove the language embeddings and use phoneme as inputs to build the unified text representations for Chinese and English. To retain multiple hypotheses about input text representation, we propose a multi-stream text encoder based on multi-stream Transformer [17] and CBHG (1-D convolution (Cov1D) bank + highway network + bidirectional gated recurrent unit (GRU)) [1]. As shown in Fig.1, the input text sequences are fed into N-stream text encoders in parallel to generate respective intermediate feature representation vectors. Finally, the representation vectors are added with stream weight β_n. Though the stream weights could be learned by attention mechanisms, we use the same stream weight for all streams and set $N = 2$ for simplicity here.

2.1.1. *Multi-stream Transformer*

Since Transformer [24] is able to capture the long range dependency effectively, the first component for multi-stream text encoder is based on Transformer. However, since the multi-layer single-stream Transformer models [24] indeed force the fusion of alternatives at every processing stage (after every layer). There is just no storage for alternative hypotheses. In order to further retain multiple hypotheses about input representation which can then be combined at the end, we employ multi-stream Transformer here.

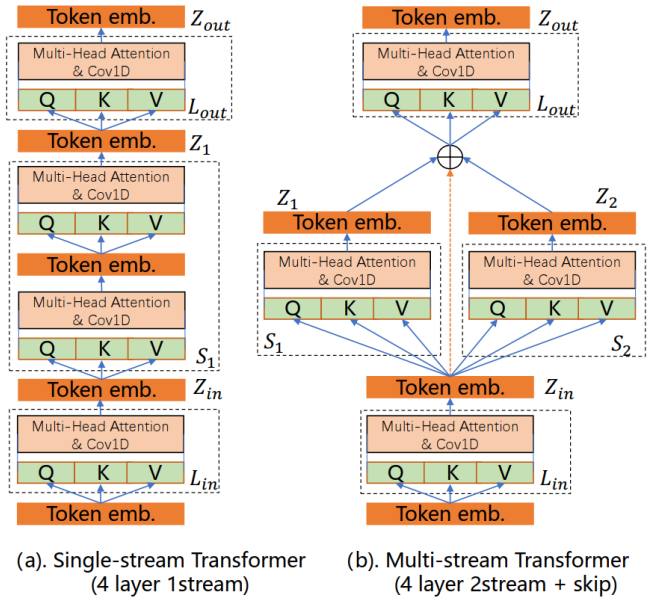

Fig. 2. Multi-stream Transformer encoder, where skip connection is shown as a dashed line. Each block is consisted of one multi-head attention and one Cov1D layer, and Add and LayerNorm [24] are omitted for simplicity.

Fig.2 shows a baseline single-stream Transformer encoder (2a) and a multi-stream Transformer encoder with two parallel streams containing one layer each (2b). Formally, we use L_{in}, L_{out} to represent the input and output of the transformer layer respectively, and use S_i to denote i-th stream with output Z_i. Then the output for the multi-stream encoder with a skip connection can be calculated as:

$$Z_{out} = L_{out}(\sum_{i=1}^{k} S_i(Z_{in}) + Z_{in}) \quad (1)$$

where k is the number of parallel streams. As shown in Fig.2(b), we employ a skip connection from the first joint layer to the last joint layer, which allows the model to factor in the initial jointly computed representation, and improves learnability at the same time.

2.1.2. *CBHG*

Another component for the multi-stream text encoder is based on CBHG module. In this module, the local and contextual information could be explicitly captured by convolutional filters, while the sequential information could be effectively modeled by a bidirectional GRU.

2.2. Efficient Speaker Representations

We extend our previous works [21] and further propose an efficient speaker representation method with random sampling mechanism and a language adversarial classify, aiming to extract speaker embedding features that are independent of content information and language identity.

Specifically, for each text and audio pairs $\{x_m^i, y_m^i\}, i \in [1, U]$, where U is the number of utterances for speaker m from all languages, then an audio utterance y^r is randomly sampled from all the utterances within the same speakers m,

$$y^r = y^{Random(U)} \quad (2)$$

where $Random(U)$ could be any positive integer that randomly generated from $[1, U]$. Then the randomly sampled audio utterance y^r is fed into the reference encoder directly to generate a fixed-length vector, namely reference embedding. In order to design a more robust speaker latent space, we follow the same methods in our previous works and use M speaker tokens to represent the variety of speaker. Thus each speaker embedding could be represented by a set of combination weights over these speaker tokens. After that, a speaker classifier is added on the speaker embeddings to ensure the speaker embeddings to be independent of contents and utterances within the same speakers, as well as to be more discriminative for different speakers. Finally, since it's common for each speaker to speak only one language, speaker embedding above may be correlated with language. In order to obtain language-independent speaker embedding, we insert a gradient reversal layer [22] prior to the language classifier. Another point that differs from our previous works [21] is that we merely employ speaker embedding here rather than concatenating it with "one-hot" representation, which is proved to be helpful for cross-lingual transfer since the pronunciations could be learned from other target-lingual speakers and be transferred more easily without "one-hot" representation.

2.3. Zero-shot Cross-lingual Transfer Strategies

From Sec 2.2, it's known that the proposed models are able to disentangle pronunciation and speaker representations. Specifically, the speaker characteristics could be captured using source-lingual data of the new speaker, while the pronunciation of cross-lingual could be learned by other speakers' target-lingual data and transferred easily thanks to the shared pronunciation learning structure. Then an initial and direct transfer strategy is to fine-tune the models using only the source-lingual data of the new speaker.

To achieve better generalization and cross-lingual synthesis, another transfer strategy is to fine-tune the models using the source-lingual data of the new speaker and all the target-lingual data of existing speakers. In most cases, the amount of source-lingual data for the new speaker is very limited, whereas the data amount for target-lingual data of other speakers is very large. To alleviate the data imbalance problem [18] during zero-shot cross-lingual transfer, the following language-balanced sampling strategy is adopted:

$$p^i = (U_i / \sum_l^L U_l)^\alpha \tag{3}$$

where U_l is the number of utterances for language l, $\alpha \in [0, 1]$,$l \in [1, L]$, in which $\alpha = 0$ represents a uniform distribution and $\alpha = 1$ denotes the true data distribution. We use $\alpha = 0.1$ here.

3. EXPERIMENTS

We use internal multi-speaker corpus of English and Chinese for training the proposed multi-speaker and multi-lingual TTS models. For Chinese, the multi-speaker corpus is about 120 hours in total and recorded by 60 professional Chinese speakers. The English data sets consist of 50 speaker, with each speaker recorded 500 sentences. Gender, age and content coverage are all considered in the design of these two multi-speaker corpus. We use 16-bit, 16 kHz sampling rate for all experiments. The spectral feature analysis operates on 20-ms windows with a 10-ms frame offset using LPCNet [25], with each frame of acoustic feature includes 30 dimensional Bark-frequency cepstral coefficients (BFCC), a pitch period, and a pitch correlation (between 0 and 1). For subjective tests, the audio naturalness is evaluated by 20 human testers in 5-point mean opinion score (MOS) and AB preference test. The speaker similarity is also evaluated by 5-point MOS.

3.1. Model Details

Our model is based on our previous works [21] and we further extend it to support multi-lingual TTS modeling. We use phoneme for text representation here. For speaker encoder, the architectures and hyper parameters are similar as that in [21] except we add a language adversarial classifier and remove the "one-hot" speaker representation. The language adversarial classifier is made up of a fully-connected network with one 256 unit hidden layer followed by a softmax layer. The speaker token number M is set to 10. For multi-stream text encoder, the hyper parameters of CBHG based stream is kept the same as that in [21], while the multi-stream Transformer based stream has 6 self-attention blocks totally (including 4 parallel streams of 1 layer each, 1 input layer and 1 out layer). Each block includes 8-head self-attention and a Cov1D layer. Both the dimension of input embedding and position embedding is set to 256. The architectures and hyper parameters of decoder and attention modules are all kept the same as that in [21]. The entire multi-speaker and multi-lingual TTS model is trained jointly using the Adam algorithm [26] with a learning rate decay and a batch size of 16, which starts from 0.001 and decays exponentially after 50K steps. Gradient cliping with factor 0.5 is applied to the gradient reversal layer during training. For speaker adaptation, a fixed learning rate of 0.0001 is used. During inference, we randomly choose one speech audio from the training set as a reference audio and use it for speaker embedding extraction.

To reconstruct audios, an "average" LPCNet vocoder [25] is trained using the multi-speaker and multi-lingual data set. The "average" LPCNet model is fine-tuned and used as a vocoder for target speaker by updating the entire model using corresponding data.

3.2. Evaluation of Zero-shot Cross-lingual Transfer

Though the proposed models are able to support voice cloning across any languages in training set, we focus on synthesizing Chinese speakers' English speech here since it is the most common situation in our practice. Therefore, we adapt our models to two unseen Chinese speakers (including 1 male and 1 female) without any English data firstly. Each speaker recorded 400 Chinese utterances. For simplicity, we randomly choose one unseen speaker for ablation studies.

3.2.1. Evaluation of multi-stream encoder

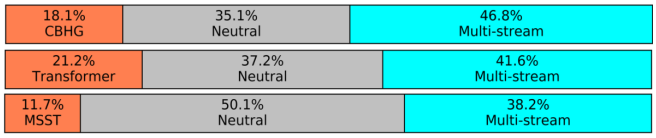

Fig. 3. The results of AB preference for different text encoders, with confidence intervals of 95% and p-value < 0.0001 from t-test.

Two baseline models, single-stream of CBHG [1] and Transformer [24] based encoder are built for comparisons. We randomly select 30 test utterances with 10 intra-lingual utterances and 20 cross-lingual utterances as the evaluation set (not included in training set). We first conduct AB preference tests on these 30 audio pairs generated by our proposed models and these two baseline models, respectively. The AB preference test results are shown in Fig.3, from which we could find the proposed multi-stream model receives the most preferences. This can be explained by that our

proposed multi-stream text encoder could take advantages of both Transformer and CBHG, resulting in more explicit text sequence modeling for both long-term dependencies and local contextual information. Furthermore, for multi-stream based encoder, we also replace the multi-stream Transformer with single-stream Transformer (MSST), and the AB preference results in Fig.3 demonstrate the advantages over the single-stream Transformer.

3.2.2. *Evaluation of speaker representation methods*

In this section, three different speaker latent space modeling methods are compared, including "one-hot" representation in [18], our previous speaker encoder (P-SE) method in [21] (the differences are presented in Sec.2.2) and the proposed speaker encoder (Proposed-SE) method. The AB preference test results are shown in Fig.4, from which we could find that the proposed method receives much more preferences than [18]. This can be explained by that comparing with one-hot representation by a fixed lookup-table, the speaker embeddings extracted by the proposed methods should contain speaker characteristics information that may be useful for decoding corresponding speaker's acoustic features. The AB preference test results between P-SE and Proposed-SE further demonstrates the importance of decoupling the language information from speaker representation. Another remarkable advantage over P-SE [21] and "one-hot" [18] is that the proposed method is helpful for cross-lingual transfer since the cross-lingual pronunciations could be learned from other target-lingual speakers and be transferred more easily without "one-hot" representation. Lastly, the proposed speaker representation can be calculated offline in advance, so it does not bring any additional computational cost during inference.

Fig. 4. The results of AB preference for different speaker representation methods, with confidence intervals of 95% and p-value < 0.0001 from t-test.

3.2.3. *Evaluation of the zero-shot cross-lingual transfer strategies*

Two zero-shot cross-lingual transfer strategies are investigated here. The first one (Baseline) is to refine the "average" models using only the source-lingual data of the new speaker. Another transfer strategy (Proposed) is to refine the models using the source-lingual data of the new speaker and all the target-lingual data of existing speakers with a language-balanced sampling strategy. The AB preference results presented in Fig.5 show the superiority of the proposed transfer strategy. Examination of testers' comments also show the proposed model performs much better on overall naturalness, similarity and sounds more stable.

Fig. 5. The results of AB preference for different transfer strategies, with confidence intervals of 95% and p-value < 0.0001 from t-test.

3.2.4. *Overall performance*

We then combine the best configuration mentioned above and make zero-shot cross-lingual transfer for 2 unseen Chinese speakers. For comparison, we use "one-hot" speaker representation without other techniques as baseline [18]. We do MOS tests for both naturalness and speaker similarity to evaluate the proposed method. The results are shown in Tab. 1. It's seen that both naturalness and speaker similarity of the generated audios are greatly improved for all speakers using the proposed methods, with an average absolute improvement of 0.38 and 0.27 in MOS respectively.

Table 1. The MOS of naturalness and speaker similarity for zero-shot cross-lingual transfer, with 95% confidence intervals computed from t-test.

Models	Speaker	Baseline	Proposed
Naturalness	female	3.89 ± 0.02	4.17 ± 0.07
	male	3.65 ± 0.04	4.13 ± 0.09
Similarity	female	4.25 ± 0.04	4.57 ± 0.03
	male	4.12 ± 0.08	4.33 ± 0.05

3.3. Extension to Few-shot Cross-lingual Transfer

We further extend the proposed methods to build bilingual TTS models for two unseen speakers (donated as female-bi and male-bi in Tab. 2), with each speaker recorded 1,000 Chinese utterances and 300 English utterances. Different from zero-shot cross-lingual transfer, we update the models using only the bilingual data of the new speakers. We do MOS tests for naturalness to evaluate the proposed methods and the results are shown in Tab. 2. It's found that our proposed model can achieve significantly better results than the baseline [18] in terms of the naturalness, which indicates the effectiveness of the proposed methods for few-shot cross-lingual transfer.

Apart from that, we also train a mono-lingual multi-speaker English TTS model and adapt it with the 300 English utterances above. We then conduct an AB preference tests on the generated audios pairs by this mono-lingual model and the proposed multi-lingual model. The AB preference result in Fig. 6 further demonstrates the superiority of the proposed multi-lingual model than the mono-lingual model.

Table 2. The naturalness MOS for few-shot cross-lingual transfer, with 95% confidence intervals computed from t-test.

Models	female-bi	male-bi
Baseline	4.24 ± 0.05	4.13 ± 0.05
Proposed	4.38 ± 0.05	4.30 ± 0.06

Fig. 6. The results of AB preference, with confidence intervals of 95% and p-value < 0.0001 from t-test.

4. CONCLUSIONS

We propose a novel method for zero-shot cross-lingual TTS task by using multi-stream text encoder and efficient speaker representation. Specifically, a unified multi-stream text encoder that enjoys both advantages of multi-stream Transformer and CBHG is proposed to retain multiple hypotheses about input representations. Then the speaker encoder with a random sampling mechanism and a language adversarial loss is proposed to extract speaker embedding features that are independent of both content information and language identity. Meanwhile, different zero-shot cross-lingual transfer strategies are also explored. Experimental results demonstrate that our proposed model not only can achieves a significant improvement on both naturalness and similarity for zero-shot cross-lingual transfer, but also improves the naturalness for few-shot cross-lingual transfer. In future, we will extend our work to more languages and speakers.

5. REFERENCES

[1] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al., "Tacotron: Towards end-to-end speech synthesis," in *INTERSPEECH*, 2017.

[2] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al., "Natural TTS synthesis by conditioning WaveNet on mel spectrogram predictions," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4779–4783.

[3] Yibin Zheng, Jianhua Tao, Zhengqi Wen, and Jiangyan Yi, "Forward–backward decoding sequence for regularizing end-to-end tts," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2067–2079, 2019.

[4] Yuxuan Wang, RJ Skerry-Ryan, Ying Xiao, Daisy Stanton, Joel Shor, Eric Battenberg, Rob Clark, and Rif A Saurous, "Uncovering latent style factors for expressive speech synthesis," *arXiv preprint arXiv:1711.00520*, 2017.

[5] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron J Weiss, Rob Clark, and Rif A Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *Proceedings of the 35th International Conference on Machine Learning,ICML 2018*, vol. 80, pp. 4700–4709.

[6] Yuxuan Wang, Daisy Stanton, Yu Zhang, RJ Skerry-Ryan, Eric Battenberg, Joel Shor, Ying Xiao, Fei Ren, Ye Jia, and Rif 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, ICML 2018*. vol. 80, pp. 5167–5176, PMLR.

[7] Zvi Kons, Slava Shechtman, Alexander Sorin, Carmel Rabinovitz, and Ron Hoory, "High quality, lightweight and adaptable TTS using LPCNet," in *INTERSPEECH*, 2019, pp. 176–180.

[8] Yutian Chen, Yannis Assael, Brendan Shillingford, David Budden, Scott Reed, Heiga Zen, Quan Wang, Luis C Cobo, Andrew Trask, Ben Laurie, et al., "Sample efficient adaptive text-to-speech," in *7th International Conference on Learning Representations, ICLR*. 2019, OpenReview.net.

[9] Heiga Zen, Norbert Braunschweiler, Sabine Buchholz, Mark JF Gales, Kate Knill, Sacha Krstulovic, and Javier Latorre, "Statistical parametric speech synthesis based on speaker and language factorization," *IEEE transactions on audio, speech, and language processing*, vol. 20, no. 6, pp. 1713–1724, 2012.

[10] Bo Li and Heiga Zen, "Multi-language multi-speaker acoustic modeling for LSTM-RNN based statistical parametric speech synthesis," in *INTERSPEECH*, 2017, pp. 2648–2472.

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

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

[13] Liumeng Xue, Wei Song, Guanghui Xu, Lei Xie, and Zhizheng Wu, "Building a mixed-lingual neural TTS system with only monolingual data," in *INTERSPEECH*, 2019, pp. 2168–2172.

[14] Yu Zhang, Ron J Weiss, Heiga Zen, Yonghui Wu, Zhifeng Chen, RJ Skerry-Ryan, Ye Jia, Andrew Rosenberg, and Bhuvana Ramabhadran, "Learning to speak fluently in a foreign language: Multilingual speech synthesis and cross-language voice cloning," *arXiv preprint arXiv:1907.04448*, 2019.

[15] Yuewen Cao, Xixin Wu, Songxiang Liu, Jianwei Yu, Xu Li, Zhiyong Wu, Xunying Liu, and Helen Meng, "End-to-end code-switched TTS with mix of monolingual recordings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2019, pp. 6935–6939.

[16] Omer Sagi and Lior Rokach, "Ensemble learning: A survey," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 8, no. 4, pp. e1249, 2018.

[17] Mikhail Burtsev and Anna Rumshisky, "Multi-stream transformers," *arXiv preprint arXiv:2107.10342*, 2021.

[18] Jingzhou Yang and Lei He, "Towards universal text-to-speech.," in *INTERSPEECH*, 2020, pp. 3171–3175.

[19] Andrew Gibiansky, Sercan Arik, Gregory Diamos, John Miller, Kainan Peng, Wei Ping, Jonathan Raiman, and Yanqi Zhou, "Deep Voice 2: Multi-speaker neural text-to-speech," in *Advances in neural information processing systems*, 2017, pp. 2962–2970.

[20] Yaniv Taigman, Lior Wolf, Adam Polyak, and Eliya Nachmani, "VoiceLoop: Voice fitting and synthesis via a phonological loop," in *International Conference on Learning Representations, ICLR*, 2018.

[21] Yibin Zheng, Xinhui Li, and Li Lu, "Investigation of fast and efficient methods for multi-speaker modeling and speaker adaptation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6618–6622.

[22] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[23] Qiao Tian, Zewang Zhang, Heng Lu, Ling-Hui Chen, and Shan Liu, "FeatherWave: An efficient high-fidelity neural vocoder with multi-band linear prediction," in *INTERSPEECH*, 2020, pp. 2458–2462.

[24] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu, "Neural speech synthesis with transformer network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, vol. 33, pp. 6706–6713.

[25] Jean-Marc Valin and Jan Skoglund, "LPCNet: Improving neural speech synthesis through linear prediction," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5891–5895.

[26] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.